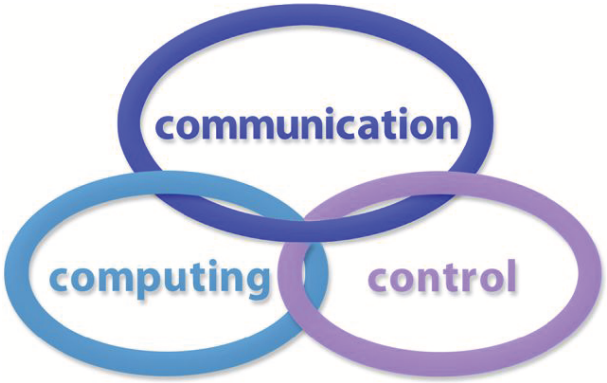
INTERNATIONAL JOURNAL

of

COMPUTERS, COMMUNICATIONS & CONTROL

With Emphasis on the Integration of Three Technologies

IJCCC



Year: 2010 Volume: 5 Number: 4 (November)

Agora University Editing House

CCC Publications

www.journal.univagora.ro

International Journal of Computers, Communications & Control

EDITORIAL ADDRESS:
R&D Agora Ltd. / S.C. Cercetare Dezvoltare Agora S.R.L.
Piaţa Tineretului 8, Oradea, jud. Bihor, Romania, Zip Code 410526
Tel./ Fax: +40 359101032
E-mail: ijccc@univagora.ro, rd.agora@univagora.ro, ccc.journal@gmail.com
Journal website: www.journal.univagora.ro

DATA FOR SUBSCRIBERS
Supplier: Cercetare Dezvoltare Agora Srl (Research & Development Agora Ltd.)
Fiscal code: RO24747462
Headquarter: Oradea, Piata Tineretului Nr.8, Bihor, Romania, Zip code 410526
Bank: MILLENNIUM BANK, Bank address: Piata Unirii, str. Primariei, 2, Oradea, Romania
IBAN Account for EURO: RO73MILB0000000000932235
SWIFT CODE (eq.BIC): MILBROBU

International Journal of Computers, Communications & Control

Mario de J. Pérez Jiménez
Dept. of CS and Artificial Intelligence
University of Seville
Sevilla, Avda. Reina Mercedes s/n, 41012, Spain
marper@us.es

Dana Petcu
Computer Science Department
Western University of Timisoara
V.Parvan 4, 300223 Timisoara, Romania
petcu@info.uvt.ro

Radu Popescu-Zeletin
Fraunhofer Institute for Open
Communication Systems
Technical University Berlin, Germany
rpz@cs.tu-berlin.de

Imre J. Rudas
Institute of Intelligent Engineering Systems
Budapest Tech
Budapest, Bécsi út 96/B, H-1034, Hungary
rudas@bmf.hu

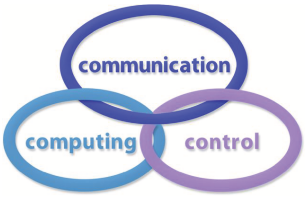
Athanasios D. Styliadis
Alexander Institute of Technology
Agiou Panteleimona 24, 551 33
Thessaloniki, Greece
styl@it.teithe.gr

Gheorghe Tecuci
Learning Agents Center
George Mason University
University Drive 4440, Fairfax VA 22030-4444, U.S.A.
tecuci@gmu.edu

Horia-Nicolai Teodorescu
Faculty of Electronics and Telecommunications
Technical University "Gh. Asachi" Iasi
Iasi, Bd. Carol I 11, 700506, Romania
hteodor@etc.tuiasi.ro

Dan Tufiş
Research Institute for Artificial Intelligence
of the Romanian Academy
Bucharest, "13 Septembrie" 13, 050711, Romania
tufis@racai.ro

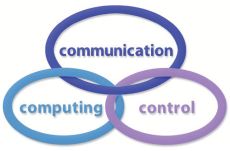
Lotfi A. Zadeh
Department of Computer Science and Engineering
University of California
Berkeley, CA 94720-1776, U.S.A.
zadeh@cs.berkeley.edu



International Journal of Computers, Communications & Control

Short Description of IJCCC

Title of journal: International Journal of Computers, Communications & Control
Acronym: IJCCC
International Standard Serial Number: ISSN 1841-9836, E-ISSN 1841-9844
Publisher: CCC Publications - Agora University
Starting year of IJCCC: 2006
Founders of IJCCC: Ioan Dzitac, Florin Gheorghe Filip and Mişu-Jan Manolescu
Logo:



Number of issues/year: IJCCC has 4 issues/odd year (March, June, September, December) and 5 issues/even year (March, September, June, November, December). Every even year IJCCC will publish a supplementary issue with selected papers from the International Conference on Computers, Communications and Control.
Coverage:

- Beginning with Vol. 1 (2006), Supplementary issue: S, IJCCC is covered by Thomson Reuters - SCI Expanded and is indexed in ISI Web of Science.
- Journal Citation Reports/Science Edition 2009:
  - Impact factor = 0.373
  - Immediacy index = 0.205
- Beginning with Vol. 2 (2007), No.1, IJCCC is covered in EBSCO.
- Beginning with Vol. 3 (2008), No.1, IJCCC, is covered in SCOPUS.

Scope: IJCCC is directed to the international communities of scientific researchers in universities, research units and industry. IJCCC publishes original and recent scientific contributions in the following fields: Computing & Computational Mathematics; Information Technology & Communications; Computer-based Control.
Unique features distinguishing IJCCC: To differentiate from other similar journals, the editorial policy of IJCCC encourages especially the publishing of scientific papers that focus on the convergence of the 3 "C" (Computing, Communication, Control).
Policy: The articles submitted to IJCCC must be original and previously unpublished in other journals. The submissions will be revised independently by at least two reviewers and will be published only after completion of the editorial workflow.

Contents

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 4, pp. 418-431

Probabilistic Proximity-aware Resource Location in Peer-to-Peer Networks Using Resource Replication

M. Analoui, M. Sharifi, M.H. Rezvani

Morteza Analoui, Mohsen Sharifi,
Mohammad Hossein Rezvani
Iran University of Science and Technology (IUST)
16846-13114, Hengam Street, Resalat Square,
Narmak, Tehran, Iran
Email: {analoui,msharifi,rezvani}@iust.ac.ir

Abstract: Nowadays, content distribution has received remarkable attention in distributed computing researches and its applications typically allow personal computers, called peers, to cooperate with each other in order to accomplish distributed operations such as query search and acquiring digital contents. In a very large network, it is impossible to perform a query request by visiting all peers. There are some works that try to find the location of resources probabilistically (i.e. non-deterministically). They all have used inefficient protocols for finding the probable location of peers who manage the resources. This paper presents a more efficient protocol that is proximity-aware in the sense that it is able to cache and replicate the popular queries proportional to distance latency. The protocol dictates that the farther the resources are located from the origin of a query, the more should be the probability of their replication in the caches of intermediate peers. We have validated the proposed distributed caching scheme by running it on a simulated peer-to-peer network using the well-known Gnutella system parameters. The simulation results show that the proximity-aware distributed caching can improve the efficiency of peer-to-peer resource location services in terms of the probability of finding objects, overall miss rate of the system, fraction of involved peers in the search process, and the amount of system load.
Keywords: Distributed systems, Peer-to-Peer network, Content Distribution, Resource Location, Performance Evaluation.

1 Introduction

1.1 Motivation

A peer-to-peer (P2P) system is a distributed system consisting of interconnected peers who are able to self-organize into network topologies with the purpose of sharing resources such as CPU or bandwidth, capable of adapting to dynamic conditions of network, without requiring the support of a global centralized server [1]. The P2P systems are classified as unstructured and structured. In the structured systems such as CAN [2] the overlay topology is tightly controlled and files are placed at exact locations. These systems provide a distributed routing table, so that queries can be routed to the corresponding peer who manages the desired content. Unlike the structured systems, in the unstructured systems such as Gnutella [3] and KazaA [4] searching mechanisms are employed to discover the location of the resources. Each peer owns a set of resources to be shared with other peers. In general, the shared resource can be any kind of data which make sense, even records stored in a relational database. The most significant searching mechanisms include brute force methods (e.g. flooding the network with propagating queries in a breath-first or depth-first manner until the desired content is discovered) [1], probabilistic searches [5], routing indices [6], randomized gossiping, and so on.

Most of the current P2P systems such as Gnutella and KazaA fall within the category of P2P "content distribution" systems. A typical P2P content distribution system creates a distributed storage medium and allows doing services such as searching and retrieving query messages which are known as "resource location" services. The area of "content distribution systems" has a large overlap with the issue of "resource location services" in the literature. Fig. 1 illustrates a possible topology of the unstructured P2P networks. Each super-peer is a powerful master node that acts as a local central indexer for files which are shared by its local peers, whereas acts as an ordinary peer for other super-peers. In graph representation, each pair of peers is connected by an edge representing a TCP connection between them. The number of neighbors of a super-peer is called its out-degree. If two nodes are not

connected by an edge, they could communicate through an indirect path which passes across some other nodes. The length of a path through which two nodes communicate with each other is known as hop-count. Upon delivery of a query request message to a super-peer, it looks for matches over its local database. If any matches are found, it will send a single response message back to the node which has requested the query. If no match is found, the super-peer may forward the query request to its neighbor super-peers.

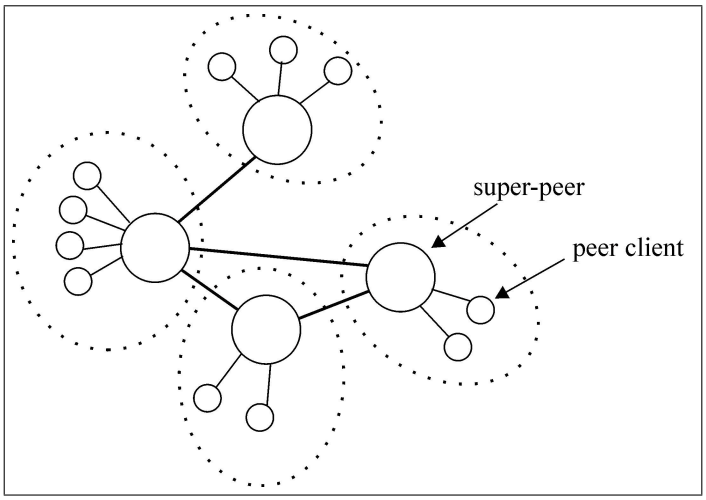


Figure 1: A typical super-peer network

In general, the performance of the P2P systems is strictly evaluated by metrics such as response time, number of hops, aggregated load, throughput, overall miss rate, fraction of participating nodes in the search operation, and so on. To meet these requirements, previous researches resorted to heuristics to locate the resources by incorporating proximity concerns.

1.2 Challenges

There already exists significant body of researches toward proximity-aware P2P systems. The proximity-aware resource location method in the P2P unstructured systems has been investigated in [7,8]. The proposed method uses flooding mechanism to forward a query request to all neighbors of a peer. It uses the hop-count as the proximity metric. In order to reduce the number of broadcast messages in the network, the header of each query message contains a time-to-live (TTL) field whose value is decremented at each hop. Finally, when the TTL reaches zero, the query message is dropped from the network. After locating the resource, a direct connection is established between the originating peer and the destined peers and the file is downloaded. The flooding approach employed in [7] is probabilistic in the sense that each peer replicates the query to its neighbors with a fixed probability.

An analytical study on the impact of proximity-aware methodology for the content distribution is presented in [9]. They have evaluated the performance of video streaming P2P network via key scalability metric, namely network load. Similar to the previous works, they have used the pair-wise latency of the peers as the proximity criterion.

Regard to the aforementioned works, in general, there are two strands of work concerning the proximity-aware methodology. First, there are works on content distribution via constructing the P2P topology [9]. Second, there are works on resource location services [7, 8]. These works assume a given topology setting such as mesh or tree for the P2P system. It has been shown in [10, 11] that finding an optimal-bandwidth topology for the P2P network is an NP-complete problem. So, we shall not try to solve the NP problem of topology construction here. Instead, we will try to optimize the proximity-aware resource locating problem within the given topology setting in the P2P system.

1.3 Contributions

In this paper, we are concerned with the design of a resource location service by using scalable proximity-aware distributed caching mechanism. We define the resource location service as "given a resource name, find with a proximity probability, the location of peers who manage the resource."

We use round-trip time (RTT) latency distance as the criterion for the probabilistic caching of each query. Each peer, upon receiving a query, at first searches its local cache. If the query is found, the peer returns it to the original requesting peer along with the reverse path which is traversed by the query. In this order, the so called query is cached in the memory of each intermediate node using replication method based on the proposed proximity-aware distributed caching mechanism. The probability of the resource replication and updating of the caches in each intermediate node is proportional to the latency distance between that node and the location where the resource is found. To the best of our knowledge, there has been no investigation on designing the proximity-aware probabilistic caching in the P2P systems.

The rest of the paper is organized as follows. Section 2 presents our proposed proximity-aware resource location mechanism along with its specification such as resource replication. Section 3 provides an analytical study of the probabilistic search method along with numerical results. Section 4 presents the experimental validation of the proposed mechanism. Finally, we discuss the related works in Section 5, and conclude in Section 6.

2 Proximity-aware Distributed Caching

Each pair of nodes is associated with a latency distance representing the average RTT experienced by communication between them. The latency distance corresponding to a specific pair of nodes may be measured either directly through ping messages, or estimated approximately through a virtual coordinate service [12]. Due to large number of nodes in a P2P system, we have adopted the latter approach to measure the latency between each pair of nodes. The virtual coordinate service which has been used in [12] is provided by VIVALDI [13], a distributed protocol developed at MIT. Due to space limitations, we do not explain the details of the virtual coordinate service here. Interested readers can refer to [12, 13] for it. As sated above, some works use the hop-count as the criterion for the distance estimation rather than using VIVALDI estimation method.

Every super-peer in our system has a local index table (LIT) that points to locally managed resources (such as files, Web pages, processes, and devices). Each resource has a location-independent globally unique identifier (GUID) that can be provided by developers of the P2P network using different means. In a distributed online bookstore application, developers could use ISBNs as GUIDs [8]. Each super-peer has a directory cache (DC) that points to the presumed location of resources managed by other super-peers. An entry in the DC is a pair (id, loc) in which id is the GUID of a resource and loc is the network address of a super-peer who might store the resource locally. Each peer has a local neighborhood defined as the set of super-peers who have connected to it. Table 1 and Table 2 provide a high-level description of the proposed proximity-aware distributed caching mechanism. The QuerySearch (QS) procedure describes the operations in which a source is searching a resource, namely . The string path $s_1, ..., s_m$ is the sequence of super-peers who have received this message so far. This sequence is used as a reverse path to the source. The header of each query message contains a TTL field which is used to control the depth of the broadcast tree. For example, Gnutella has been implemented with a TTL parameter equal to 7. The QueryFound (QF) procedure indicates that the resource being searched by source has been found at super-peer . In this procedure, the max_latency is the latency distance between the super-peer who manages and the farthest super-peer in the reverse path.

Figure 2 illustrates a typical unstructured P2P computing system. A given resource is managed by nodes $S_3, S_5, S_7, S_8, S_9, S_{13}$, and S_{18}. The resource is saved as a file on disk memory corresponding to clients who are clustered by aforementioned super-peers. Inspecting the DC of super-peers S_{11}, S_{12}, and S_{16} for example, reveals that the resource is located in the super-peer S18. The LITs corresponding to the nodes $S_1, S_2, S_4, S_6, S_{10}, S_{11}, S_{12}, S_{14}, S_{15}, S_{16}$, and S_{17} are empty; indicating that they are not themselves managers of the resource . Also, the DCs corresponding to the nodes $S_1, S_{10}, S_{11}, S_{12}, S_{14}, S_{15}$, and S_{17} are empty; indicating that they do not know the address of the owner of the resource res.

Each super-peer upon receiving the QS message, at first searches within its LIT. If it finds the resource in the LIT, it will return a QF message. The QF message is forwarded to the source following the reverse path which has been used by the QS message. It updates the DCs corresponding to each of the intermediate nodes as well. The contribution of our work emerges at this point where the QF message updates the LIT in each of the intermediate nodes using replication of resources based on the proposed proximity-aware distributed caching mechanism. The probability of resource replication and updating of the LIT corresponding to each intermediate node is proportional to the latency distance between that node and the location where the resource has been found. To this end, each intermediate node *r* performs the following actions with a probability proportional to the latency distance between itself and the peer who has been found as the manager of the resource:

1) establishing a TCP connection with the super-peer who manages the resource.

2) downloading the resource object and saving into client c who has enough available space.
3) updating the LIT by adding the entry *(res,c)* to it.

Table 1: QuerySearch message received by super-peer *r*.

```
QuerySearch(source, res, (s_1, ..., s_m), TTL)
begin
  If res ∈ LIT then
  begin
    max_latency= max{(lat(r,s_1),lat(r,s_2),...,lat(r,s_m))}
    send QueryFound(source, res, max_latency, (s_1, ..., s_m-1), r) to s_m
  end
  else if (res, loc) ∈DC then
    /* send request to presumed location */
    Send QuerySearch(source, res, (s_1, ..., s_m, r), TTL-1) to loc
    else if (TTL > 0) then
      for v_i = v_1 to v_m do      /* v_i ∈ N(r) */
      begin
        max_latency= max{(lat(r,v_1),lat(r,v_2),...,lat(r,v_m))}
        Send QuerySearch(source, res, (s_1, ..., s_m, r), TTL-1) with probability p to v_i.
        /* The probability p is proportional to lat(r,v_i) / max_latency */
      end
end
```

Table 2: QueryFound message received by super-peer *r*.

```
QueryFound (source, res, max_latency, (s_1, ..., s_m), v)
begin
  if r ≠ source then
  begin
    add (res, v) to DC
    with probability proportional to lat(r,v) / max_latency do
    begin
      Connect to super-peer v to get resource res from it
      find a local client, c, with enough available memory
      add (res, c) to LIT
    end
    send QueryFound(source, res, max_latency, (s_1, ..., s_m-1), v) to s_m
  end
  else   /* end of query search process */
    connect to super-peer v to get resource res from it.
end
```

If the super-peer does not find the resource in its LIT but finds it in the DC, it will send a QS message to the super-peer who is pointed to by that DC. If this super-peer no longer has the resource, the search process will be continued from that point forward. If a super-peer does not find the resource in its LIT or DC, it will forward the request to each super-peer in its neighborhood with a certain probability p which is called the *broadcasting probability*. This probability could vary with the length of the path that the request traverses.

Figure 3 illustrates how a QS message would be propagated in the network. In the figure, the maximum number of nodes to be traversed by a QS message is defined to be equal to 3 hops (apart from the source node). Similar to Gnutella, our system uses a Breath-First-Search (BFS) mechanism in which the depth of the broadcast tree is limited by the TTL criterion. The difference is that in Gnutella every node receiving a query forwards the message to all of its neighbors, while in our proposal, the propagation is performed probabilistically and is done if the query is not found neither in the LIT nor in the DC of a node.

In Fig. 3, the QS message originating from source S_1 is probabilistically sent to super-peers S_2, S_3, and S_4 due to search for the resource *res*. The super-peer S_3 finds the resource in its LIT, but S_2 and S_4 do not find such an entry, hence probabilistically forward the message to the nodes registered in their DCs. Note that the super-peer S_4 does not forward the message to S_{10} because, for example, in this case the probability of forwarding is randomly selected to be zero.

Figure 4 illustrates an example of returning QF messages in a reversed path from the location where the resource *res* is found, to the node who has originated the query request. The QF message is routed to the source (node S_1) following the reverse path which is used by the QS message. The QF message updates the corresponding DC of each intermediate node based on the proposed proximity-aware distributed caching mechanism. The probability

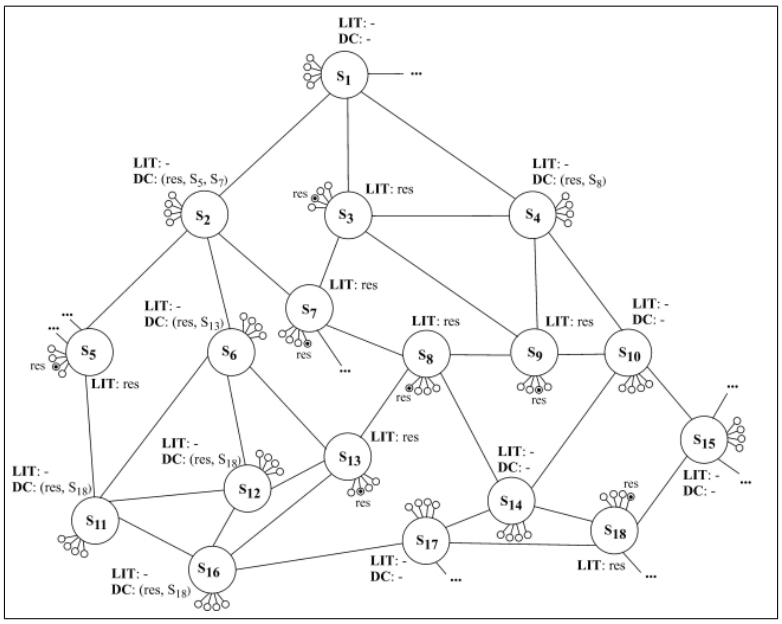


Figure 2: The topology of a typical unstructured P2P system

of replication and caching the resource object in the LIT of each intermediate node is proportional to the latency distance between that node and the location where the resource is found. The closer is the intermediate node to the discovered resource; the less will be the probability of caching the resource in the node's LIT. This probability is shown by a graphical representation with partially boldfaced circles. In the sequence of nodes which consists of S_1, S_2, S_6, and S_{13}, the node S_6 caches the address of the resource *res* with the least probability; whereas node $_1$ caches it with the most probability. The probability of caching the resource *res* by S2 is larger than that of S_6 and smaller than that of S1.

3 The Analytic Study of the Proposed Mechanism

Considering that the corresponding super-peer of a query is located at the root of a tree of height d, each super-peer r has k_r neighbors, where k_r follows a power-law distribution. In such a distribution, the majority of nodes have relatively few local connections to other nodes, but a significant small number of nodes have large wide-ranging sets of connections. The power-law distribution gives small-world networks a high degree of fault tolerance, because random failures are most likely to eliminate nodes from the poorly connected majority [14]. Hence, each query originating from a client must visit at most d-1 super-peers. Note that the levels of the tree are numbered 1,...,d from the root down. Let q_i be the probability that a super-peer at level i has a local index for the resource *res*. Let R(i) be the number of super-peers at level i who receive a QuerySearch message from upstream super-peers, and S(i) be the number of super-peers at level i who may forward the QuerySearch message to downstream super-peers one level down. So, it must be that R(1)=0 and q_1=0.

The reason for $q_1 = 0$ lies in the fact that the resource *res* only may be found in the LIT of the super-peers who are located in the levels 2,...,d of the broadcast tree.

Let us consider j super-peers at level i-1 may not find the resource in their LIT and forward query request message one level down. Each of j super-peers, say r , located at level i-1 can select at most k_r super-peers among its neighbors to send a query request message. Let us suppose that each super-peer forwards these query messages independently to the children who are located at level i-1 with probability p_i. Let $m_1, ..., m_j$ be the number of level i children to receive the query request message from super-peers 1,...,j located at level i-1, respectively. Let us assume n super-peers receive these query requests at level i. So, we can define the tuple S as follows

$$S(j,n,k_1,...,k_j) = \{\overrightarrow{m} = (m_1,...,m_j) | \sum_{s=1}^{j} m_s = n \;\&\; 0 \leqslant m_s \leqslant k_s\} \tag{1}$$

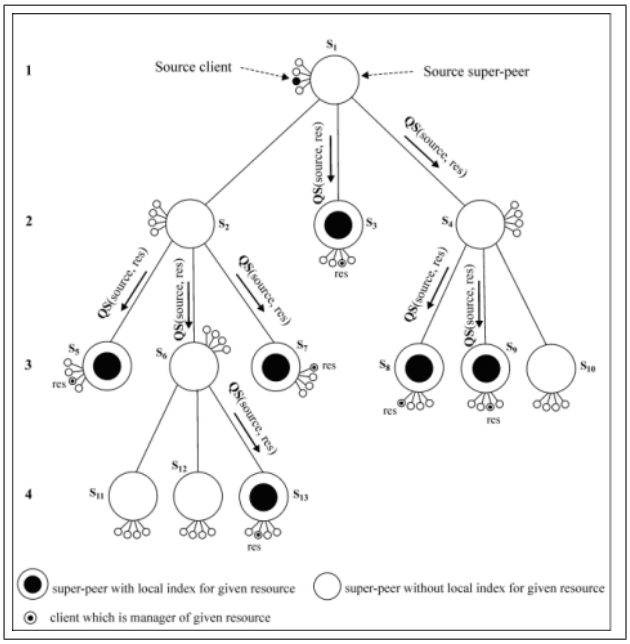


Figure 3: Forwarding a QS message using maximum hop-count equal to 3

Where, $k_1,...,k_j$ are out-degrees of nodes 1,...,j which are located at level i-1 , respectively. Now we are in a position to define conditional probability Pr[R(i)=n|S(i-1)=j] . This is the probability of receiving the *QuerySearch* messages by n super-peers at level i given that j super-peers at level i-1 may forward the messages one level down. With respect to the above premises we have:

$$Pr[R(1)=n|S(i-1)=j] = \sum_{\vee \overrightarrow{m} \in S(j,n,k_1,...k_s)} \prod_{s=1}^{j} \binom{k_s}{m_s} \times p_i^{m_s} \cdot (1-p_i)^{k_s m_s} \tag{2}$$

For example let us consider that,j-3, n=5, $k_1 = 2$, $k_2 = 3$ and $k_3 = 2$. Three nodes at level i-1 forward the QuerySearch message to five super-peers at level i. Thus, S(3,4,2,3,2)={(0,3,2), (2,1,2),(2,3,0),(2,2,1), (1,2,2),(1,3,1)}. This example shows that Eq. (2) does not depend on the order of appearance of values $m_1,...,m_j$ in $\overline{m}$.

We can derive the probability Pr[R(i)=n] that n super-peers receive a QuerySearch message at level i as follows

$$Pr[R(i)=n] = \sum_{j=0}^{n_{i-1}} Pr[R(i)=n|S(i-1)=j] \cdot Pr[S(i-1)=j] \tag{3}$$

Where, n=0,...,$\sum_{r=1}^{j} k_r$, and n_{i-1} is the total number of nodes located at level i-1. Eq. (3) can be computed recursively by following conditions:

Pr[R(1)=1]=1, Pr[R(1=0)]=0, Pr[R(i)=0|S(i-1)=0]=1, and

Pr[R(i)=n|S(i-1)=0]=0 for n>0.

We are now in a position to compute the average number of super-peers, , involved in query search (apart from the source) as follows

$$\overline{N} = \sum_{i=2}^{d} \sum_{n=0}^{n_i} n \cdot Pr[R(i)=n] \tag{4}$$

Where, n_i is the total number of nodes located at level i. Note that S(i-1) is not necessarily equal to R(i-1). The probability Pr[S(i-1)=j] in Eq. (3) can be computed by conditioning on r super-peers that receive *QuerySearch* messages at level i-1. So, we have:

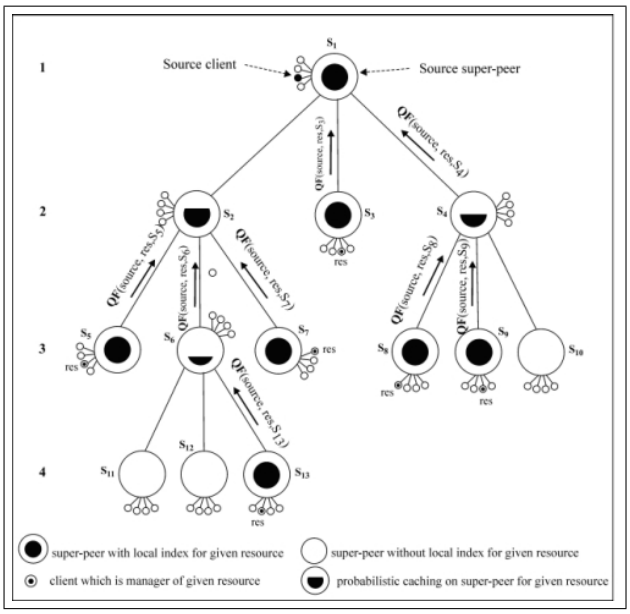


Figure 4: Forwarding a QS message using maximum hop-count equal to 3

$$Pr[S(i-1)=j]=\sum_{r=j}^{n_{i-1}} Pr[S(i-1)=j|R(i-1)=r].Pr[R(i-1)=r] \tag{5}$$

The above equation can be simplified as follows

$$Pr[S(i-1)=j]=\sum_{r=j}^{n_{i-1}} \binom{r}{j} q_{i-1}^{r-j}(1-q_{i-1})^{j}.Pr[R(i-1)=r] \tag{6}$$

The probability P_f that a local index for a resource is found can be computed as

$$P_f=1-\prod_{i=2}^{d} Pr[\textit{no local index is found at level } i] \tag{7}$$

Formally, the above equation can be expressed as follows

$$P_f=1-\prod_{i=2}^{d}\sum_{n=0}^{n_i}(1-q_i)^n.Pr[R(i)=n] \tag{8}$$

Now, the probability that an entry for the resource is not found, namely P_{nf}, can be computed as follows

$$P_{nf}(s)=\sum_{n=0}^{n_i}(1-q_s)^n.Pr[R(s)=n] \tag{9}$$

For computing the average number of hops required to find a local index for a resource, namely $\overline{H}$, we first define $P_f^{min}(i)$ with the assumption that the first level wherein a resource is found at level i:

$$P_f^{min}(i)=\frac{\prod_{s=2}^{i-1}P_{nf}(s)][1-P_{nf}(i)]}{P_f} \tag{10}$$

Now, the average number of hops can be computed as follows

$$\overline{H}=\sum_{i=2}^{d}(i-1)P_f^{min}(i) \tag{11}$$

As mentioned earlier, Eq. (11) is approximately equivalent to mean latency distance. Before proceeding, let us define F as the ratio between the average number of super-peers, namely $\overline{N}$, involved in the query operation and the total number of super-peers except the originating super-peer:

$$F = \frac{\overline{N}}{\sum_{i=2}^{d} n_i} \tag{12}$$

Another important factor in our analysis is the load of query requests imposed by other nodes on each super-peer node. Assuming that in a deterministic query search case, the overall load on each node is $1_{Byte/sec}$, then in a probabilistic case, the overall load will be $1_{Byte/sec}$. The reason is that in the probabilistic search case, each node sees a fraction F of the requests generated by each peer. We anticipate that the idea of caching the local indices, proposed by us, causes the value of F to tend to very low values, thus resulting in a reduction in the processing load of each peer.

Now, we provide the numerical results of the above analytical model. We assume d=5. For each node j, the out-degree, namely k_j, comes from a power-law distribution with $\alpha = 1.5$. Figure 5 shows the variation of P_f and F versus p. This figure assumes a fixed message broadcasting probability, i.e., $p_i = p$ for i=2,...,d. Also the probability of caching in the DCs at each level is assumed to be $q_2 = 0.5$, $q_3 = 0.25$, $q_4 = 0.1$, and $q_5 = 0.01$. As the broadcasting probability p increases, the probability that a directory entry for the resource is found increases and exceeds 0.97 for a value of equal to 0.7. At that point, only 12% of the super-peers would participate in the search (i.e., F= 0.12).

Figure 6 shows the variation of F versus p_f. The assumptions used in this figure are the same as those of Fig. 5. It can be seen from Fig. 5 that by adjusting the broadcasting probability, one can find the probability that the resource is found. Given this, one can tune the fraction of participating nodes from Fig. 6.

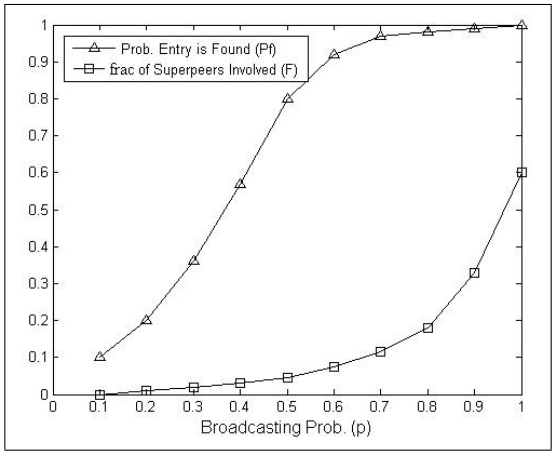


Figure 5: Probability of finding entry and fraction of participating super-peers vs. different broadcasting probabilities

4 Experimental Validation

We have performed a large number of experiments to validate the effectiveness of our proximity-aware distributed caching scheme. We evaluated the performance of the system with a file-sharing application based on several metrics. These metrics include fraction of involving super-peers in the query search, probability of finding an entry in DCs, overall cache miss ratio, average number of hops to perform query requests, and system load. All of the aforementioned metrics, except system load, are already defined in the previous sections. The *load* metric is defined as the amount of work an entity must do per unit of time. It is measured in terms of two resource types: incoming bandwidth, and outgoing bandwidth. Since the availability of the incoming and the outgoing bandwidths is often asymmetric, we have treated them as separate resources. Also, due to heterogeneity of the system, it is useful to study the aggregate load, i.e., the sum of the loads concerning to the all nodes in the system. All of the results are averaged over 10 runs of experiments and have been come up with 95% confidence intervals. We followed the general routine devised in [15] for the efficient design of the P2P network. So, as the first step, we had to

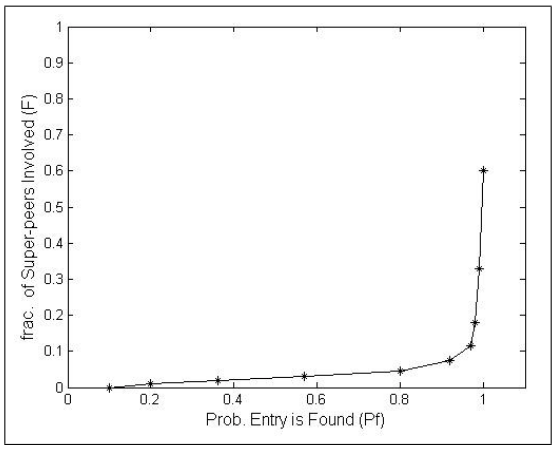


Figure 6: Fraction of super-peers involved vs. probability of finding entry

generate an instance topology based on a power-law distribution. We used the PLOD algorithm presented in [16] to generate the power-law topology for the network. The second step was calculating the expected cost of actions. Among three "macro actions", i.e., query, join, and update, which exist in a cost model [15], we have restricted our attention to the query operations. Each of these actions is composed of smaller "atomic" actions for which the costs are given in [15]. In terms of bandwidth, the cost of an action is the number of bytes being transferred. We used the specified size of the messages for Gnutella protocol in such a way that is defined in [15]. For example, the query messages in Gnutella include a 22-byte Gnutella header, a 2 byte field for flags, and a null-terminated query string. The total size of a query message, including Ethernet and TCP/IP headers, is therefore 82 plus the query string length. Some values, such as the size of a metadata record are not specified by the protocol, rather are functions of the type of the data which is being shared. The values which are used from [15] are listed in Table 3.

Table 3: Gnutella bandwidth costs for atomic actions [15]

Atomic Action	Bandwidth Cost (Bytes)
Send Query	82 + query length
Recv. Query	82 + query length
Send Response	80+28 #addresses+76 #results
Recv Response	80+28 #addresses+76 #results

To determine the number of results which are returned to a super-peer r, we have used the query model developed in [17] which is applicable to super-peer file-sharing systems. The number of files in the super-peer's index depends on the particular generated instance topology I. We have used the so called query model to determine the expected number of returned results, i.e. $E[N_r|I]$. Since the cost of the query is a linear function of $[N_r|I]$ and also since the load is a linear function of the cost of the queries, we can use these expected values to calculate the expected load of the system [15].

In the third step, we must calculate the system load using the actions. For a given query originating from node s and terminating in node r we can calculate the expected cost, namely C_{sr}. Then, we need to know the rate at which the query action occurs. The default value for the query rate is 9.26×10^{-3} which is taken from the general statistics provided by [15] (see Table 4). The query requests in our experiments have been generated by a workload generator. The parameters of the workload generator can be set up to produce uniform or non-uniform distributions. Considering the cost and the rate of each query action, we can now calculate the expected load which is incurred by node r for the given network instance I as follows

$$E[M_r|I] = \sum_{s \in Network} E[C_{sr}|I].E[F_s] \tag{13}$$

Where, F_s is the number of the queries submitted by the node s in the time unit, and $E[F_s]$ is simply the query rate per user.

Let us define Q as the set of all super-peer nodes. Then, the expected load of all such nodes, namely M_Q , is defined as follows

$$E[M_Q|I] = \frac{\sum_{n \in Q} E[M_n|I]}{|Q|} \tag{14}$$

Also, the aggregate load is defined as follows

$$E[\overline{M}|I] = \sum_{n \in Network} E[M_n|I] \tag{15}$$

We ran the simulation over several topology instances and averaged E[M|I] over these trials to calculate E[E[M|I]]=E[M] . We came up with 95% confidence intervals for E[M|I]. The settings used in our experiments are listed in Table 4. In our experiments, the network size was fixed at 10000 nodes. As mentioned before, the generated network has a power-law topology with the average out-degree of 3.1 and TTL=7. These parameters reflect Gnutella topology specifications which have been used by many researchers so far. For each pair of the super-peers *(s,r)*, the latency distance *lat(s,r)* was generated using a normal distribution with an average $\mu = 250_{ms}$ and a variance $\delta = 0.1$ [12]. Then, to find the pair-wise latency estimation, namely *est(s,r)* , we ran the VIVALDI method over the generated topology.

Table 4: Experimental settings

Name	Default	Description
Graph type	Power-law	The type of network, which may be strongly connected or power-law
Graph size	10000	The number of peers in the network
Cluster size	10	The number of nodes per cluster
Avg. out-degree	3.1	The average out-degree of a super-peer
TTL	7	The time-to-live of a query message
Query rate	9.26×10^{-3}	The expected number of queries per user per second

In order to be able to compare the results with the previous works, we chose a cache size per super-peer equal to 1% of the total number of the resources managed by the all super-peers. In a distributed system with highly variant reference patterns, it is better to use frequency-based replacement policies. The frequency-based policy takes into account the frequency information which indicates the popularity of an object. The Least-Frequency-Used (LFU) is a typical frequency-based policy which has been proved to be an efficient policy [18]. In LFU, the decision to replace an object from the cache is made by the frequency of the references to that object. All objects in the cache maintain the reference count and the object with the smallest reference count will be replaced. The criterion for replacing an object from the cache is computed as follows

$$Cost_{Object} = Frequency_{Object} \times Recency_{Object} \tag{16}$$

Where, $Frequency_{Object}$ and $Recency_{Object}$ denote the *access frequency* and the *elapsed time from recent access*, respectively. If the cache has enough room, LFU will store the new object in it. Otherwise, LFU selects a candidate object which has the lowest $Cost_{Object}$ value among all cached objects. Then, LFU will replace the candidate object by the new object if the of the new object is higher than that of the candidate object. Otherwise, no replacement occurs.

Figure 7 shows the experimental results concerning the effect of the resource replication on the fraction of participating super-peers, namely F, and the probability of finding objects, namely P_f, versus various broadcasting probabilities. It can be seen from the figure that P_f attains high values for much smaller values of p. The experimental result in Fig. 7 shows a trend similar to what the analytical study provides in Fig. 5. By adjusting the broadcasting probability, one can tune the probability of finding the resource. In the case of using resource replication, P_f achieves larger values in comparison with the case in which the resource replication is not used. In contrast to P_f, the metric F achieves smaller values in the case of using the resource replication in comparison with the case in which the resource replication is not used. Its cause lies in the fact that in the case of using the resource replication method, some intermediate nodes replicate the queries in their local disks (cache the queries into their LIT); leading to a decrease in the LITs miss ratio, thus resulting an increase in the probability of finding the queries. Such nodes do not need to propagate the *QuerySearch* message to other super-peers anymore.

Fig. 8 shows the effect of using the resource replication on the cache miss ratio as a function of the broadcasting probability p. In the both cases of Fig. 8, the cache miss ratio decreases by an increase in p. Its cause lies in the fact that when p increases, more super-peers participate in the search; hence it is more likely that the resource is

found by more than one super-peer. So, more DCs of the intermediate nodes in the reverse path to the source will be aware of the resource; leading to a decrease in the DCs miss ratio. The use of the resource replication decreases the miss ratio compared to the case in which the resource replication is not used. However, the amount of the reduction in the miss ratio is not remarkable in both cases for the values of p greater than 0.8. At this point, the use of resource replication method yields a miss ratio of 0.72; giving 20% improvement over the 0.9 miss ratio when no resource replication is used.

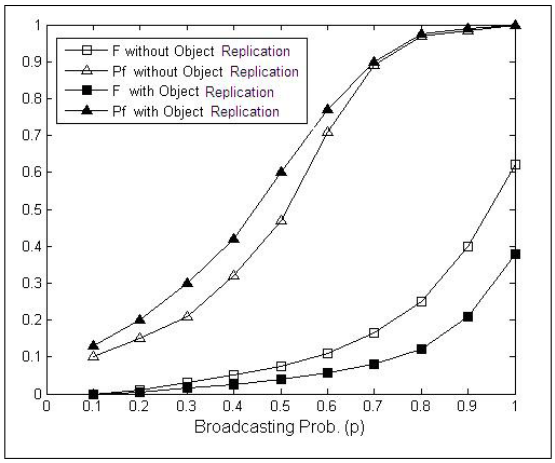


Figure 7: The effect of resource replication on the fraction of participating peers and the probability of finding objects for various broadcasting probabilities

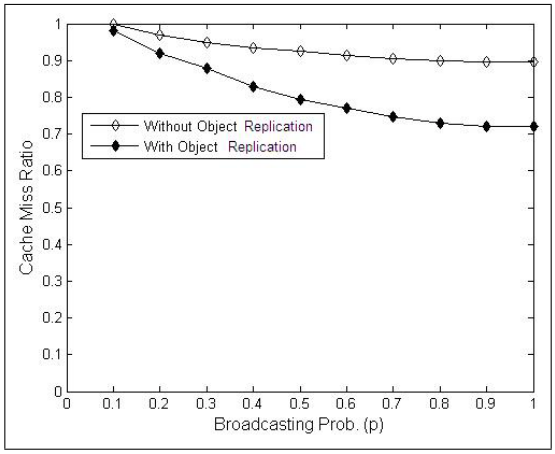


Figure 8: The effect of resource replication on overall cache miss ratio for various broadcasting probabilities

Figure 9 shows the average number of the required hops to find the resource, namely H, which is normalized by the total number of super-peers (except the original source). The figure shows the effect of the resource replication method in various broadcasting probabilities. It can be seen in both curves of Fig. 9 that the average number of hops initially increases until reaches to a maximum point and then begins to decrease. A higher broadcasting probability means that the super-peers who are located further away from the original source are contacted and the resource tends to be found further away from the original source. As p continues to increase, the increased values of hit ratio concerning to intermediate DCs allow the resource to be found in locations where are closer to the original source; hence decreasing the value of H. It is clear from Fig. 9 that the use of resource replication reduces the number of hops needed to find the resource. For example, in a reasonable practical point of broadcasting probability, such as 0.7, it yields a 31% improvement, whereas the hop ratio decreases from 0.08 to 0.055.

Figure 10 shows the effect of resource replication on the total required bandwidth of the system, i.e. the required incoming and outgoing bandwidth of super-peers, for various broadcasting probabilities. By increasing

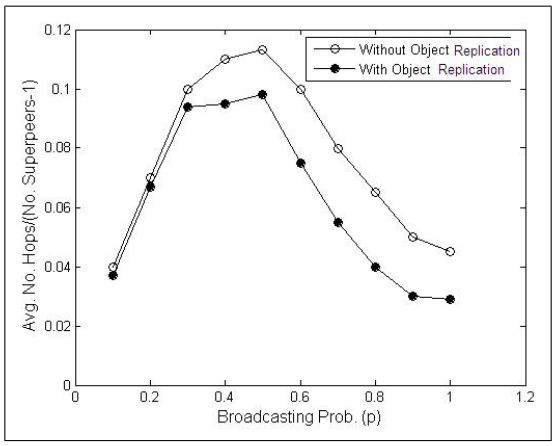


Figure 9: The effect of resource replication on hop ratio for various broadcasting probabilities

the broadcasting probability, some additional costs are imposed to the system. The most important costs include the cost of sending queries to each super-peer, a startup cost for each super-peer as they process the query, and the overhead of additional packet headers for individual query responses. Some of these factors are mentioned in the literature by prior researchers. Interested readers can find useful hints in [15]. The upper curve in Fig. 10 shows the required bandwidth in the absence of resource replication. In this case, as the broadcasting probability p increases, the required bandwidth of super-peers increases and reaches to 7.7 108 bps for a value of equal to 0.8. From this point forward, the growing of bandwidth occurs more slightly until reaches to 7.9×10^8 bps at the value of p equal to 1. The lower curve in Fig. 10 shows an improvement in the required bandwidth in the presence of resource replication. In this case, the required bandwidth decreases to 6.6×10^8 bps for a value of p equal to 0.8, resulting in a 14% improvement in comparison with the same point in the upper curve.

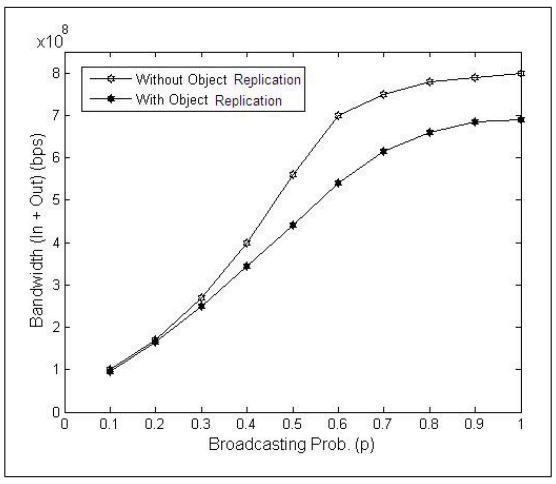


Figure 10: The effect of resource replication on total required bandwidth for various broadcasting probabilities

5 Related Works

Unstructured architectures are divided into three sub-classes [1]: 1) hybrid decentralized, 2) purely decentralized, and 3) partially centralized. In a hybrid decentralized P2P system, all peers connect to a central directory server that maintains a table for their IP address, connection bandwidth and other information, and another table that keeps the list of files that each peer holds. Upon receiving a request from a peer, the server searches for any matches in its index table and returns a list of peers who hold the matching file. Then a direct connection is

established between the originating peer and the peers who hold the requested files to download them. Although the implementation of hybrid decentralized systems is easy, they suffer from vulnerabilities and attacks as well as scalability problems. Napster [19] is an example of such systems. Gnutella [3] is a well-known example of purely decentralized system.

A significant research toward proximity-aware resource location services in typical Gnutella-based unstructured P2P system has been done in [7, 8]. The proximity metric in this works is the TTL of the query messages. Forwarding the queries is done with a fixed probability. When a query message is reached to a peer, its TTL is decremented. The forwarding of the query messages will be stopped if its TTL is reached to zero. The analytical study on the impact of the proximity-aware methodology for making the P2P streaming more scalable has been presented in [9]. They have proposed a geometric model which maps the network topology from the space of discrete graph to the continuous geometric domain, meanwhile capturing the power-law property of the Internet. They found that, although random peer selection methods can maximally save the server resources, they introduce the maximum load to the network.

In systems where availability is not guaranteed, such as Gnutella [3], resource location techniques can afford to have loose guarantees [20]. Current search techniques in "loosely controlled" P2P systems are rather inefficient because they impose a heavy load on system as well as high response times. The main motivation for many researches in the area of P2P systems was early "loosely controlled" systems such as Gnutella, Freenet [14], Napster [19], and Morpheus [21]. Other resource location techniques for "loosely guaranteed" systems are mentioned in [20]. In the other proposed techniques which mentioned in [20] each node maintains "hints" as to which nodes contain data that answer certain queries, and route messages via local decisions based on these hints. This idea itself is similar to the philosophy of hints which is used by Menasce et al. in [7]. CAN [2] is an example of systems with "strong guarantee" that employs search techniques. These systems can locate an object by its global identifier within a limited number of hops. It is concluded from the literature that selecting a resource locating methodology depends on the type of system which is planned. A complete literature survey relevant to search techniques is collected in [20] which interested readers can refer to it for more detail.

The resource locating techniques for partially centralized P2P networks are addressed in [15, 20]. They also evaluate some proposed query search broadcasting policies using Gnutella system and compare their performance with each other. The first query search policy evaluated by Yang et al. on Gnutella is called iterative deepening in which, a query is sent iteratively to more nodes until the query is answered. The second proposed technique, directed Breath First Search (DBFS) technique, forwards a limited set of nodes selected to maximize the probability that the query is answered. Their experimental analysis shows that if nodes are allowed to answer queries on behalf of other nodes, then the number of nodes that process a query will be reduced without decreasing the number of results. This very nice conclusion has formed our original motivation for designing distributed P2P caching protocol based on super-peers. The last technique which is investigated by them is called local indices technique in which, nodes maintain simple indices over other client's data. Queries are then processed by a smaller set of nodes. This is also quite similar to the concept of the super-peer nodes which we have adopted in our proposal.

The performance of hybrid P2P systems such as Napster is investigated by Yang and Garcia-Mollina in [17]. Morpheus [21] is another hybrid P2P system whose architecture is similar to Gnutella. Upon joining a new peer to the system, P2P network contacts a centralized server which then directs it to a super-peer. The authors of [17] study the behavior and performance of hybrid P2P systems and develop a probabilistic model to capture the query characteristic of such systems.

6 Conclusions

In this paper we have targeted the scalable proximity-aware location service for P2P systems. The proposed protocol provides a scalable distributed caching mechanism to find the peers who manage a given resource. The proposed mechanism enhances the mechanisms which have been proposed in previous researches by replicating objects based on latency distance metric, resulting in less aggregate load of the system. The simulation results showed that the use of probabilistic resource discovery service in P2P systems combined with latency-aware probabilistic resource replication, improves the overall performance of the system in terms of aggregated load, throughput, response time, number of hops, and number of contributing peers in the search process. Using the proposed mechanism yields at most 20% improvement in miss ratio in comparison with the case in which no resource replication is used. Also, in reasonable practical points of broadcasting probability, it yields about 30% reduction in hop ratio as well as 14% reduction in required bandwidth of the system.

Bibliography

[1] S. Androutsellis-Theotokis, D. Spinellis, *A Survey of Peer-to-Peer Content Distribution Technologies*, ACM Computing Surveys, vol. 36, no. 4, pp. 335-371, 2004

[2] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker, *A Scalable Content Addressable Network*, Proc. ACM Sigcomm, August 2001.

[3] M. Ripeanu, I. Foster, A. Iamnitchi, *Mapping the Gnutella Network: Properties of Large-Scale Peer-to-Peer Systems and Implications for System Design*, IEEE Internet Computing, 6(1), February 2002.

[4] http://www.kazaa.com.

[5] Q. Lv, P. Cao, E. Cohen, K. Li, and S. Shenker, *Search and Replication in Unstructured Peer-to-Peer Networks*, the 16th ACM International Conference on Supercomputing (ICS'02). New York, NY., 2002.

[6] A. Crespo, H. Garcia-Molina, *Routing Indices for Peer-to-Peer Systems*, Proc. of Int. Conf. on Distributed Computing Systems, Vienna, Austria, 2002.

[7] D.A. Menascé, L. Kanchanapalli, *Probabilistic Scalable P2P Resource Location Services*, ACM Sigmetrics Performance Evaluation Rev., Volume 30, No. 2, pp. 48-58, 2002.

[8] D. Menascé, *Scalable P2P Search*, IEEE Internet Computing, Volume 7, No. 2, March/April 2003.

[9] L. Dai, Y. Cao, Y. Cui and Y. Xue, *On Scalability of Proximity-Aware Peer-to-Peer Streaming*, in Computer Communications, Elsevier, vol. 32, no 1, pp. 144-153, 2009.

[10] Y. Zhu, B. Li., *Overlay Networks with Linear Capacity Constraints*, IEEE Transactions on Parallel and Distributed Systems, 19 (2), pp. 159-173, February 2008.

[11] Y. Zhu, B. Li, K. Q. Pu., *Dynamic Multicast in Overlay Networks with Linear Capacity Constraints*, IEEE Transactions on Parallel and Distributed Systems, Vol. 20, No. 7, pp. 925-939, 2009.

[12] G.P. Jesi, A. Montresor, O. Babaoglu, *Proximity-Aware Superpeer Overlay Topologies*, IEEE Transactions on Network and Service Management, September 2007.

[13] F. Dabek, R. Cox, F. Kaashoek, and R. Morris., *VIVALDI: A Decentralized Network Coordinate System*, The SIGCOMM '04, Portland, Oregon, August 2004.

[14] I. Clarke, S. G. Miller, T. W. Hong, O. Sandberg, and B. Wiley, *Protecting Free Expression Online with Freenet*, IEEE Internet Computing , Volume 5, No. 1, pp. 40-49, 2002.

[15] B. Yang, H. Garcia-Molina, *Designing a Super-Peer Network*, Proc. Int'l Conf. Data Eng. (ICDE), pp. 49-63, Mar. 2003.

[16] C. Palmer, J. Steffan, *Generating network topologies that obey power laws*, The GLOBECOM 2000, November 2000.

[17] B. Yang, H. Garcia-Molina, *Comparing Hybrid Peer-to-Peer Systems*, Proc. 27th Int. Conf. on Very Large Data Bases, Rome, 2001.

[18] J.W. Song, K.S. Park, S.B. Yang, *An Effective Cooperative Cache Replacement Policy for Mobile P2P Environments*, In proceeding of IEEE International Conference on Hybrid Information Technology (ICHIT'06), Korea, Vol. 2, pp. 24-30, 2006.

[19] http://www.napster.com.

[20] B. Yang, H. Garcia-Molina, *Improving Search in Peer-to-Peer Networks*, The 22nd International Conference on Distributed Computing Systems (ICDCS'02), Vienna, Austria, 2002.

[21] http://www.morpheus-os.com.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 4, pp. 432-446

How to Write a Good Paper in Computer Science and How Will It Be Measured by ISI Web of Knowledge

R. Andonie, I. Dzitac

Răzvan Andonie
Department of Computer Science
Central Washington University, USA
and
Department of Electronics and Computers
Transylvania University of Braşov, Romania
E-mail: andonie@cwu.edu

Ioan Dzitac
Department of Mathematics-Informatics
Aurel Vlaicu University of Arad
310330 Arad, Romania
and
Cercetare Dezvoltare Agora
Piatţa Tineretului, 8, Oradea, Romania
Email: ioan.dzitac@uav.ro

Abstract: The academic world has come to place enormous weight on bibliometric measures to assess the value of scientific publications. Our paper has two major goals. First, we discuss the limits of numerical assessment tools as applied to computer science publications. Second, we give guidelines on how to write a good paper, where to submit the manuscript, and how to deal with the reviewing process. We report our experience as editors of International Journal of Computers Communications & Control (IJCCC). We analyze two important aspects of publishing: plagiarism and peer reviewing. As an example, we discuss the promotion assessment criteria used in the Romanian academic system. We express openly our concerns about how our work is evaluated, especially by the existent bibliometric products. Our conclusion is that we should combine bibliometric measures with human interpretation.
Keywords: scientific publication, publication assessment, plagiarism, reviewing, bibliometric indices.

1 Introduction

Faculty work generally falls into three categories: research, teaching, and service. Assessment protocols have considered, to a varied extent, scholarly activities performed in each of these areas. Faculty assessment is conducted for purposes of reappointment, promotion, the awarding of tenure, and professional development.

During the last decades, a societal focus on the work of university faculty as a measure of return on the public's investment in higher education stimulated a reevaluation of how faculty performance ought to be measured and assessed. The development of workable assessment systems is difficult largely due to the fact that the value of assessment is often controversial:

- Assessment methods are defined differently from discipline to discipline.
- Assessment methods depend on the communication of standards upon which judgments of quality will be based and acceptable mechanisms for documenting faculty work.
- As members of a profession, faculty reserve the right to be the sole judges of the quality of the work performance of those claiming membership among their ranks.

At different levels, non-faculty administrators are also involved in the assessment process of faculty. There are cases when non-faculty are judging the scientific activity of faculty solely based on criteria like number of publications and impact factors, without having the expertise to pertinently judge these publications.

This creates a possible conflict and, in many cases, we can observe tensions in the faculty-administrator relationship [1]. The conflict starts from a communication failure between the two groups. Basically,

faculty and administrators, share the same goals. Beside the psychological aspect (faculty do not like to be judged by non-academics), another cause of tension is the difficult question "How do we measure academic performance?"

Research performance is typically measured in terms of productivity, relying largely on the use of quantitative measures such as the number of publications, value of grants, or other creative works produced over a specified period of time. Many universities use indexing systems, like Thompson Scientific, as a main assessment tool for publications. But how much can we trust such a numerical criterion? Is it enough to count the number of citations of your paper to judge its value?

Our paper has two major goals. First, we focus on assessment techniques for scientific publications. We discuss the limits of numerical assessment tools. We particularly analyze the specific aspects of computer science (CS) publications knowing that cross-disciplinary comparisons should be generally avoided.

Second, we give guidelines on how to write a good paper, where to submit the manuscript, and how to deal with the reviewing process. We report our experience as editors of IJCCC. From this perspective, we also analyze two important aspects of publishing: plagiarism and peer reviewing.

We illustrate with the promotion assessment criteria used in the Romanian academic system. Finally, we discuss the "publish or perish practice" from the perspective of the current publication assessment techniques.

2 Assessment of CS scientific publications

Books, which some disciplines do not consider important scientific contributions, can be a primary vehicle in CS. We discuss here only dissemination of scientific results by conference proceedings and journals and we start with an important statement: *The order in which a CS publication lists authors is generally not significant. In the absence of specific indications, it should not serve as a factor in researcher evaluation.*

In the CS publication culture, prestigious conferences are a favorite tool for presenting original research, unlike disciplines where the prestige goes to journals and conferences are for raw initial results. Acceptance rates at very selective CS conferences are between 13% and 20%. Can we tell from the acceptance rate alone how good a conference is? The answer is negative. For example, the International Joint Conference on Neural Networks (IJCNN) is a much better conference than shown by its 2008 acceptance rate, which was 58%. As a regular reviewer of IJCNN, one of the authors of this paper (R.A.) considers that about 80% of the submitted papers are at least acceptable. *We cannot tell how selective a conference (or a journal) is only by the percentage of papers it accepts because far fewer bad papers are submitted to the best conferences and journals.*

CS journals have their role, often to publish deeper versions of papers already presented at conferences. While many researchers use this opportunity, others have a successful career based largely on conference papers.

There is an increasing tendency to numerically measure the quality of a paper. The starting point would be data from citation databases, such as Institute for Scientific Information's Web of Science, that can be analyzed to determine the popularity and impact of specific articles, authors, and publications. In ISI Web of Science citation metrics information is available only when records on the publication list have been added from the Web of Science. Usually metrics are "Sum of the Times Cited", "Average Citations per Item" and "h-index". According to Hirsch [2], h is the number of articles greater than h that have at least h citations. An h-index of 20 means that there are 20 items that have 20 citations or more. The accuracy of these metrics largely depend on the accuracy of the ISI database.

The Journal Citation Reports (JCR), published annually by Thomson Reuters, provides quantitative tools for ranking journals in accordance to statistical information from citation data. Ranking is performed, for instance, by the *impact factor*, which is a measure of the frequency with which the "average

article" in a journal has been cited in a particular year or period. The annual JCR impact factor is a ratio between citations and recent citable items published. Only citations from papers indexed by ISI Web of Science are considered.

Thus, the impact factor of a journal is calculated by dividing the number of current year citations to the source items published in that journal during the previous two years. The impact factor $JCR(J,Y)$ of journal J in year Y, is

$$c(Y;Y-2,Y-1)/p(Y-2,Y-1),$$

where $p(Y-2,Y-1)$ is the number of articles published in journal J in the previous two years ($Y-1$ and $Y-2$), and $c(Y;Y-2,Y-1)$ is the number of citations in year Y of papers published during the previous two years in journal J.

Publication quality is just one aspect of research quality, impact is one aspect of publication quality, and the number of citations is one aspect of impact. Citation counts rely on databases such as ISI, CiteSeer, ACM Digital Library, Scopus, and Google Scholar. They, too, have limitations. An issue of concern to computer scientists is the tendency to use publication databases that do not adequately cover CS, such as Thomson Scientific's ISI Web of Science. The principal problem is what ISI counts [3]. The results make Niklaus Wirth, Turing Award winner, appear for minor papers from indexed publications, not his seminal 1970 Pascal report. As another example, Knuth's milestone book series does not even figure. Other evidences of ISI's shortcomings for CS are [3]:

- ISI's internal coverage (i.e., percentage of citations of a publication in the same database) is over 80% for physics or chemistry, but only 38% for CS. Therefore, we should not make cross-disciplinary comparisons based on number of citations.
- ISI does not index many top conferences (for instance, The International Conference on Software Engineering (ICSE), the top conference in the field) but indexes SIGPLAN Notices, an unrefereed publication.
- ISI's "highly cited researchers" list includes many prestigious computer scientists but leaves out such iconic names as Wirth, Parnas, Knuth and all the ten 2000-2006 Turing Award winners except one.
- Any evaluation criterion, especially quantitative, must be based on clear, published criteria. The methods by which ISI selects documents and citations are not published or subject to debate.

The problem for computer scientists is that assessment relies on often inappropriate and occasionally outlandish criteria. Evaluation criteria, like ISI's impact factor or conference acceptance rates are flawed. Assessment criteria must themselves undergo assessment and revision. We should at least try to base it on metrics acceptable to the profession [3]: "Publication counts only assess activity. Giving them any other value encourages "write-only" journals, speakers-only conferences, and Stakhanovist research profiles favoring quantity over quality."

3 Where to publish your work: conference vs journal

The first thing to start your research is to know what the major journals and conferences in that field are. The rule of thumb is to read "good" papers and submit your papers to "good places". How to recognize a good journal or conference? It is quite easy if you already went through the reviewing process of that publication: a good journal/conference tends to have rigorous review process. If you are a graduate student, work with your mentors to understand what constitutes good versus bad conference/journal.

When ranking conferences, you should look at the following factors: acceptance rate, review process, program committee, who the publisher of the proceedings is, and which database is indexing the published proceedings.

This is an example of a good conference (see T. N. Vijaykumar [14]):

> The ACM/IEEE International Symposium on Computer Architecture (ISCA) is the top forum for architecture and has been so since 1975. ISCA papers are 10-12 pages in length with detailed results, and go through around 5-6 double-blind reviews by the top experts on the topic. The acceptance rate is 15-20%, decided by a National Science Foundation (NSF) - panel-style, 20-person program committee. ISCA takes only 30-35 papers a year (there are no short papers, no posters).

For ranking journals, we have to look at the JCR impact factor, publishing house, and editors. The ISI ranking system is based on the JCR impact factor (see Fig. 1).

UPDATE MARKED LIST

Ranking is based on your journal and sort selections.

Mark	Rank	Abbreviated Journal Title (linked to journal information)	ISSN	JCR Data						Eigenfactor™ Metrics	
				Total Cites	Impact Factor	5-Year Impact Factor	Immediacy Index	Articles	Cited Half-life	Eigenfactor™ Score	Article Influence™ Score
☐	41	ASIAN J CONTROL	1561-8625	359	0.600	0.686	0.237	76	4.7	0.00093	0.156
☐	42	ASSEMBLY AUTOM	0144-5154	212	0.562	0.526	0.075	40	5.4	0.00058	0.130
☐	43	IMA J MATH CONTROL I	0265-0754	324	0.535		0.037	27	8.6	0.00101	
☐	44	INF TECHNOL CONTROL	1392-124X	54	0.495		0.054	37		0.00018	
☐	45	P I MECH ENG I-J SYS	0959-6518	247	0.447	0.414	0.033	90	5.6	0.00079	0.120
☐	46	MODEL IDENT CONTROL	0332-7353	110	0.435	0.648	0.000	6	7.9	0.00030	0.270
☐	47	AT-AUTOM	0178-2312	176	0.378		0.050	60	5.3	0.00017	
☐	47	CONTROL CYBERN	0324-8569	415	0.378	0.560	0.000	14	9.0	0.00128	0.256
☐	49	J DYN CONTROL SYST	1079-2724	201	0.377	0.565	0.080	25	7.3	0.00083	0.304
☐	50	INT J COMPUT COMMUN	1841-9836	95	0.373		0.205	39		0.00011	
☐	51	INTELL AUTOM SOFT CO	1079-8587	102	0.349	0.301	0.136	44	6.2	0.00019	0.061
☐	52	T I MEAS CONTROL	0142-3312	194	0.340	0.586	0.000	29	6.8	0.00052	0.194
☐	53	INT J ROBOT AUTOM	0826-8185	144	0.339	0.438	0.029	35	7.1	0.00041	0.133
☐	54	MEAS CONTROL-UK	0020-2940	168	0.329	0.392	0.029	34	6.7	0.00039	0.100
☐	55	REV IBEROAM AUTOM IN	1697-7912	25	0.291		0.026	39		0.00012	
☐	56	J SYST ENG ELECTRON	1004-4132	195	0.269		0.000	196	3.1	0.00038	
☐	57	AUTOMAT REM CONTR+	0005-1179	984	0.251	0.238	0.121	165	>10.0	0.00205	0.118
☐	58	COMPUT CONTROL ENG	0956-3385	177	0.158	0.325		0	>10.0	0.00025	0.078
☐	59	CONTROL ENG	0010-8049	92	0.024	0.038	0.042	95		0.00024	0.021

Figure 1: ISI Journal Ranking: IJCCC has impact factor 0.373.

For example, let us compute the IJCCC JCR 2009 impact factor. We have: 34 items published in IJCCC (in 4 regular issues) in 2007; 35+ 89 = 124 items published in IJCCC (in 4 regular issues + 1 supplementary issue) in 2008. The total number of articles published in 2007 and 2008 in IJCCC is $p(2007, 2008) = 34 + 124 = 158$. In 2009, there are $c(2009; 2007, 2008) = 22 + 37 = 59$ citations to items published in 2007 and 2008. Hence, $JCR(IJCCC, 2009) = c(2009; 2007, 2008)/p(2007, 2008) = 59/158 = 0.373$.

IJCCC is a new journal, founded in 2006. Authors use different journal title abbreviations, and this makes journal identification by ISI problematic. In addition to this, since the supplementary 2008 issue contains the ICCCC 2008 proceedings, many citations appear as "Proceedings of ICCCC 2008", without mentioning the journal. These are two reasons why the ISI Web of Science database contains incorrect IJCCC entries, which influences the impact factor of our journal. We recognize here the traditional "garbage in - garbage out" problem.

A solution would be to use for each journal its unique ISSN. This is certainly not very easy, because all indexing systems have to change, and authors may have to include the ISSN's in their list of publications. However, we think that the effort is worth, since this would make bibliometric indicators more precise.

Here are examples of important journals and conferences, for different CS domains:

- Database: IEEE Trans on Knowledge and Data Engineering, ACM Trans on Database Systems, Int'l Conf on VLDB.
- Software Engineering: IEEE Trans on Software Engineering, ACM Trans on Software Eng. and Methodology, IEEE Int'l Conf on Software Engineering.
- Computer Networks: IEEE/ACM Trans on Networking, IEEE INFOCOM, ACM Mobicom.
- Parallel/Distributed Systems: IEEE Trans on Parallel and Distributed Systems, ACM Trans on Computer Systems, ICDCS, IPDPS.
- Neural Networks: IEEE Trans. on Neural Networks, Neural Computation, NIPS, IJCNN, ICANN, IWANN, ESANN.

Should we submit to a journal or conference? In the CS context, this question deserves a discussion.

3.1 Why prefer a conference

According to Patterson *et al.* [4], in CS, conference publication is preferred to journal publication, at least for experimentalists. This was the recommendation (a memo) of the Computer Research Association (CRA) in 1999. The CRA memo asserts that conference publication is superior to journal publication in computer science. According to the memo, the typical conference submission receives four to five evaluations, whereas the typical journal submission receives only two to three evaluations.

Computing researchers are right to view conferences as an important archival venue and use acceptance rate as an indicator of future impact. Papers in highly selective conferences, with acceptance rates of 30% or less, should continue to be treated as first-class research contributions with impact comparable to, or better than, journal papers [5]. This distinguishes computer science from other academic fields where only journal publication carries real weight. There are two main reasons to publish in the proceedings of selective conferences:

- Conferences are more timely than journals.
- Conferences have higher standards of novelty. Journals often only require 20-30% of the material to be new, compared to an earlier conference version.

Conference selectivity serves two purposes: pick the best submitted papers and signal prospective authors and readers about conference quality. Is there a connection between conference acceptance rate and impact factor, where impact is measured by the number of citations received? The answer is positive, up to some threshold. Adopting the right selectivity level helps attract better submissions and more citations. Chen and Konstan [5] found, with respect to ACM-wide data, that acceptance rates of 15-20% seem optimal for generating the highest number of future citations for both the proceedings as a whole and the top papers submitted. Conferences rejecting 85% or more of their submissions risk discouraging overall submissions and inadvertently filtering out high-impact research.

3.2 Why prefer a journal

Many universities evaluate faculty on the basis of journal publications because, in most scientific fields, journals have higher standards than conferences. Journals may have longer page limits and journal reviews tend to be more detailed. Many times, conference committees enlist inexperienced graduate students as reviewers of papers in order to meet the quota for reviews. Because conference papers are

limited in length, and because a large number of papers must be reviewed within a short time, the quality of reviews of conference papers is generally low. In contrast, for journals, because there are usually no page limits, authors can explain their ideas completely. Editors can choose qualified reviewers carefully. Reviewers can take adequate time to write thorough reviews.

By polishing a manuscript for journal publication, the author minimizes the number of errors and improves the clarity of the exposition. Thus, journal papers are more likely to be correct and readable than conference papers. Journals are more widely distributed through libraries than conference proceedings, which go out of print quickly. In all disciplines, the criteria for quality include innovation, thoroughness, and clarity, appraised through rigorous peer review. Across disciplines, there are common standards for the evaluation and documentation of publicly presented scholarly work [6]. According to some authors, computer science is not sufficiently different from other engineering disciplines to warrant evaluation on completely different grounds. The evaluation of the scholarship of academic computer scientists should continue to emphasize publications in rigorously refereed, archival scientific journals.

The "conferences vs journal" debate is far from over and was recently relaunched in Communications of the ACM. Studying the metadata of the ACM Digital Library, Chen and Konstan [5] found that papers in low-acceptance-rate conferences have higher impact than in high-acceptance-rate conferences within ACM. Highly selective conferences - those that accept 30% or less of submissions - are cited at a rate comparable to or greater than ACM jounals.

According to Vardi [7], unlike every other academic field, computer science uses conferences rather than journals as the main publication venue. This has led to a great growth in the number of low level conferences. Some call such conferences "refereed conferences" but we all know this is just an attempt to mollify promotion and tenure committees. The reviewing process performed by program committees is done under extreme time and workload pressures, and it does not rise to the level of careful refereeing. Only a small fraction of conference papers are followed up by journal papers.

4 How to write a good paper and how to deal with the editor

Ask two questions before starting: *i)* What is new in your work?, and *ii)* What are you going to write? Emphasize on the originality and significance of your work. Organize your thinking and decide the structure (outlines) of your paper. Stick on your central points throughout the whole paper and remove all unnecessary discussions. There are many good papers on "How to write a good paper", and one of the best known was authored by Robert Day [8]. One could find there some general guidelines which always help:

- Start writing the day you start the research and maintain a good bibliographic database (use BibTeX and LaTeX).
- Think about where to submit early.
- Don't try and prove you are smart and avoid the kitchen sink syndrome.
- Start from an outline.
- Work towards making your paper a pleasure for the reviewer to read.
- Obey the guide to authors.

The structure of a CS paper is not different than the structure of any scientific publication: Title, Abstract, Introduction, Background, Related Work, System Model & Problem Statement, Methods / Solutions, Simulations / Experiments, Conclusion, Acknowledgment, and References. Almost everybody knows this. However, there are some simple rules of thumb which can make life easier.

According to the "Hourglass Model" [9], a paper should start from general, and go through particular back to general (Fig. 2).

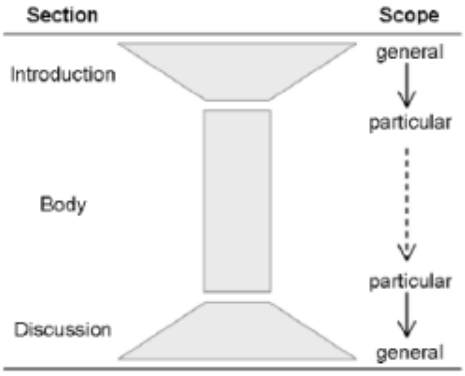


Figure 2: Hourglass Model (Swales [9]).

1. Choose a right title. The title should be very specific, not too broad. The title should be substantially different from others. Avoid general titles, e.g., "Research on data mining", "Contributions to Information Theory", "Some research on job assignment in cluster computing", or "A new framework for distributed computing".

2. Write a concise abstract. An abstract should tell:

- Motivation: Why do we care about the problem and the results?
- Problem statement: What problem is the paper trying to solve and what is the scope of the work?
- Approach: What was done to solve the problem?
- Results: What is the answer to the problem?
- Conclusions: What implications does the answer imply?

A good hint is pack each of these part into one sentence.

3. Organization of your paper.

- Plan your sections and subsections. Use a top-down writing method. Use a sentence to represent the points (paragraphs) in each subsections.
- Writing details: expand a sentence in the sketch into a paragraph.
- Keep a logical flow from section to section, paragraph to paragraph, and sentence to sentence.

4. Introduction: the most difficult part. This is why one of the authors of the present paper (R. A.) prefers to write the Introduction in the final stage and, whenever he writes a paper with students, he prefers to write this section entirely himself.

An Introduction should:

- Establish a territory:
  - bring out the importance of the subject
  - make general statements about the subject
  - present an overview on current research on the subject

- Establish a niche:
  - oppose an existing assumption
  - reveal a research gap
  - formulate a research question
  - continue a tradition, or propose a completely new approach
- Occupy the niche:
  - sketch the intent of the own work
  - outline important characteristics and results of your own work
  - give a brief outlook on the structure of the paper

5. Related work and list of references. Use a proper selection of references. Show your knowledge in the related area. Give credit to other researchers (reviewers are usually chosen from the references). Cite good quality work, particularly when citing your own work, and up to date work. Related work should be organized to serve your topic. Emphasize on the significance and originality of your work.

6. Your conclusions. A research paper should be circular in arguments, i.e., the conclusion should return to the opening, and examine the original purpose in the light of the research presented.

Assuming that you have decided where to submit, and your paper is ready. What is the next step after writing a nice letter to the editor (if it is a journal) with your manuscript submitted electronically? Most probably, your paper will be rejected, or conditionally accepted after a major review. It is almost impossible to have your paper accepted without any modification suggested (except if your name is Donald Knuth or David Patterson!). Even an acceptance "with minor modifications" is rare.

The best scientists get rejected and/or have to make major revisions. It is unreasonable to get defensive, unless it is really called for. You should address EVERY aspect of the reviewers concerns. Make it obvious to the reviewer through the *Summary of Changes* and the revised manuscript itself of the changes you have made. Do not add new science unless it is called for.

A good referee report is immensely valuable, even if it tears your paper apart. Remember, each report was prepared without charge by someone whose time you could not buy. All the errors found are things you can correct before publication. Appreciate referee reports, and make use of them. An author who feels insulted and ignores referee reports wastes an invaluable resource and the referees' time.

Finally, we have to remember what you put in the literature is your scientific legacy after all else is gone.

5 Plagiarism and innovation

Since IJCCC is a young journal, it is reasonable to believe that our review process is less professional than for a top ranked journal like IEEE Transactions, and our journal attracts many authors who are in their early research stages. Such authors are usually under terrible pressure to get something published or not to finish their PhD, or not to be promoted (and possibly lose their jobs). The matter becomes serious for them, and some authors try everything to save their career, including plagiarism.

After receiving your paper, editors and reviewers have to deal with a very unpleasant task of which authors are probably not aware: they have to detect possible plagiarism. Only after your paper passes this preliminary screening it is considered for the true review. As IJCCC editors we have to reject about 8% because of detected plagiarism. Authors of these papers are blacklisted and not considered for future publication. We do not have any statistics about plagiarism frequencies at other publications, but this would certainly be of interest. We may even imagine a "plagiarism world map"! Therefore, we consider important to discuss plagiarism here.

The rules of what constitutes plagiarism and how it should be dealt with are not always clear. According to the IEEE Institute print edition, there are five level of plagiarism:

> "1. Uncredited verbatim copying of a full paper. Results in a violation notice in the later article's bibliographic record and a suspension of the offender's IEEE publication privileges for up to five years. 2. Uncredited verbatim copying of a large portion (up to half) of a paper. Results in a violation notice in the later article's bibliographic record and a suspension of publication privileges for up to five years. 3. Uncredited verbatim copying of individual elements such as sentences, paragraphs, or illustrations. May result in a violation notice in the later article's bibliographic record. In addition, a written apology must be submitted to the original creator to avoid suspension of publication privileges for up to three years. 4. Uncredited improper paraphrasing of pages or paragraphs (by changing a few words or phrases or rearranging the original sentence order). Calls for a written apology to avoid suspension of publication privileges and a possible violation notice in the later article's bibliographic record. 5. Credited verbatim copying of a major portion of a paper without clear delineation of who did or wrote what. Requires a written apology, and to avoid suspension, the document must be corrected."

The guidelines also make recommendations for dealing with repeated offenses.
According to the IJCCC Author Guidelines, submissions to IJCCC must represent original material:

> "Papers are accepted for review with the understanding that the same work has been neither submitted to, nor published in, another journal or conference. If it is determined that a paper has already appeared in anything more than a conference proceeding, or appears in or will appear in any other publication before the editorial process at IJCCC is completed, the paper will be automatically rejected.
>
> Papers previously published in conference proceedings, digests, preprints, or records are eligible for consideration provided that the papers have undergone substantial revision, and that the author informs the IJCCC editor at the time of submission.
>
> Concurrent submission to IJCCC and other publications is viewed as a serious breach of ethics and, if detected, will result in immediate rejection of the submission.
>
> If any portion of your submission has previously appeared in or will appear in a conference proceeding, you should notify us at the time of submitting, make sure that the submission references the conference publication, and supply a copy of the conference version(s) to our office. Please also provide a brief description of the differences between the submitted manuscript and the preliminary version(s).
>
> Editors and reviewers are required to check the submitted manuscript to determine whether a sufficient amount of new material has been added to warrant publication in IJCCC. If you have used your own previously published material as a basis for a new submission, then you are required to cite the previous work(s) and clearly indicate how the new submission offers substantively novel or different contributions beyond those of the previously published work(s). Any manuscript not meeting these criteria will be rejected. Copies of any previously published work affiliated with the new submission must also be included as supportive documentation upon submission."

Whereas plagiarism can more or less measured and there are even software tools available which can help for this, the hardest part to judge as a reviewer is the level of innovation: How much innovation is enough to accept a paper?

According to Patterson *et al.* [4]:

> "When one discovers a fact about nature, it is a contribution per se, no matter how small. Since anyone can create something new (in a synthetic field), that alone does not establish a contribution. Rather, one must show that the creation is better. Accordingly, research in computer science and engineering is largely devoted to establishing the "better" property."

The degree of innovation required depends on the policy of the publication and how selective the conference/journal is. For example, let us illustrate with a good journal. The IEEE Transactions on Neural Networks is ranked 13th overall in terms of impact factor (2.62) among all electrical and electronic engineering journals (206 journals), according to the latest Journal Citation Report (see [10]). The average time between submission and publication (in print) is 18.8 months, which implies that the average time between final acceptance of a paper and publication is approximately 8 months. The conditions to have a paper accepted for IEEE Transactions on Neural Networks are posted in the authors guidelines:

- Full Papers are characterized by novel contributions of archival nature in developing theories and/or innovative applications of neural networks and learning systems. The contribution should not be of incremental nature, but must present a well-founded and conclusive treatment of a problem. Well organized survey of literature on topics of current interest may also be considered.
- Brief Papers report sufficiently interesting new theories and/or developments on previously published work in neural networks and related areas. The contribution should be conclusive and useful.

It is important to read very carefully these guidelines before submitting a paper. Words like "incremental" research are important and should be understood clearly. Editors, like accountants, are serious people and they do not play with words.

According to Qiu [11]:

> "One-third of more than 6,000 surveyed across six top Chinese institutions admitted to plagiarism, falsification or fabrication. Many blamed the culture of *jigong jinli* - seeking quick success and short-term gain - as the top reason for such practices."
>
> "Most academic evaluation in China - from staff employment and job promotion to funding allocation - is carried out by bureaucrats who are not experts in the field in question. When that happens, counting the number of publications, rather than assessing the quality of research, becomes the norm of evaluation."
>
> "To critics such as Rao Yi, dean of the life-science school at Peking University in Beijing, the lack of severe sanctions for fraudsters, even in high-profile cases, also contributes to rampant academic fraud."

We discover the same situation in India [12]:

> "The resulting push to publish, combined with ignorance about what exactly constitutes plagiarism and research misconduct, has led to a rise in such incidents in the last eight to 10 years."
>
> "Meanwhile, the lack of both federal and institutional mechanisms that could detect and punish instances of misconduct have compounded the problem, say some scientists."

Actually, plagiarism appears in all countries, but it is more visible in countries: *i)* with high level of corruption, where plagiarism is not punished, *ii)* where only the number of papers is the measure of success, and *iii)* where plagiarism is not considered a major offense.

As editors and reviewers, we spend sometimes more time with detecting plagiarism than with judging the novelty of a paper.

6 The task of the reviewer

There is an endless stream of research papers submitted to conferences, journals, and other periodicals. Many such publications use impartial, external experts to evaluate papers. This approach is called peer review, and the reviewers are called referees. Refereeing is a public service, one of the professional obligations of a computer science and engineering professional. Unfortunately, referees typically learn to produce referee reports without any formal instruction; they learn by practice [13]. The quality of a publication is also determined by the quality of the reviews. Good publications attract the best reviewers and keep this way, in a positive feedback, a high publication standard. For an acceptance rate of 33%, it is fair to ask each published author to provide at least nine good reviews for submitted papers, assuming that each submitted paper has three reviewers.

Beside detecting plagiarism, editors have to face another administrative problem: they have to find good reviewers. Since IJCCC is less prestigious than an IEEE or ACM journal, it is perhaps less attractive for a good computer scientist to collaborate with us. As IJCCC editors, we have difficulties in motivating and recruiting good reviewers. The name of the editor can help. In our case, many of our international professional friends have accepted to write reviews simply because of our personal relationship. *One rule we try to apply is to let all authors from Romania be reviewed by non-Romanian residents.* The goal is to make our review process unbiased. The most reliable reviewers are experts in their postdoc stage. Senior computer scientists are less willing to meet the review deadlines. Our review process is blind, but not double-blind: reviewers do know the author's name. The simple blind review process is possibly more biased, but it has a big advantage: plagiarism is easier to detect.

Reading a paper as a referee is closer to what a teacher or professor does when grading a paper than what a scientist or engineer does when reading a published work. As a referee you must read the paper carefully and with an open mind, checking and evaluating the material with no presumption as to its quality or accuracy. If you want to be taken seriously as a referee, you must have a middle-of-the-road. A referee who always says "yes" or always says "no" is not helpful.

Don't waste that effort on a detailed critique of a badly flawed paper that can never be made publishable. Finding one or more fatal and uncorrectable flaws excuses the referee from checking all subsequent details. Your report should not be insulting. Don't refer to the author as "idiot" nor to the paper as "trash". Your review should be directed at the paper, not the author. In all cases, the evaluation should be objective and fair. The more psychologically acceptable the review, the more useful it will be.

After comparing the paper to an appropriate standard (not your own standards, which may be high or low), you should be able to put it into one of these categories:

> (1) Major results; very significant (fewer than 1 percent of all papers). (2) Good, solid, interesting work; a definite contribution (fewer than 10 percent). (3) Minor, but positive, contribution to knowledge (perhaps 10-30 percent). (4) Elegant and technically correct but useless. This category includes sophisticated analyses of flying pigs. (5) Neither elegant nor useful, but not actually wrong. (6) Wrong and misleading. (7) So badly written that technical evaluation is impossible.

But what are the standards of a journal or conference? You should compare the paper with the average paper in that specific journal or conference, not with the best or worst. Of course, in some cases the average is too low and needs to be raised by critical refereeing.

As a reviewer, you should be alert to the author who tries to publish the same work in all its various combinations, permutations, and subsets, and to the author who adds the "least publishable unit" of new material to each paper.

7 Case study - the IJCCC reviewing process

From the many hundreds of emails received by us from authors, we have selected some representative ones.

An "excuse letter" for detected plagiarism:

> "Respected Sir,
> I am Mrs. J. from ***. This paper was originally prepared in 2008 during my course work period as a conceptual paper. By the time I was not aware of the act of plagiarism. And this paper was submitted only by me without the knowledge of my supervisor. But, later in the middle of 2009 I came to know that preparing an article in this manner is avoided. Regarding this I sent a mail to the journal office stated that when can it be published. But I came to know that this paper has been sent for review process and you got the result as such.
> So, I request you to forgive me for my activity which I have been done unknowingly and Now I know to prepare the articles which exhibit only my own findings and I am sure that hereafter this type of work will not be done by me.Once again I apologize for my action and Sorry for the inconvenience."

Here are four hilarious submission letters, with their typo and language errors:

> "Dear Sir / Madam
> This is two paper when you received my email just reply me.
> **notes : which time i can recieved the final result.
> Thank you very much"

> "Hi,Dear professor i have send a new paper for your's journal
> . . . , Msc. Faculty member and Head Research group"

> "Dear Sir
> Pl find my paper attched to this mail id. . . . "

> "Knowing the importance of your journal, we want to submit to you the advances of our research in the area in order to share them with your readers.
> Hoping to hear from you soon"

Certainly, a nice submission letter is not a sufficient condition for a manuscript to be accepted. But can we expect a good paper from an author who does not know how to write a simple letter?

Here is a nice professional submission letter:

> "Dear Dr. . . . ,
> Please find attached our paper entitled . . . . This is joint work of . . . . I will serve as corresponding author. Please accept it as a candidate for the publication in IJCCC.
> This manuscript is the authors' original work and has not been published nor has it been submitted simultaneously elsewhere. All authors have checked the manuscript and have agreed to the submission.
> Thank you for your consideration.
> Best regards,"

Finally, here is the first part of a good Summary of Changes document addressed to us after a major review:

"Dear Editors:

I am the author of the paper entitled .... I revised my paper according to your suggestions and here is the explanation of the changes:

1. Section 1, paragraph 5, the first sentence is changed from "a fuzzy QoS routing protocol proposed to..." to "we present a Fuzzy controller based QoS Routing Algorithm...".

2. In Abstract, "NS2" is explained as network simulation version 2 explicitly.

... 11. In Simulation section, the nodes are assigned classes randomly and we removed the class distribution item in Table 2 accordingly."

8 Case study - promotion requirements in Romania

In an effort to uniformly regulate promotion requirements in Romanian universities, the Romanian Ministry for Education [15] asks for a minimum number of published papers indexed by the ISI Web of Science citation system or other major citation indexing service. Under this relatively flexible umbrella, for each disciplines there are specific standards, in an attempt to automatize academic ranking. The ranking procedures are many times ambiguous and contradictory because of the possible exempts. Exempts are frequently modified, in accordance to the acting Minister of Education. For instance, one ISI indexed paper may be replaced by several papers indexed by other citation indexing services.

Everybody is asking for "ISI papers". Each year Thomson Reuters evaluates approximately 2,000 journals for possible coverage in Web of Science. ISI Web Of Science covers over 10,000 of the highest impact journals worldwide and over 110,000 conference proceedings. These are defined as *ISI indexed papers*. For CS, ISI indexed papers are the papers indexed by Science Citation Index Expanded. According to the present promotion regulations of the Romanian Ministry for Education, the required ISI indexed papers can be journal or conference papers. Among the many good publications covered by ISI Web of Science there also journals and proceedings of questionable quality.

Most promotion standards, including the basic criteria of the Romanian Ministry for Education, consider the number of ISI indexed papers, but not other publication assessment indicators, like impact factor and h-index. This Stakhanovist criterion favors quantity over quality. Physicists are sophisticated and they use more assessment indicators [16]: number of authors, number of citations, and impact factor. It is not easy to be a physicist in Romania, especially when you have to prepare your promotion portfolio. But, after all, let us mention that the author of the h-index is Jorge E. Hirsch, a physicist!

One may think that replacing the publication counter by the impact factor of the journal, or by the number of citations of the paper, would be sufficient to accurately quantify scholarship. At Thomson Reuters' web site [17], we find the following warning: "The impact factor should be used with informed peer review. In the case of academic evaluation for tenure it is sometimes inappropriate to use the impact of the source journal to estimate the expected frequency of a recently published article."

Using excessively the ISI indexing scheme to evaluate CS papers has additional drawbacks:

- As we have mentioned before, this creates from the very beginning a handicap for computer scientists since ISI does do not adequately cover CS.
- Another weakness of the ISI indexing scheme in CS is its poor coverage of high impact conferences, knowing that computer science uses conferences rather than journals as the main publication venue.
- A third weakness is the temptation to perform cross-disciplinary comparison.

Observation: One of the promotion requirements of the Romanian Ministry for Education is the publication of books as "first author". As we have mentioned before, the order in which a CS publication lists authors is generally not significant. For articles, these requirements do not refer to the order of authors.

9 Conclusions: The current publication and review model is killing research

How efficient are bibliometric measures, like impact factor and h-index? The UK government is considering using bibliometrics in its Research Excellence Framework, a process which will assess the quality of the research output of UK universities and on the basis of the assessment results, allocate research funding. The bibliometric indicators of research quality were tested during 2009-09 [18]. The bibliometrics pilot exercise was conducted with 22 higher education institutions and covered 35 units of assessment from the 2008 Research Assessment Exercise. Both Thomson Reuters Web of Science and Elsevier's Scopus databases were used. The pilot exercise showed that citation information is not sufficiently robust to be used formulaically or as a primary indicator of quality; but there is considerable scope for it to inform and enhance the process of expert review.

According to [19], German universities distribute money to researchers by a formula that includes the Thomson impact factor. Each point of impact factor is worth about 1000 Euros. In Pakistan, researchers receive bonuses of up to US$20,000 a year depending on the sum of the impact factors of the journals in which they publish. And the critique addressed to the Thomson impact factor, which is embedded in a commercial product, continues [19]:

> "To an extent that no one could have anticipated, the academic world has come to place enormous weight on a single measure that is calculated privately by a corporation with no accountability, a measure that was never meant to carry such a load. Yes, some of us benefit from this flawed system-in addition to other rewards that come from publishing in high-impact journals, we collect nice cash bonuses. But none of this changes the fact that evaluating research by a single number is embarrassing reductionism, as if we were talking about figure skating rather than science."

Definitely, we have to express openly our concerns about how our work is evaluated, especially by commercial bibliometric products. Not only that these products are expensive, but their misuse reduces us to figures in different statistics and rankings.

While numeric criteria trigger strong reactions, peer review is strongly dependent on evaluators' choice and availability (the most competent are often the busiest), can be biased, and does not scale up. The solution is in combining techniques, subject to human interpretation. For instance, extract, first, a citation record for the individual candidate via one of the free Internet search engines (e.g., Google Scholar). Second, ask for evaluations concerning the significance of a candidate's work from carefully selected (i.e., impartial and highly qualified) scientific peers.

The pressure to publish is too large for most to ignore. Grants don't get funded unless we splatter our names across journals and conferences the world over. Grad students don't graduate. Assistants and adjuncts don't get tenure. Your CV is fewer than 5 pages? You must be stupid. Join more vacuous clubs, dues-hungry societies, and enter more regional poster conferences.

Too much time is spent writing papers rather than developing research. Too much time is spent calculating impact factors and finding out who is indexing what. Evaluation criteria, like ISI's impact factor or conference acceptance rates are flawed. The reviewing process is inherently flawed and may kill good papers. It is hard to find good reviewer, willing to do this voluntary work.

What is the solution? One option would be to slow down. Without the pressure to publish a number of ISI indexed papers each year, regardless where and how important they are, we might get thorough, lengthy, reproducible publications. Is this not what publishing is about? What do we gain from publishing incremental research papers? There are more people writing papers than people who have time to verify their results.

Bibliography

[1] M. Del Favero and N. J. Bray, "Herding cats and big dogs: Tensions in the faculty-administrator relationship," in *Higher Education: Handbook of Theory and Research*, J. C. Smart, Ed. Springer, 2010, vol. 25, pp. 477–541. ISBN 978-90-481-8597-9 (print), 978-90-481-8598-6 (electronic), DOI: 10.1007/978-90-481-8598-6-13.

[2] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 46, pp. 16 569–16 572, November 2005. ISSN-0027-8424, DOI:10.1073/pnas.0507655102

[3] B. Meyer, C. Choppy, J. Staunstrup, and J. van Leeuwen, "Viewpoint research evaluation for computer science," *Commun. ACM*, vol. 52, no. 4, pp. 31–34, 2009, ISSN:0001-0782

[4] D. Patterson, L. Snyder, and J. Ullman, "Best practices Memo: Evaluating computer scientists and engineers for promotion and tenure," *Computer Research Association*, 1999.

[5] J. Chen and J. A. Konstan, "Conference paper selectivity and impact," *Commun. ACM*, vol. 53, no. 6, pp. 79–83, 2010, ISSN: 0001-0782

[6] C. E. Glassick, M. T. Huber, and G. I. Maeroff, *Scholarship Assessed: Evaluation of the Professoriate*. Jossey-Bass, 1997.

[7] M. Y. Vardi, "Conferences vs. journals in computing research," *Commun. ACM*, vol. 52, no. 5, pp. 5–5, 2009, ISSN: 0001-0782.

[8] R. A. Day, *How To Write & Publish a Scientific Paper*. Oryx Press, 1998.

[9] J. Swales, *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, 1990, ISBN-10: 0521338131; ISBN-13: 978-0521338134.

[10] M. M. Polycarpou, "Editorial: A new era for the IEEE Transactions on Neural Networks," *Neural Networks, IEEE Transactions on*, vol. 19, no. 1, pp. 1–2, January 2008, ISSN 1045-9227.

[11] J. Qiu, "Publish or perish in China," *Nature*, vol. 463, pp. 142–143, 2010, ISSN: 0028-0836; EISSN : 1476-4687

[12] S. Neelakantan, "In India, plagiarism is on the rise," *GlobalPost*, pp. 142–143, October, 18th, 2009.

[13] A. J. Smith, "The task of the referee," *IEEE Computer*, vol. 23, pp. 65–71, 1990.

[14] [Online]. Available: `http://cobweb.ecn.purdue.edu/~vijay/papers/acceptance.html`

[15] [Online]. Available: `www.edu.ro/`

[16] [Online]. Available: `www.fizica.unibuc.ro/fizica/`

[17] [Online]. Available: `thomsonreuters.com/products_services/science/free/essays/impact_factor/`

[18] [Online]. Available: `www.hefce.ac.uk/pubs/hefce/2009/09_39/`

[19] A. Wilcox, "Rise and fall of the Thomson impact factor," *Epidemiology*, vol. 19, pp. 373–374, 2008, ISSN: 1044-3983. Online ISSN: 1531-5487.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 4, pp. 447-457

Decentralized Controller Design for Forbidden States Avoidance in Timed Discrete Event Systems

A. Aybar

Aydın Aybar
Anadolu University, Dept. of Electrical and Electronics Engineering
26555, Eskişehir, Turkey. E-mail: aaybar@anadolu.edu.tr

Abstract: A decentralized controller design approach is developed for the timed discrete event systems which are modelled by timed automata in this work. An approach, called *augmentation*, is presented to obtain the new modelling method such that each unit delay of any event represents a pair of new state and event. The augmented automata model, obtained by using this approach, is considered to design a decentralized controller. This controller design approach is developed such that the local controller is designed for each subautomaton, obtained by using overlapping decompositions and expansions and these controllers are then combined to obtain a decentralized controller for the given timed automaton. The designed decentralized controller guarantees the unreachability of a forbidden state in the considered automaton.

Keywords: Discrete event systems, Automata, Time delays, Decentralized controller.

1 Introduction

Although automata and Petri nets are known as common modelling methods for discrete event systems (see, [1]– [4]), these models were first presented without time notation. Since there exist time delays in the dynamic systems, time notation is a necessity for the modelling methods of the discrete event systems [5,6]. Time notation was used for automata (see, [7]). In this timed automata model, a class of finite state automata was extended with a set of clocks. The clocks were chosen as real values and timed event, denoted by a pair of an event and its occurence time, were used to determine the reachability of any state. Afterwards, the timed automata model was used by many works (for example, [8–10]). Moreover, the basic supervisory controller approaches were presented for these timed systems, modelled by timed automata (for example, [11–13]).

It is known that the computational complexity of a supervisory controller design depends on the number of states and clocks for the timed automata model [10]. Moreover, the computational complexity increases exponentially with the number of states of the untimed automata model [14] (also see, [15]). Thus, a controller design for timed automata (especially, large scale automata have more number of states and events), can be more complex. An approach, called augmentation, is first introduced for timed automata in order to decrease the computational complexity, depending on the time and/or clock, of a controller design.

This approach, based on [16, 17], is described such that each unit delay of any event represents a pair of new state and event, and then these pairs are added to the original automaton. A new modelling model is introduced such that the augmented automaton is obtained by adding the pairs of events and states, corresponding to unit time delays, to the original automaton in this work. In [16, 17], the strecthing approach was developed for timed Petri nets. In this developed approach, each delay, assigned to a transition, denotes a pair of new place and transition. Using the similarity between automata and Petri nets, we first develop the augmentation approach in this work. Although any event of automata can be related to any transition of Petri net, there exist some differences between these models (for example, a marking vector of Petri net is corresponding to a state of automaton).

The augmented automaton is used to design a decentralized controller in this work. An algebraic approach, which gives the state space representation for automata, is also developed to determine the state vectors.

Our aim is a decentralized controller design which prevents the occurence of the forbidden states. To facilitate the controller design, we use the approach of overlapping decompositions. The overlapping decompositions approach was first introduced by [18] for the case of continuous-state systems (systems described by differential or difference equations with continuous state variables). This approach was then used for discrete event systems by ([4, 14, 19, 20]).

2 Preliminaries

2.1 Mathematical Model

The timed automata model is represented by $\mathcal{A}(Q,\Sigma,\mathcal{C},q_0,D)$. Here, Q is the set of states, Σ is the set of events, $\mathcal{C}: Q\times Q\to\Sigma\cup\{0\}$ is the connection matrix, q_0 is the initial state at the initial time, and D is set of the time delays of the events such that $d_e\in\mathcal{R}^+$ is the time delay of the event $e\in\Sigma$, where $\mathcal{R}^+$ is set of nonnegative real numbers.

The connection matrix is given as

$$\mathcal{C}(q_i,q_j):=\begin{cases} e, & \text{if } q_i \text{ is obtained when event } e \text{ occurs at state } q_j \\ 0, & \text{otherwise}\end{cases}, \text{ for } q_i,q_j\in Q$$

$\mathcal{C}(q_i,q_j)=0$ denotes no connection between two states q_i and q_j. $\mathcal{C}(q_i,q_j)=e$ denotes a connection between these states via e. It is assumed that the connection of between two states is done by only one event.

In this work, the vector-matrix form is used to determine new state. The state vector at time τ is denoted by $S(\tau)$

$$S(\tau):=\begin{cases}\Lambda_Q(q), & \text{if } \tau=\Gamma(q), \text{ for any } q\in Q\\ \mathcal{Z}, & \text{otherwise}\end{cases}$$

Here, $\Gamma(q)$ denotes the obtained time of state q (it is assumed that each event occurs immediately as it becames possible), $\Lambda_Q: Q\to\{0,1\}^{|Q|}$, $\Lambda_Q(q):=\begin{cases}1, & \text{if } q=[Q]_j\\ 0, & \text{otherwise}\end{cases}$, $j\in\{1,...,|Q|\}$ where, $[Q]_j$ denotes the j^{th} element of Q, and $|Q|$ indicates the number of the elements of Q, τ denotes the global time, Z, which is zeros vector, denotes that the occurence of the event e has not finished at τ or the considered event can not occured at the given state.

The state equation is given as follows:

$$S(\tau)=(\mathcal{C}\vee S(\tau_e))\wedge\mathcal{O}(e,\tau_e),\ e\in\Sigma \tag{1}$$

It is assumed that the initial state $S(\tau_0)=\Lambda_{\tilde{Q}}(q_0)$ and there exists an event e such that $\tau_e=\Gamma(q)$ for $q\in Q$ (there is one exception such that if there exists deadlock in the considered automaton, no event can occur at deadlock state), where τ_e denotes the occurence time of the event e. Note that, $\vee$ and $\wedge$ are used respectively. Here,

- The event function is defined as $\mathcal{O}(e,\tau_e):=e\odot\phi(\tau-\tau_e-d_e)$, where, $\phi:\mathcal{R}^+\to\{0,1\}$; $\phi(x)=\begin{cases}1, & \text{if } x\geq 0\\ 0, & \text{otherwise}\end{cases}$.
- The operation $\odot$ on the set $\ddot{\Sigma}:=\Sigma\cup\{0,1\}$ is defined as $0\odot 0=0$, $e_k\odot 0=0$, $0\odot e_k=0$, $0\odot 1=0$, $1\odot 0=0$, $1\odot 1=0$, $e_k\odot 1=e_k$, $1\odot e_k=e_k$, $e_k\odot e_k=0$, $e_k\odot e_l=0$, $e_l\odot e_k=0$.

- The operation $\vee$ on the matrices $H \in \ddot{\Sigma}^{|Q|\times|Q|}$ and $R \in \ddot{\Sigma}^{|Q|}$, $F = H \vee R$, is defined as $F(i) = \sum_{k=1}^{|Q|} H(i,k) \odot R(k)$, $i \in \{1,...,|Q|\}$.
- The operation $\otimes$ on the set $\ddot{\Sigma}$ is defined as $0 \otimes 0 = 0$, $e_i \otimes 0 = 0$, $0 \otimes e_i = 0$, $0 \otimes 1 = 0$, $1 \otimes 0 = 0$, $1 \otimes 1 = 1$, $e_i \otimes 1 = e_i$, $1 \otimes e_i = e_i$, $e_i \otimes e_i = 1$, $e_i \otimes e_j = 0$, $e_j \otimes e_i = 0$.
- For $F \in \ddot{\Sigma}^{|Q|}$ and $e \in \Sigma$, $F \wedge e = [F(1) \otimes e \ ... \ F(|Q|) \otimes e]^T$.

In the given example automaton, shown in Fig. 1a, the set of states is $Q = \{q_0, q_1, q_2, q_3, q_4\}$, the set of events is $\Sigma = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7\}$, and the initial state is q_0. The set of time delays, assigned to the events, is given as $d_{e_1} = d_{e_2} = d_{e_6} = 2$ sec., $d_{e_3} = 3$ sec., $d_{e_4} = d_{e_5} = d_{e_7} = 1$ sec. Let the occurence time of e_3 be $\tau_{e_3} = 5$ sec. and $S(5) = \Lambda_Q(q_2) = [0\ 1\ 0\ 0\ 0]^T$. The state vector is obtained as

$$S(\tau) = (\mathcal{C} \vee S(5)) \wedge \mathcal{O}(e_3, 5) = \left(\begin{bmatrix} 0 & 0 & 0 & e_4 & 0 \\ e_1 & 0 & 0 & 0 & 0 \\ 0 & e_3 & 0 & e_5 & 0 \\ 0 & e_2 & 0 & 0 & e_6 \\ e_7 & 0 & 0 & 0 & 0 \end{bmatrix} \vee \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right) \wedge (e_3 \odot \phi(\tau - 5 - 3))$$

$$= \begin{bmatrix} 0 \\ 0 \\ e_3 \odot 1 \\ 0 \\ 0 \end{bmatrix} \wedge (e_3 \odot \phi(\tau - 5 - 3)) = \begin{cases} [0\ 0\ 1\ 0\ 0]^T, & \text{if } \tau \geq 8 \\ [0\ 0\ 0\ 0\ 0]^T, & \text{if } 5 < \tau < 8 \end{cases}$$

q_2 is obtained when the occurence of event e_3 is completed (q_2 is not yet obtained in time interval between 5 sec. and 8 sec.). In this work, a new model is introduced in next section and used to design a decentralized controller.

2.2 Overlapping Decompositions and Expansions

Overlapping decompositions and expansions [21] have been widely used to design decentralized controllers for continuous-state systems. These concepts have also been used to design supervisory controllers for discrete event systems modeled by Petri nets [19] and by automata [14]. To our best knowledge, overlapping decompositions and expansions of discrete event systems modelled by automata or formal languages have been first introduced in [14]. In the given approach, *overlapping subautomata* of an automaton are first identified by examining the topological structure of the given automaton. These subautomata are identified such that the only interconnection between the subautomata are through the overlapping part, i.e., no event should connect two states in different subautomata, unless one of these states is in the overlapping part of the two subautomata. As an example, the automaton (Fig. 1a) can be decomposed into two subautomata as shown in Fig. 1b-1c ([14]).

After an overlapping decomposition of the original automaton is obtained, the expansions of the automaton is explained as follows [14]:

i) A state or an event in the overlapping part of *n* subautomata is repeated *n* times and each repeated state/event is assigned to a different subautomaton.

ii) Two events are introduced between any two repeated states, such that when such an event occurs the state changes from one repeated state to the other. Note that, a delay of each new event is assigned to the biggest common divisor of time delays of the original automaton in this work.

iii) If the initial state is in the overlapping part of the original automaton, then the initial state of the expanded automaton can be chosen as any one of the repeated states of the original initial state.

Otherwise, the initial state of the expanded automaton is chosen as the initial state of the original automaton.

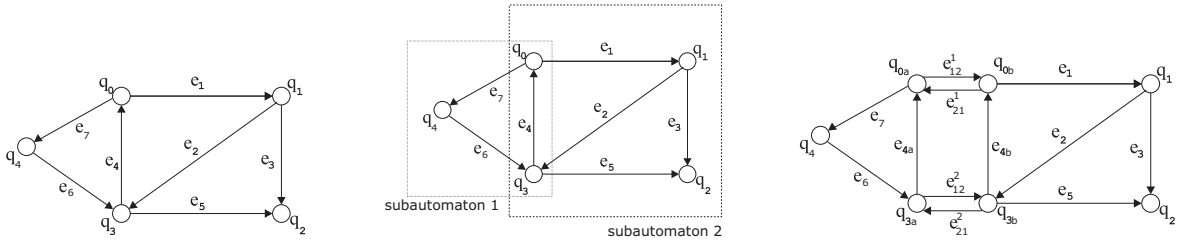


Figure 1. (a) Example automaton (b) Overlappingly decomposed automaton (c) Expanded automaton

As a result of this procedure, an expanded automaton, which consists of α disjoint subautomata, is obtained from an original automaton which was decomposed into α overlapping subautomata.

The set of states of the expanded automaton is given by $\tilde{Q} := \cup_{i=1}^{\alpha} Q_i$, where Q_i is the set of states of the i^{th} subautomaton. The set of events of the expanded automaton is given by $\tilde{\Sigma} := \check{\Sigma} \cup \grave{\Sigma}$. Here, $\check{\Sigma} = \cup_{i=1}^{\alpha} \Sigma_i$, where Σ_i is the set of events of the i^{th} subautomaton and $\grave{\Sigma}$ is the set of additional events introduced between the repeated states. As an example, the states q_0, q_3, and the event e_4 are repeated, and new events e^1_{12}, e^1_{21}, e^2_{12} and e^2_{21} are added to the repeated states in the expanded automaton in Fig. 1a. Then, the time delay of these events is determined as one second.

3 New Model for Timed Automata

Although the usage of time in the mathematical model is a necessity for the real world system, the computational complexity of time delay systems increases because of the defined all processes and functions need more memories and time. We introduce the augmentation approach. Using this approach, a new model is obtained for timed automata and called as the augmented automata, where each event has only unit time delay.

The augmentation approach is defined such that time delays are represented by new states and events in this work. The augmented automaton, $\bar{A}_T(\bar{Q}, \bar{\Sigma}, \bar{\mathcal{C}}, q_0)$, is introduced, where, $\bar{\mathcal{C}} : \bar{Q} \times \bar{Q} \to \bar{\Sigma}$, $\bar{Q} := Q \cup (\bigcup_{e \in \Sigma} \Delta_S(e))$, and $\bar{\Sigma} := \Sigma \cup (\bigcup_{e \in \Sigma} \Delta_E(e))$ are given following items.

- The time delays of the events are scaled such that $d^s_e := d_e/\lambda$, for $e \in \Sigma$ and $d_e \in D$, where λ indicates the biggest common divisor of time delays. Note that, the set of the scaled time delays of the events is denoted by D_s. It is assumed that $d^s_e \geq 1$, for all $e \in \Sigma$ in this work.

- For the event $e \in \Sigma$ such that $\mathcal{C}(q_i, q_j) = e$ and $d^s_e = 1$, the input connection from e to the state is hold such as $\bar{\mathcal{C}}(q_i, q_j) = \mathcal{C}(q_i, q_j) = e$ for $q_i, q_j \in Q$. Note that, if $\mathcal{C}(q_a, q_b) = 0$, then $\bar{\mathcal{C}}(q_a, q_b) = 0$.

- For the event $e^* \in \Sigma$, and $d^s_{e^*} > 1$, $\delta_{e^*} := d^s_{e^*} - 1$ numbers new events and states are defined such as $f^{e^*}_1, f^{e^*}_2, \ldots, f^{e^*}_{\delta_{e^*}}$, and $p^{e^*}_1, p^{e^*}_2, \ldots, p^{e^*}_{\delta_{e^*}}$. The sets are constructed by using these events and states as $\Delta_E(e^*)$ and $\Delta_S(e^*)$, respectively.

- The pairs are constructed by using the new events and states for any event $e^* \in \Sigma$, $d^s_{e^*} > 1$, such that $(f^{e^*}_i, p^{e^*}_i)$ for $i \in \{1, 2, \ldots, \delta_e\}$. For $\bar{\mathcal{C}}(q_k, q_n) = e^*$, the connections are described such as from q_n to $f^{e^*}_1$, from $f^{e^*}_1$ to $p^{e^*}_1$, from $p^{e^*}_1$ to $f^{e^*}_2$, ... from $f^{e^*}_{\delta_{e^*}}$ to q_k. Hence, the new connection matrix is constructed for the new automaton model such as $\bar{\mathcal{C}}(p^{e^*}_1, q_n) = f^{e^*}_1$, $\bar{\mathcal{C}}(p^{e^*}_2, p^{e^*}_1) = f^{e^*}_2$, ..., $\bar{\mathcal{C}}(q_k, p^{e^*}_{\delta_{e^*}}) = e^{e^*}_{\delta_{e^*}}$.

As a result of the above procedure, we obtain the augmented automaton which has more events and states but each event has only unit time delay.

We introduce the algebraic approach for the augmented automaton. Let $\bar{S}_n$ be denote the present state vector and $\bar{S}_{n+1}$ be denote the next state vector (for $n \in \{0,1,2,...\}$, $\bar{S}_0 = \Lambda_{\bar{Q}}(q_0)$ denotes the initial state vector). The state equation is defined as follows:

$$\bar{S}_{n+1} = (\bar{\mathcal{C}} \vee \bar{S}_n) \wedge \bar{e}, \;\; \bar{e} \in \bar{\Sigma}. \tag{2}$$

It is possible to obtain p_k^e as the current state. It shows that the occurence of the event e has not finished yet and also the duration time is determined as $k * \lambda + \tau_e$ for the event e. Compared to (1), the evaluation of the above equation (2) is much simpler, since it does not require the time notation and the event function $\mathcal{O}$.

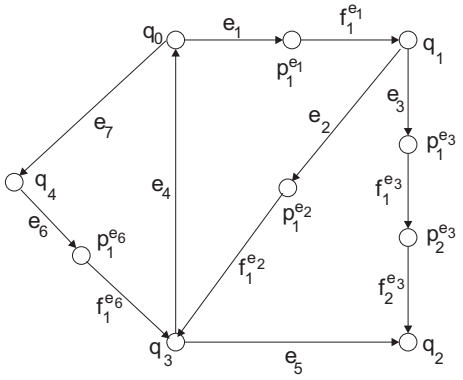


Figure 2. Augmented automaton

For example, we obtain the augmented automaton (Fig. 2.) for the given timed automata (Fig. 1a). The set of states is $\bar{Q} = \{q_0, q_1, q_2, q_3, q_4\} \cup \{p_1^{e_1}, p_1^{e_2}, p_1^{e_3}, p_2^{e_3}, p_1^{e_6}\}$, where, $\Delta_S(e_1) = \{p_1^{e_1}\}$, $\Delta_S(e_2) = \{p_1^{e_2}\}$, $\Delta_S(e_3) = \{p_1^{e_3}, p_2^{e_3}\}$, and $\Delta_S(e_6) = \{p_1^{e_6}\}$, the set of events is $\bar{\Sigma} = \{e_1, e_2, e_3, e_4, e_5, e_6\} \cup \{f_1^{e_1}, f_1^{e_2}, f_1^{e_3}, f_2^{e_3}, f_1^{e_6}\}$, where $\Delta_E(e_1) = \{f_1^{e_1}\}$, $\Delta_E(e_2) = \{f_1^{e_2}\}$, $\Delta_E(e_3) = \{f_1^{e_3}, f_2^{e_3}\}$, and $\Delta_E(e_6) = \{f_1^{e_6}\}$, and the connection matrix is given as

$$\bar{\mathcal{C}} = \begin{bmatrix} 0 & 0 & 0 & e_4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & f_1^{e_1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & e_5 & 0 & 0 & 0 & 0 & f_2^{e_3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & f_1^{e_2} & 0 & 0 & f_1^{e_6} \\ e_7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ e_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & e_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & e_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & f_1^{e_3} & 0 & 0 \\ 0 & 0 & 0 & 0 & e_6 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

4 Decentralized Controller Design

A decentralized controller for the forbidden states avoidance is developed for the considered automaton in this section.

4.1 Centralized Control

The centralized controller guarantees the unreachability of a forbidden state for the original automaton ($\mathcal{F}$ denotes the set of the forbidden states). Now, we consider a centralized controller design for the original augmented automaton (OAA).

In the OAA, the set of forbidden states is taken as $\bar{\mathcal{F}} = \mathcal{F}$. It is possible that there exists a state which only leads to any element of the set $\bar{\mathcal{F}}$. Thus, the set $\bar{\mathcal{F}}$ is extented by these sets and a new set, denoted by $\bar{\mathcal{G}}$, is obtained by the following algorithm. This algorithm, called as FS, requests the set of the forbidden states and the definition. Note that, this algorithm also finds the deadlock states, in which no event can occur, and adds these deadlock states to the set $\bar{\mathcal{G}}$. In this work, each element of $\bar{\mathcal{G}}$ is called as forbidden state.

A controller for the OAA is defined as

$$\bar{K}(\bar{S}_n, \bar{e}) = \bar{K}(\Lambda_{\bar{Q}}(q), \bar{e}) = \begin{cases} 0, & \text{if } \bar{S}_{n+1} \in \bar{\mathcal{G}}_v \\ 1, & \text{otherwise} \end{cases}, \ \bar{e} \in \bar{\Sigma} \tag{3}$$

where, $\bar{S}_n = \Lambda_{\bar{Q}}(q)$, $\bar{S}_{n+1} = (\bar{\mathcal{C}} \vee \bar{S}_n) \wedge \bar{e}$ and $\bar{\mathcal{G}}_v := \bigcup_{q \in \bar{\mathcal{G}}} \Lambda_{\bar{Q}}(q)$ denotes the set of the state vectors, corresponding to states of $\bar{\mathcal{G}}$. Note that, if q_f is a forbidden state, $q_f \in \bar{\mathcal{G}}$, then $\Lambda_{\bar{Q}}(q_f)$ is called as forbidden state vector, $\Lambda_{\bar{Q}}(q_f) \in \bar{\mathcal{G}}_v$. Once $\bar{K}(\bar{S}_n, \bar{e}) = 0$ denotes disabling event $\bar{e} \in \bar{\Sigma}$, $\bar{K}(\bar{S}_n, \bar{e}) = 1$ denotes enabling event $\bar{e}$. Then, this controller guarantees the unreachability of an element of $\bar{\mathcal{G}}$.

The OAA with the controller can be also called as controlled automaton, denoted by $\bar{\mathcal{A}}_T^K(\bar{Q}, \bar{\Sigma}, \bar{\mathcal{C}}, q_0, \bar{K})$. The controlled state equation, which is obtained by adding this controller to the equation (2), is given as follows:

$$\bar{S}_{n+1} = (\bar{\mathcal{C}} \vee \bar{S}_n) \wedge (\bar{e} \otimes \bar{K}(\bar{S}_n, \bar{e})), \ \bar{e} \in \bar{\Sigma} \tag{4}$$

Thus, any element of of $\bar{\mathcal{G}}$ does not occur in this controlled automaton.

Algorithm to construct the set $\bar{\mathcal{G}}$

$\bar{\mathcal{G}}$ **= FS(**$\bar{A}_T, \bar{\mathcal{F}}$**)**
$\bar{\mathcal{G}} = \bar{\mathcal{F}}$
Do Loop Construction
 $\hat{\mathcal{F}} = \emptyset$
 For $i = 1$ to $|\bar{Q}|$
  If $[\bar{Q}]_i \notin \bar{\mathcal{G}}$ Then
   $cnt = 0$
   For $j = 1$ to $|\bar{Q}|$
    If $\bar{C}(j,i) = 0$ Or $[\bar{Q}]_j \in \bar{\mathcal{G}}$ Then
     $cnt = cnt + 1$
     If $cnt = |\bar{Q}|$ Then
      $\hat{\mathcal{F}} \leftarrow \hat{\mathcal{F}} \cup \{[\bar{Q}]_j\}$
     End
    End
   End
  End
 End
 If $\hat{\mathcal{F}} = \emptyset$ Then
  Exit Loop Construction
 End
 $\bar{\mathcal{G}} \leftarrow \bar{\mathcal{G}} \cup \hat{\mathcal{F}}$
Loop Construction

Here, both $\cup$ and $\dot{\cup}$ are used to denote the *set union*. $L \dot{\cup} M$ is used, rather than $L \cup M$, whenever it is known apriori that $L \cap M = \emptyset$. To evaluate $N = L \cup M$, the set C is first initialized as L; for each element (from first to last), m, of M, it is then checked whether $m \in L$. If $m \notin L$, then m is added to set N. To evaluate $N = L \dot{\cup} M$, on the other hand, elements of L and of M are simply appended to form N.

4.2 Decentralized Control

Now, we first consider to design a controller for each disjoint subautomaton. Then, a controller of the expanded augmented automaton (EAA) is obtained by using these controllers of subautomata. Finally, a decentralized controller is designed by using the controller of the EAA for the OAA.

It is known that the augmented subautomata are easily obtained by using overlapping decompositions and expansions. Let $\bar{A}_{i_T}(\bar{Q}_i, \bar{\Sigma}_i, \bar{\mathcal{C}}_i, q_{i_0})$ be denote the i^{th} subautomaton. Now, some definitions and notation are given such that the EAA is denoted by $\tilde{A}_T(\tilde{Q}, \tilde{\Sigma}, \tilde{\mathcal{C}}, \tilde{q}_0)$, where, $\tilde{Q} := \cup_{i=1}^{\alpha} \bar{Q}_i$, $\tilde{\Sigma} := \cup_{i=1}^{\alpha} \bar{\Sigma}_i \cup \grave{\Sigma}$, and the connection matrix can be easily determined by using new sets of states and events.

$\Psi_{\bar{Q}} : \bar{Q} \to \tilde{Q}$, $\Psi_{\bar{Q}}(q)$ denotes the set of states in the EAA which corresponds to the state q in the OAA and $\Psi_{\bar{\Sigma}} : \bar{\Sigma} \to \tilde{\Sigma}$, $\Psi_{\bar{\Sigma}}(e)$ denotes the set of events in the EAA which corresponds to the event e in the OAA. Also, we define $\Psi_{\bar{\Sigma}}^{-1} : \tilde{\Sigma} \to \bar{\Sigma}$ and $\Psi_{\bar{Q}}^{-1} : \tilde{Q} \to \bar{Q}$ such that $e = \Psi_{\bar{\Sigma}}^{-1}(\tilde{e}) \iff \tilde{e} \in \Psi_{\bar{\Sigma}}(e)$ and $q = \Psi_{\bar{Q}}^{-1}(\tilde{q}) \iff \tilde{q} \in \Psi_{\bar{Q}}(q)$.

The set of the forbidden states for the i^{th} augmented subautomaton is obtained as $\bar{\mathcal{F}}_i := \tilde{\mathcal{F}} \cap \bar{Q}_i$, where $\tilde{\mathcal{F}} = \bigcup_{\bar{q} \in \bar{\mathcal{F}}} \Psi_{\bar{Q}}(\bar{q})$. For the i^{th} subautomaton, $\bar{\mathcal{G}}_i$ and $\bar{\mathcal{G}}_{i_v}$ are obtained by using the algorithm *FS*. Note that, this algorithm needs the definition of the i^{th} subautomaton, and the set $\bar{\mathcal{F}}_i$.

It is possible to design a controller, $\bar{K}_i$ for $\bar{A}_{i_T}$, if the initial state of this subautomaton is not a forbidden state ($q_{i_0} \notin \bar{\mathcal{G}}_i$). Since this repeated state is used for the interconnection between the subautomata (see, Section 2.2), it is assumed that any repeated state in the i^{th} subautomaton is not element of $\bar{\mathcal{G}}_i$ for all $i \in \{1, ..., \alpha\}$ ($\tilde{q} \notin \bar{\mathcal{G}}_i$ for $\tilde{q} \in \tilde{Q}_i^0$ which denotes the set of repeated states in the i^{th} subautomaton).

The controller for the EAA is designed by using local controllers, $\bar{K}_i$ for all $i \in \{1, ..., \alpha\}$, where α denotes the number of subautomata,

$$\tilde{K}(\Lambda_{\tilde{Q}}(\tilde{q}), \tilde{e}) = \begin{cases} \bar{K}_i(\Lambda_{\bar{Q}_i}(\tilde{q}), \tilde{e}), & \text{if } \tilde{e} \in \bar{\Sigma}_i \\ 1, & \text{otherwise} \end{cases}, \ \tilde{q} \in \tilde{Q} \tag{5}$$

Note that, $\tilde{\mathcal{G}} := \bigcup_{i \in \{1,...,\alpha\}} \bar{\mathcal{G}}_i$. Consequently, the controlled state equation is obtained by adding this controller to the state equation, for $\tilde{S}_n = \Lambda_{\tilde{Q}}(\tilde{q})$,

$$\tilde{S}_{n+1} = (\tilde{\mathcal{C}} \vee \tilde{S}_n) \wedge (\tilde{e} \otimes \tilde{K}(\tilde{S}_n, \tilde{e})), \ \tilde{e} \in \tilde{\Sigma}$$

Theorem 1: $\tilde{K}$ avoids the existence of the elements of $\tilde{\mathcal{G}}$ in $\tilde{A}_T^K$.
Proof: Let $\tilde{q} \in \tilde{Q}$ and $\tilde{e} \in \tilde{\Sigma}$. This state is also an element of any subautomaton, $\tilde{q} \in \bar{Q}_k$, for $k \in \{1, ..., \alpha\}$.

i) If there is no relation between $\tilde{q}$ and $\tilde{e}$, then $\tilde{K}(\Lambda_{\tilde{Q}}(\tilde{q}), \tilde{e}) = 1$ because of its definition (see, the equation (5)). In this case, $\tilde{e}$ is not occured at $\tilde{q}$ and also this value of $\tilde{K}$ does not affect the controlled state equation because of the definition of operation $\otimes$.

ii) If $\tilde{e} \in \grave{\Sigma}$, then $\tilde{K}(\Lambda_{\tilde{Q}}(\tilde{q}), \tilde{e}) = 1$. In this case, the next state is also repeated state and not a forbidden state ($\tilde{q}^o \notin \bar{\mathcal{G}}_j$ for $\tilde{q}^o \in \tilde{Q}_j^0, \forall j \in \{1, ..., \alpha\}$).

iii) If $\tilde{e} \in \bar{\Sigma}_k$, then $\tilde{K}(\Lambda_{\tilde{Q}}(\tilde{q}), \tilde{e}) = \bar{K}_k(\Lambda_{\bar{Q}_k}(\tilde{q}), \tilde{e})$. Let the next state, $\tilde{q}^+$, be obtained by using $\tilde{e}$ from $\tilde{q}$. If $\tilde{q}^+ \in \bar{\mathcal{G}}_k$, then $\bar{K}_k(\Lambda_{\bar{Q}_k}(\tilde{q}), \tilde{e}) = 0$ and $\tilde{q}^+ \in \tilde{\mathcal{G}}$ because of definition of $\tilde{\mathcal{G}}$ [$\tilde{K}(\Lambda_{\tilde{Q}}(\tilde{q}), \tilde{e}) = 0$]. Otherwise, $\bar{K}_k(\Lambda_{\bar{Q}_k}(\tilde{q}), \tilde{e}) = 1$ and $\tilde{q}^+ \notin \tilde{\mathcal{G}}$ [$\tilde{K}(\Lambda_{\tilde{Q}}(\tilde{q}), \tilde{e}) = 1$]. Note that, $\tilde{K}(\Lambda_{\tilde{Q}}(\tilde{q}), \tilde{e}) = 1$ if there is no relation between $\tilde{q}$ and $\tilde{e}$. □

We now obtain a controller, $\bar{K}$, for the OAA, by using the controller, $\tilde{K}$, for the EAA as follows:

$$\bar{K}(\Lambda_{\bar{Q}}(\bar{q}), \bar{e}) = \prod_{\tilde{q} \in \Psi_{\bar{Q}}(\bar{q})} \prod_{\tilde{e} \in \Psi_{\bar{\Sigma}}(\bar{e})} \tilde{K}(\tilde{\Lambda}(\tilde{q}), \tilde{e}), \ \bar{q} \in \bar{Q}, \ \bar{e} \in \bar{\Sigma} \tag{6}$$

Furthermore, the controlled state equation is given as $\bar{S}_{n+1} = (\bar{\mathcal{C}} \vee \bar{S}_n) \wedge (\bar{e} \otimes \bar{K}(\bar{S}_n, \bar{e}))$, $\bar{e} \in \bar{\Sigma}$, where, $\bar{S}_n = \Lambda_{\bar{Q}}(\bar{q})$, in the OAA.

Theorem 2: $\bar{K}$ guarantees the unreachability of a forbidden state in $\bar{\mathcal{A}}_T^K$.

Proof: It is known that the sets $\bar{\mathcal{G}}_j$ for all $j \in \{1, ..., \alpha\}$ and $\tilde{\mathcal{G}}$ are determined. Any repeated event in the EAA is only connected to the repeated states because of overlapping decomposition approach.

i) if $\bar{q}^\ddagger \in \bar{Q}$, $\Psi_{\bar{Q}}(\bar{q}^\ddagger) = \{\bar{q}^\ddagger\}$ and $\bar{e}^* \in \bar{\Sigma}$, $\Psi_{\bar{\Sigma}}(\bar{e}^*) = \{\bar{e}_a^*, \bar{e}_b^*, ..., \bar{e}_x^*\}$, then
$\bar{K}(\Lambda_{\bar{Q}}(\bar{q}^\ddagger), \bar{e}^*) = \tilde{K}(\Lambda_{\tilde{Q}}(\bar{q}^\ddagger), \bar{e}_a^*)$.
$\tilde{K}(\Lambda_{\tilde{Q}}(\bar{q}^\ddagger), \bar{e}_b^*)$. ... $\tilde{K}(\Lambda_{\tilde{Q}}(\bar{q}^\ddagger), \bar{e}_x^*)$. Since any event, in the overlapping part, is only connected to the states in the overlapping part, $\bar{K}(\Lambda_{\bar{Q}}(\bar{q}^\ddagger), \bar{e}^*) = 1$.

ii) if $\bar{q}^\ddagger \in \bar{Q}$, $\Psi_{\bar{Q}}(\bar{q}^\ddagger) = \{\bar{q}^\ddagger\}$ and $\bar{e}^* \in \bar{\Sigma}$, $\Psi_{\bar{\Sigma}}(\bar{e}^*) = \{\bar{e}^*\}$, then $\bar{q}^\ddagger$ and $\bar{e}^*$ are elements of subautomata. Note that, $\bar{K}(\Lambda_{\bar{Q}}(\bar{q}^\ddagger), \bar{e}^*) = \tilde{K}(\Lambda_{\tilde{Q}}(\bar{q}^\ddagger), \bar{e}^*) = 0$ if $\bar{q}^\ddagger$ and $\bar{e}^*$ are not in same automaton. If there is a relation between $\bar{q}^\ddagger$ and $\bar{e}^*$ in the j^{th} subautomaton, then $\bar{K}(\Lambda_{\bar{Q}}(\bar{q}^\ddagger), \bar{e}^*) = \tilde{K}(\Lambda_{\tilde{Q}}(\bar{q}^\ddagger), \bar{e}^*) = \bar{K}_j(\Lambda_{\bar{Q}_j}(\bar{q}^\ddagger), \bar{e}^*)$. In this case, if $\bar{q}^u$, which is obtained by using $\bar{e}^*$ from $\bar{q}^\ddagger$, is an element of $\bar{\mathcal{G}}_j$, then $\bar{K}_j(\Lambda_{\bar{Q}_j}(\bar{q}^\ddagger), \bar{e}^*) = 0$. Otherwise, $\bar{K}_j(\Lambda_{\bar{Q}_j}(\bar{q}^\ddagger), \bar{e}^*) = 1$.

iii) If $\bar{q}^\ddagger \in \bar{Q}$, $\Psi_{\bar{Q}}(\bar{q}^\ddagger) = \{\bar{q}_a^\ddagger, \bar{q}_b^\ddagger, ..., \bar{q}_y^\ddagger\}$ and $\bar{e}^* \in \bar{\Sigma}$, $\Psi_{\bar{\Sigma}}(\bar{e}^*) = \{\bar{e}^*\}$, then
$\bar{K}(\Lambda_{\bar{Q}}(\bar{q}^\ddagger), \bar{e}^*) = \tilde{K}(\Lambda_{\tilde{Q}}(\bar{q}_a^\ddagger), \bar{e}^*)$. $\tilde{K}(\Lambda_{\tilde{Q}}(\bar{q}_b^\ddagger), \bar{e}^*)$. .. .$\tilde{K}(\Lambda_{\tilde{Q}}(\bar{q}_y^\ddagger), \bar{e}^*)$. If $\bar{e}^*$ is not connected to any element of $\Psi_{\bar{Q}}(\bar{q}^\ddagger)$, then $\bar{K}(\Lambda_{\bar{Q}}(\bar{q}^\ddagger), \bar{e}^*) = 1$. 1 ... .1 = 1. Let $\bar{q}_l^\ddagger$ and $\bar{e}^*$ in the l^{th} subautomaton (the elements of $\Psi_{\bar{Q}}(\bar{q}^\ddagger) \setminus \{\bar{q}_l^\ddagger\}$ are in the other subautomata, $\tilde{K}(\Lambda_{\tilde{Q}}(\dot{\tilde{q}}), \bar{e}^*) = 1$, for $\dot{\tilde{q}} \in \Psi_{\bar{Q}}(\bar{q}^\ddagger) \setminus \{\bar{q}_l^\ddagger\}$). In this case, $\bar{K}(\Lambda_{\bar{Q}}(\bar{q}^\ddagger), \bar{e}^*) = 1$ .... 1 .$\bar{K}_l(\Lambda_{\bar{Q}_l}(\bar{q}_l^\ddagger), \bar{e}^*)$ .1 ...1 = $\bar{K}_l(\Lambda_{\bar{Q}_l}(\bar{q}_l^\ddagger), \bar{e}^*)$ is obtained. Here, if $\hat{\bar{q}}$, which is obtained by using $\bar{e}^*$ from $\bar{q}_l^\ddagger$, is an element of $\tilde{\mathcal{G}}$ then $\bar{K}_l(\Lambda_{\bar{Q}_l}(\bar{q}_l^\ddagger), \bar{e}^*) = 0$ [$\bar{K}(\Lambda_{\bar{Q}}(\bar{q}^\ddagger), \bar{e}^*) = 0$]. Otherwise, $\bar{K}_l(\Lambda_{\bar{Q}_l}(\bar{q}_l^\ddagger), \bar{e}^*) = 1$ [$\bar{K}(\Lambda_{\bar{Q}}(\bar{q}^\ddagger), \bar{e}^*) = 1$].

iv) If $\bar{q}^\ddagger \in \bar{Q}$, $\Psi_{\bar{Q}}(\bar{q}^\ddagger) = \{\bar{q}_a^\ddagger, \bar{q}_b^\ddagger, ..., \bar{q}_y^\ddagger\}$ and $\bar{e}^* \in \bar{\Sigma}$, $\Psi_{\bar{\Sigma}}(\bar{e}^*) = \{\bar{e}_a^*, \bar{e}_b^*, ..., \bar{e}_x^*\}$, then $\bar{K}(\Lambda_{\bar{Q}}(\bar{q}^\ddagger), \bar{e}^*)$ $= 1$ .... $1 = 1$ since any repeated state is not element of $\tilde{\mathcal{G}}$.

Note that, each element of $\bar{\mathcal{G}}$ is also an element of $\tilde{\mathcal{G}}$ because of overlapping decompositions and expansions approach and $\bar{q} \in \bar{\mathcal{G}} \Rightarrow \bar{q} \in Q_j$, for $j \in \{1, ..., \alpha\}$. Thus, $\bar{q}$ is an element of $\bar{\mathcal{G}}_j$ and also $\bar{q} \in \tilde{\mathcal{G}}$. This decentralized controller prevents the forbidden states in $\bar{\mathcal{A}}_T^K$. □

We can obtain a decentralized controller for the original automaton as follows:

$$K(\Lambda_Q(q), e) = \bar{K}(\Lambda_{\bar{Q}}(q), e), \;\; q \in Q, \; e \in \Sigma \tag{7}$$

This controller is added to the state equation (1) and $S(\tau) = (\mathcal{C} \vee S(\tau_e)) \wedge (\mathcal{O}(e, \tau_e) \otimes K(S(\tau_e), e))$, for $e \in \Sigma$ is obtained. It is known that, although the forbidden states are elements of Q, the new states may be elements of $\bar{\mathcal{G}}$. The controller of the OAA only disables the elements of Σ because of the connection of the pairs new states and events (see, Section 4). Therefore, the occurence of any event, which is disabled at any state by the controller (6), is disabled for the original automaton by the controller (7).

5 Example

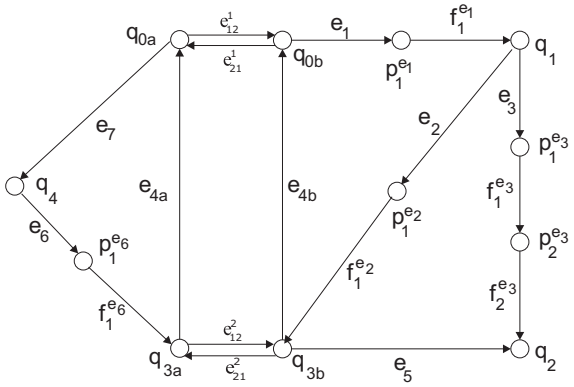


Figure 3. Expanded automaton

In this section, we design a decentralized controller, which guarantees the forbidden states avoidance, for the given timed automaton (Fig. 1a). The augmented automaton for this automaton is obtained as Fig. 2. The EAA, shown in Fig. 3, is obtained by using overlapping decompositions and expansions.

For the original timed automaton, the set of forbidden states is given as $\mathcal{F} = \{q_2\}$ and $\bar{\mathcal{F}} = \mathcal{F}$. Then, $\tilde{\mathcal{F}} = \bigcup_{\bar{q}\in\bar{\mathcal{F}}} \Psi_{\bar{Q}}(\bar{q}) = \{q_2\}$. Now, we consider two subautomata to design a decentralized controller.

In the first subautomaton, $\bar{A}_{1_T}(\bar{Q}_1, \bar{\Sigma}_1, \bar{\mathcal{C}}_1, q_{1_0})$, the set of states is $\bar{Q}_1 = \{q_{0a}, q_{3a}, q_4, p_1^{e_6}\}$, the set of events is $\bar{\Sigma}_1 = \{e_4, e_{6a}, e_7, f_1^{e_6}\}$, and the initial state is $q_{1_0} = q_{0a}$. In the second subautomaton, $\bar{A}_{2_T}(\bar{Q}_2, \bar{\Sigma}_2, \bar{\mathcal{C}}_2, q_{2_0})$, the set of states is $\bar{Q}_2 = \{q_{0b}, q_2, q_{3b}, p_1^{e_1}, p_1^{e_2}, p_1^{e_3}, p_2^{e_3}\}$, the set of events is $\bar{\Sigma}_2 = \{e_1, e_2, e_3, e_{4b}, f_1^{e_1}, f_1^{e_2}, f_1^{e_3}, f_2^{e_3}\}$, and the initial state is $q_{2_0} = q_{0b}$. The connection matrices are

$$\bar{\mathcal{C}}_1 = \begin{bmatrix} 0 & e_{4a} & 0 & 0 \\ 0 & 0 & 0 & f_1^{e_6} \\ e_7 & 0 & 0 & 0 \\ 0 & 0 & e_6 & 0 \end{bmatrix} \quad \text{and} \quad \bar{\mathcal{C}}_2 = \begin{bmatrix} 0 & 0 & 0 & e_{4b} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & f_1^{e_1} & 0 & 0 & 0 \\ 0 & 0 & 0 & e_5 & 0 & 0 & 0 & f_2^{e_3} \\ 0 & 0 & 0 & 0 & 0 & f_1^{e_2} & 0 & 0 \\ e_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & e_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & e_3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & f_1^{e_3} & 0 \end{bmatrix}.$$

For each subautomaton, $\bar{\mathcal{F}}_i = \tilde{\mathcal{F}} \cap \bar{Q}_i$ for $i \in \{1,2\}$ is obtained such as $\bar{\mathcal{F}}_1 = \emptyset$ and $\bar{\mathcal{F}}_2 = \{q_2\}$. In the subautomata, the set $\bar{\mathcal{G}}_1 = \emptyset$ is obtained by using the algorithm *FS* (note that, $\bar{\mathcal{G}}_{1_v} = \emptyset$). Thus, $\bar{K}_1(\Lambda_{\bar{Q}_1}(q_{0a}), \bar{e}^+) = \bar{K}_1([1\ 0\ 0\ 0]^T, \bar{e}^+) = 1$ for all $\bar{e}^+ \in \bar{\Sigma}_1$ and $\bar{K}_1(\Lambda_{\bar{Q}_1}(\bar{q}^*), \bar{e}^x) = 1$ for all $\bar{q}^* \in \bar{Q}_1$ and $\bar{e}^x \in \bar{\Sigma}_1$. The set $\bar{\mathcal{G}}_2 = \{q_2, p_2^{e_3}, p_1^{e_3}\}$ is obtained by using the algorithm *FS* (note that, $\bar{\mathcal{G}}_{2_v} = \{[0\,1\,0\,0\,0\,0\,0\,0]^T, [0\,0\,0\,0\,0\,0\,0\,1]^T, [0\,0\,0\,0\,0\,0\,1\,0]^T\}$). Thus, $\bar{K}_2(\Lambda_{\bar{Q}_2}(q_1), e_5) = 0$, $\bar{K}_2(\Lambda_{\bar{Q}_2}(q_{3a}), e_3) = 0$ and $\bar{K}_2(\Lambda_{\bar{Q}_2}(\bar{q}^x), \bar{e}^*) = 1$ for all $\bar{q}^x \in \bar{Q}_2 \setminus \{q_1, q_{3a}\}$ and $\bar{e}^* \in \bar{\Sigma}_2$.

Using the equation (5), the controller is obtained for the EAA as $\tilde{K}(\Lambda_{\tilde{Q}}(q_1), e_5) = \bar{K}_2(\Lambda_{\bar{Q}_2}(q_1), e_5) = 0$, $\tilde{K}(\Lambda_{\tilde{Q}}(q_{3b}), e_3) = \bar{K}_2(\Lambda_{\bar{Q}_2}(q_{3b}), e_3) = 0$, and $\tilde{K}(\Lambda_{\tilde{Q}}(\tilde{q}^\ddagger), \tilde{e}^c) = 1$, $\forall \tilde{q}^\ddagger \in \tilde{Q} \setminus \{q_1, q_{3a}\}$, $\forall \tilde{e}^c \in \tilde{\Sigma}$. It is known that $\tilde{S}_n = \Lambda_{\tilde{Q}}(q_1) = [0\ 1\ 0\ 0\ 0\ 0\ 0\ 0; 0\ 0]^T$, and $\Lambda_{\bar{Q}_2}(q_1) = [0\ 1\ 0\ 0\ 0\ 0\ 0\ 0]^T$.

Finally, a decentralized controller, which avoids the forbidden states, is designed by using (6). This controller is given as
$\bar{K}(\Lambda_{\tilde{Q}}(q_1), e_5) = \tilde{K}(\Lambda_{\tilde{Q}}(q_1), e_5) = \bar{K}_2(\Lambda_{\bar{Q}_2}(q_1), e_5) = 0$, $\bar{K}(\Lambda_{\tilde{Q}}(q_3), e_3) = \tilde{K}(\Lambda_{\tilde{Q}}(q_{3a}), e_3)\ .\ \tilde{K}(\Lambda_{\tilde{Q}}(q_{3b}), e_3) = \bar{K}_1(\Lambda_{\bar{Q}_1}(q_{3a}), e_3)\ .\ \bar{K}_2(\Lambda_{\bar{Q}_2}(q_{3b}), e_3) = 1\ .\ 0 = 0$, and $\bar{K}(\Lambda_{\tilde{Q}}(q^d), e^w) = 1$, $\forall q^d \in \tilde{Q} \setminus \{q_1, q_3\}$, $\forall e^w \in \tilde{\Sigma}$. The final result is given such that the occurence of e_5 is disabled at the state q_1 and the occurence of e_3 is

disabled at the state q_3. Thus, decentralized controller avoids state q_2. For the original timed automaton, the controller is obtained as $K(\Lambda_Q(q_3), e_3) = 0$, $K(\Lambda_Q(q_1), e_5) = 0$, and $K(\Lambda_Q(q^x), e^z) = 1$, $\forall q^x \in Q \setminus \{q_1, q_3\}$, $\forall e^z \in \Sigma$.

Now, let us compare the results of centralized and decentralized controllers for the given automaton. Both of these controllers disables the occurence of e_5 the state q_1 and the occurence of e_3 at the state q_3. The most advantage is that the size of the connection matrix for each subautomaton is smaller then the size of the connection matrix of the OAA.

6 Conclusion

A decentralized controller approach using overlapping decompositions for the timed discrete event systems. An approach, called *augmentation*, is presented to obtain the new modelling method such that each unit delay of any event represents a pair of new state and event. The augmented automaton is constructed by adding the pairs of events and states to the original automaton in this work.

The decentralized controller design approach is presented to prevent the occurence of the forbidden states. The augmented automaton is first decomposed overlappingly and expanded to obtain subautomata. Then, a controller is designed for each disjoint subautomaton. These local controllers are then combined to obtain a controller for the augmented automaton. Moreover, the state space representation is used for timed and untimed automata by the given algebraic approach.

Since the clock or timer does not used to analyse for the augmented automaton, the first advantage is that the computational complexity does not depend on clock for the timed automata. For the construction of the augmented autonmaton, the new states and events are added to the original automaton, and then the size of the connection matrix of the original automaton is smaller than the size of the connection matrix of the augmented automaton. Although this seems to be a disadvantage, the connection matrices of the augmented subautomata are only used to design the decentralized controller (i.e., the connection matrix of the augmented automaton is not used for the decentralized controller design approach). The size of the connection matrix of each subautomaton is an advantage for the decentralized approach (the number of states and events of subautomata is less than original automaton, [14, 15]).

Although the effort needed to obtain a useful the overlapping decomposition, this can be not comparable to the controller design since the decomposition may, in most cases, be easily made. Further research can also be undertaken to use this approach to design decentralized controllers for various objectives (for example, a controller can be designed such that this controller leads the given discrete event systems to marked states).

Bibliography

[1] P. J. G. Ramadge and W. M. Wonham, "The control of discrete event systems," *Proceedings of the IEEE*, vol. 77, pp. 81–98, 1989.

[2] R. S. Sreenivas and B. H. Krogh, "On Petri net models of infinite state supervisors," *IEEE Transactions on Automatic Control*, vol. 37, pp. 274–277, 1992.

[3] A. Aybar and A. İftar, "Decentralized supervisory controller design to avoid deadlock in Petri nets," *International Journal of Control*, vol. 76, pp. 1285–1295, 2003.

[4] A. Aybar and A. İftar, "Decentralized supervisory controller design for discrete-event systems using overlapping decompositions and expansions," *Dynamics of Continuous, Discrete and Impulse Systems (Series B)*, vol. 11, pp. 553–568, 2004.

[5] A. A. Desrochers and R. Y. Al-Jaar, *Applications of Petri Nets in Manufacturing Systems*, The Institute of Electrical and Electronics Engineers Inc., New York, 1995.

[6] M. Zhou and F. DiCesare, *Petri Net Synthesis for Discrete Event Control of Manufacturing Systems*, Kluwer Academic, Norwell, MA, 1993.

[7] R. Alur and D. L. Dill, "A theory of timed automata," *Theoretical Computer Science*, vol. 126, pp. 183–235, 1994.

[8] A. Gouin and J. Ferrier, "Temporal coherence of timed automata product," in *Proc. of the 1999 IEEE International Conference on Systems, Man, and Cybernetics*, October 1999, pp. 176–181.

[9] J. Krakora, L. Waszniowski, P. Pisa, and Z. Hanzalek, "Timed automata approach to real time distributed system verification," in *Proc. of the 2004 IEEE International Workshop on Factory Communication Systems*, September 2004, pp. 407–410.

[10] A. Khoumsi, "A supervisory control method for ensuring the comformance of real-time discrete event systems," *Discrete Event Dynamic Systems: Theory and Applications*, vol. 15, pp. 397–431, 2005.

[11] B. A. Bradin and W. M. Wonham, "Supervisory control of timed discrete–event systems," *IEEE Transactions on Automatic Control*, vol. 39, pp. 329–342, 1994.

[12] F. Lin and W. M. Wonham, "Supervisory control of timed discrete–event systems under partial observation," *IEEE Transactions on Automatic Control*, vol. 40, pp. 558–562, 1995.

[13] I. Açıksöz, "Time step approach for timed automata model (in turkish)," M.S. thesis, Anadolu University, Eskişehir, Turkey, June 2006.

[14] A. Aybar and A. İftar, "Overlapping decompositions of large–scale discrete–event systems," in *Proceeding CD-ROM of The 15th IFAC World Congress*, Barcelona, Spain, July 2002.

[15] K. Rudie and W. M. Wonham, "Think globally, act locally: decentralized supervisory control," *IEEE Transactions on Automatic Control*, vol. 37, pp. 1692–1708, 1992.

[16] A. Aybar and A. İftar, "Supervisory controller design for timed Petri nets," in *Proceedings of the IEEE International Conference on System of Systems Engineering*, Los Angeles, CA, U.S.A., Apr. 2006, pp. 59–64.

[17] A. Aybar and A. İftar, "Deadlock avoidance controller design for timed Petri nets using stretching," *IEEE Systems Journal*, vol. 2, pp. 178–188, 2008.

[18] M. Ikeda and D. D. Šiljak, "Overlapping decompositions, expansions, and contractions of dynamic systems," *Large Scale Systems*, vol. 1, pp. 29–38, 1980.

[19] A. Aybar and A. İftar, "Overlapping decompositions and expansions of Petri nets," *IEEE Transactions on Automatic Control*, vol. 47, pp. 511–515, 2002.

[20] A. Aybar, A. İftar, and H. Apaydın-Özkan, "Centralized and decentralized supervisory controller design to enforce boundedness, liveness, and reversibility in Petri nets," *International Journal of Control*, vol. 78, pp. 537–553, 2005.

[21] M. Ikeda and D. D. Šiljak, "Overlapping decentralized control with input, state, and output inclusion," *Control Theory and Advanced Technology*, vol. 2, pp. 155–172, 1986.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 4, pp. 458-468

Genetic Algorithm Based Feature Selection In a Recognition Scheme Using Adaptive Neuro Fuzzy Techniques

M. Bhattacharya, A. Das

Mahua Bhattacharya
Indian Institute of Information Technology & Management
Morena Link Road, Gwalior-474003, India
E-mail: mb@iiitm.ac.in

Arpita Das
Institute of Radio Physics & Electronics
University of Calcutta
92, A.P.C. Road, Kolkata-700009
E-mail: arpita.rpe@caluniv.ac.in

Abstract:
The problem of feature selection consists of finding a significant feature subset of input training as well as test patterns that enable to describe all information required to classify a particular pattern. In present paper we focus in this particular problem which plays a key role in machine learning problems. In fact, before building a model for feature selection, our goal is to identify and to reject the features that degrade the classification performance of a classifier. This is especially true when the available input feature space is very large, and need exists to develop an efficient searching algorithm to combine these features spaces to a few significant one which are capable to represent that particular class. Presently, authors have described two approaches for combining the large feature spaces to efficient numbers using *Genetic Algorithm* and *Fuzzy Clustering* techniques. Finally the classification of patterns has been achieved using adaptive neuro-fuzzy techniques. The aim of entire work is to implement the recognition scheme for classification of tumor lesions appearing in human brain as space occupying lesions identified by CT and MR images. A part of the work has been presented in this paper. The proposed model indicates a promising direction for adaptation in a changing environment.
Keywords: Adaptive neuro- fuzzy, Genetic algorithm, Feature selection, pattern recognition.

1 Introduction

The boundary detection based on Fourier Descriptors introduces a large number of feature vectors in a pattern recognition scheme. To classify different boundaries, any standard classifier needs large number of inputs and to train the classifier large number of training cycles and huge memory are also required. A complicated structure of the classifier invites the problem of over learning, and which may cause for misclassification [2]. Therefore need exists for significant feature selection for efficient pattern recognition scheme. Among many existing methods for solving feature selection problem (FSP), pruning methods for neural network [7],[8], classification trees [9] fuzzy clustering [10] may be referred. GA is an efficient search algorithm based on the mechanics of natural selection and natural genetics [1]. It combines survival of the fittest among string structures with a structured yet randomized information exchange to form a search algorithm with some of the innovative flair of human search. Since genetic algorithm is invented to simulate evolutionary processes observed in nature the goal of survival or optimization in a changing environment could be achieved [3]. However, GA [1],[4],[5],[6],[11] differs

from other searching algorithm in that sense, it does not deal with the neighborhood of a single current solution. GA use a collection (or population) of parameters, from which using selective crossover and mutation strategies, better solutions may come out. In present paper, the network architecture used for final classification is ANFIS adaptive neuro fuzzy inference system. ANFIS [13],[14] architecture for Sugeno fuzzy model is an innovative soft computing expert system that removes the limitations of conventional neural networks [12],[13],[15]. The proposed method of feature selection has been compared with Fuzzy clustering theory where GA based feature selection shows the improvement over fuzzy clustering due to natural selection mechanisms. Proposed FSP methodology combined with ANFIS classifier is an intelligent, expert system that gives the user accurate detection even in presence of additive noise. The objective of entire work is to identify the different space occupying lesions appearing in human brain as tumor / cancer lesions in different grades of benignancy / malignancy using boundary as feature. Presently a part of the work has been presented considering few pattern boundaries in order to develop an accurate classification technique using GA based feature selection.

2 Proposed Methodology

In the proposed method the significant boundary of ROI is

extracted and GA has been applied to reduce the feature vector size. These reduced and significant features are then fed to ANFIS Sugeno fuzzy network for classification. A comparative study has been conducted for efficient feature selection using both GA and FCM and finally to classify patterns using ANFIS. This study effectively gives the superior results for GA based feature selection. The method is summarized in Figure-1.

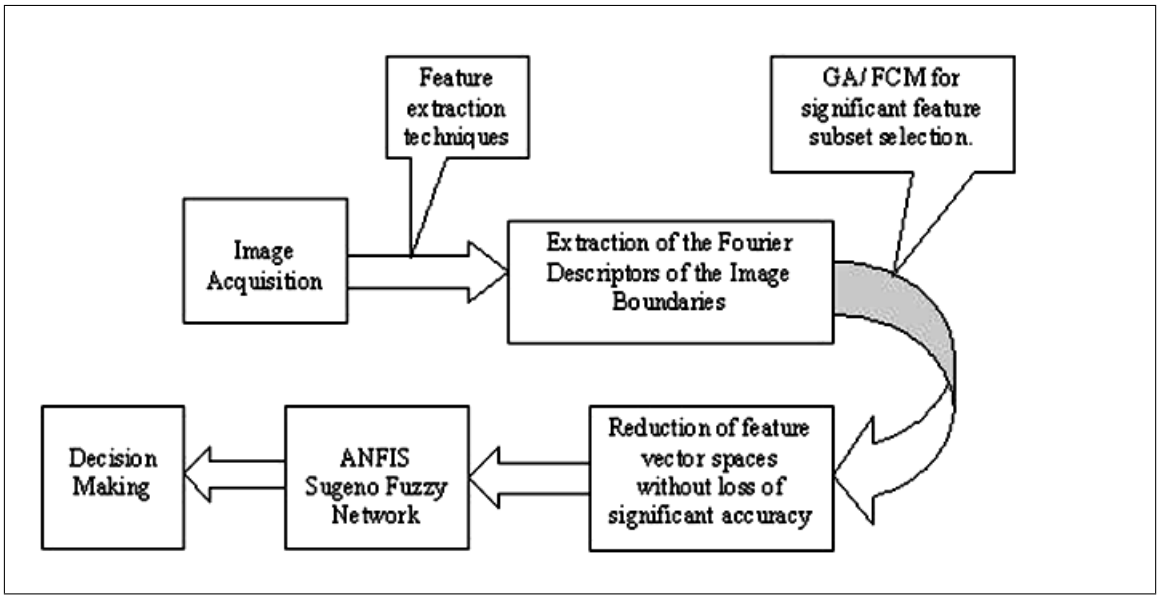


Figure 1: Proposed technique

2.1 Boundary Extraction using Fourier Descriptors

Feature selection is the choice of descriptors in a particular application. The boundary of pattern to be analyzed has been detected by implementing *Canny edge detector* and *Fourier Descriptors* of the edges then used as shape information. A figure with k-points digital boundary in the x-y plane as, $x(k) = x_k$,

$y(k) = y_k$ can be represented as

$$s(k) = [x(k)y(k)] \text{ for } k = 0, 1, 2, ..., k-1. \tag{1}$$

Each co-ordinate pair can be treated as a complex number so that,

$$s(k) = x(k) + j * y(k) \text{ for } k = 0, 1, 2, ..., k-1. \tag{2}$$

The Discrete Fourier Transform (DFT) of s(k) is given below

$$a(u) = \frac{1}{k} * \sum_{k=0}^{K-1} s(k) * e^{-\frac{j2(\pi)uk}{K}} \text{for} k = 0, 1, 2, ..., k-1. \tag{3}$$

The complex coefficient $a(u)$ is called the Fourier Descriptor of the edge points. Let us suppose that instead of all Fourier coefficients, only the first 'P' coefficients are used. This is equivalent to set $a(u) = 0$ for $u > (P-1)$. The overall global shape of the images has been identified (It can be shown that if $P \approx u/3$, approximate boundary detection would be possible). Thus a few Fourier descriptors can be used to capture the gross essence of a boundary. This property is valuable, because these coefficients carry shape information and can be used as the basis for differentiating between distinct boundary shapes.

2.2 Genetic Algorithm for Feature Selection

GA manipulates chromosomes, which are the encoded string set of parameters of a target system to be optimized. Presently different boundaries extracted from CT and MR images of section of human brain having space occupying lesions are recognized on the basis of Fourier Descriptors and which play the role of payoff values (objective function) associated with individual strings. In GA, a new set of offspring has been created in every generation on the basis of the fittest of old generation. GA efficiently exploits the historical information to speculate on new search points with expected improved performance. It is the best learned from the careful study of biological example that, where robust performance is desired, nature does it better which is the secret of adaptation and survival. GA uses three operators: selection (or reproduction), crossover and mutation to achieve the goal of evolution [1],[3].

2.3 Fuzzy C-Means Clustering Algorithm for Feature Selection

In the proposed method, fuzzy c-means clustering algorithm used for reduction of input feature vector sizes without loss of accuracy level of detection.

Algorithm 1. *Let $X = \{x_1, x_2, ..., x_n\}$ be a set of given data. A fuzzy c-partition of X is a family of fuzzy subsets of x, denotes by $P = \{A_1, A_2, ..., A_c\}$, which satisfies*

$$\sum_{i=1}^{c} A_i(x_k) = 1 \textit{ for all } k \in N_n \tag{4}$$

and $i = 1$

$$0 < \sum_{k=1}^{n} A_i(x_k) < n \textit{ for all } i \in N_c \tag{5}$$

where c is a positive integer Given a set of data $X = \{x_1, x_2, ..., x_n\}$, where x_k, in general is a vector, for all $k \in N_n$, the problem fuzzy clustering is to find a fuzzy pseudo partition and the associated cluster centers by which the structure of the data is represented as best as possible. To solve the problem of fuzzy

clustering, we need to formulate a performance index. Usually, the performance index is based upon cluster centers, $v_1, v_2, ..., v_c$ associated with the partition are calculated by the formula.

$$v_i = \frac{\sum_{k=1}^{n} [A_i(x_k)]^m x_k}{\sum_{k=1}^{n} [A_i(x_k)]^m} \tag{6}$$

for all $i \in N_c$, where $m > 1$ is a real number that governs the influence of membership grades. Observe that the vector v_i calculated by above equation is viewed as the cluster center of the fuzzy class A I, is actually weighted average of data in A_i. The performance index of a fuzzy pseudo partition P, $J_m(P)$, is defined in terms of the cluster centers by the formula

$$J_m(P) = \sum_{k=1}^{n} \sum_{i=1}^{c} [A_i(x_k)]^m \|x_k - v_i\|^2 \tag{7}$$

where $\|x_k - v_i\|^2$ represents the distance between x_k and v_i . Clearly, the smaller the value of $J_m(P)$, the better the fuzzy pseudo partition P. Thus, the goal of fuzzy c-means clustering method is to find a fuzzy pseudo partition P that minimizes the performance index $J_m(P)$.

2.4 Classification of Features using ANFIS model

A generalized ANFIS model based on Sugeno fuzzy architecture is utilized for classification of significant features. The numbers of input nodes are equal to the reduced input feature space sizes. The number of membership functions in each of the input node is continually adjusted to achieve the optimum classification results. To adapt the model with ever-changing environments, hybrid-learning rule is used.

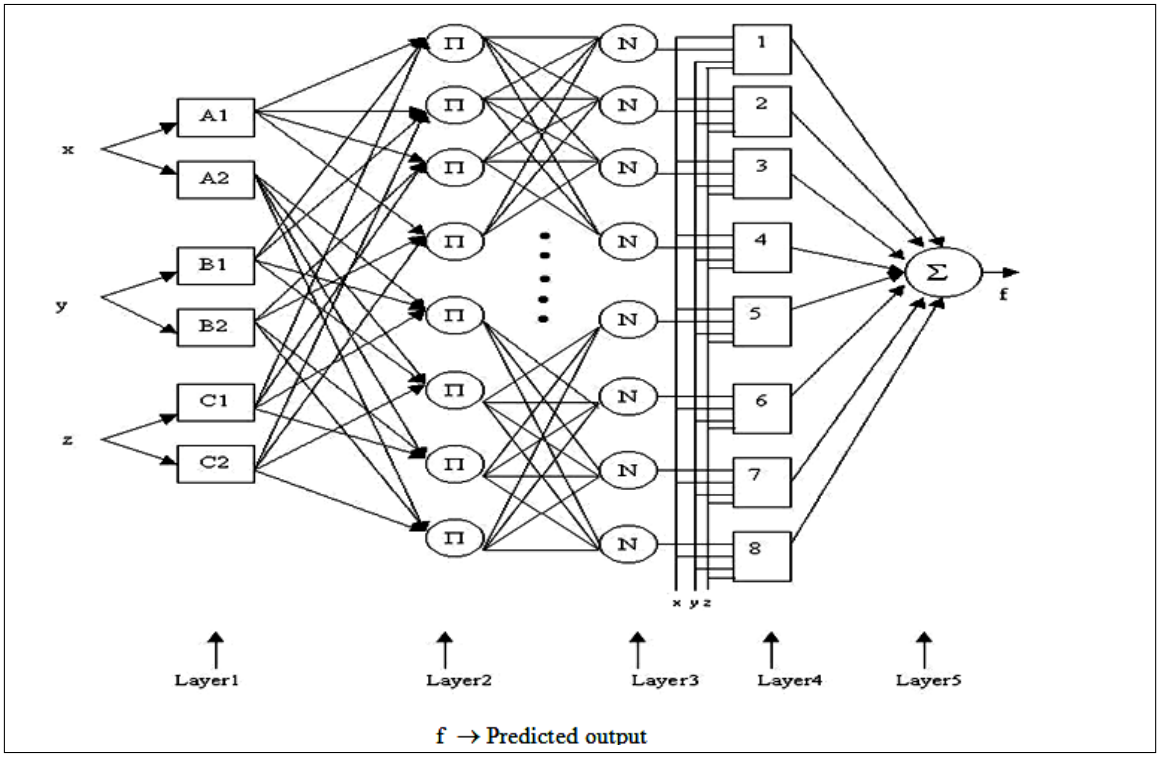


Figure 2: The ANFIS Model for Final classification.

Figure-2 illustrates the reasoning mechanism of the Sugeno fuzzy ANFIS architecture for boundary detection and texture analysis of masses respectively, where nodes of the same layer have similar functions as described below.

Layer 2. Every node I in this layer is an adaptive node with a node function $O_{1i} = \mu_{A_i}(x)$ for $i = 1,2$. and $O_{1i} = \mu_{B_i}(y)$ where x (or y) is the input to node i and A_i (or B_i) is a linguistic label (such as large or small) associated with this node. In other words O_{1i}, i is the membership grade of fuzzy set $A(A_1,A_2)$ or $B(B_1,B_2)$. Here the membership function for A can be any appropriate parameterized membership function, such as generalized bell function:

$$\mu_A(x) = \frac{1}{1+|\frac{(x-c_i)}{a_i}|^{2b}} \tag{8}$$

where a_i, b_i, c_i is the parameter set. As the values of these parameters change, the bell-shaped function varies accordingly. Parameters of this layer are referred to as premise parameters.

Layer 3. Every node in this layer is a fixed node labeled $\prod$, whose output is the product of all the incoming signals:

$$O_{2,i} = w_i = \mu_{A_i}(x)\mu_{B_i}(x) \text{ for } i = 1,2. \tag{9}$$

In general, any T-norm operator that performs fuzzy AND can be used as the node function in this layer.

Layer 4. Every node in this layer is a fixed node labeled N. The ith node calculates the ratio of the rule's firing strength to the sum of all rule's firing strengths:

$$O_{3,i} = \overline{w_i} = \frac{w_i}{(w_1+w_2)} \text{ for } i = 1,2. \tag{10}$$

For convenience, outputs of this layer are called normalized firing strengths.

Layer 5. Every node i in this layer is an adaptive node with a node function

$$O_{4,i} = \overline{w_i} f_i = \overline{w_i}(p_i x + q_i y + r_i) \tag{11}$$

where w_i is a normalized firing strength from layer 3 and p_i, q_i, r_i is the parameter set of this node. Parameters of this layer are referred to as consequent parameters.

Layer 6. The single node in this layer is fixed node labeled $\sum$, which computes the overall output as the summation of all incoming signals:

$$\text{Overall output } = O_{5,i} = \frac{\sum_i w_i f_i}{\sum_i w_i} \tag{12}$$

Thus ANFIS architecture is functionally equivalent to a Sugeno fuzzy model.

Hybrid leaning rule combines steepest decent method and least-squares estimator for fast identification of parameters in ANFIS model. For hybrid learning to be applied in a batch mode, each epoch is composed of a forward pass and a backward pass. In the forward pass, after an input vector is presented, node outputs go forward until layer 4 and consequent parameters are identified by the least squares method. In the backward pass, the error signals propagate backward and the premise parameters are updated by gradient decent. The hybrid approach converges much faster since it reduces the search space dimensions of the original pure back propagation.

2.5 Decision Making Logic

The ANFIS model is trained with targets for each of the output classes, which are well separated, and then membership functions are generated for detecting the possible range of output values. Each membership function corresponds to each of the output class; the overlapping regions between two or more classes give the possibility of existing of the particular pattern in all of the overlapped classes. But highest membership grade determines that the particular image pattern belongs to the corresponding

class. Thus to construct a boundary region for a particular class, we design a decision rule using fuzzy if-then conditions that states:
$2.5 <=$ output value $<= 7.5$, test image belongs to class-A;
$7.5 <=$ output value $<= 12.5$, test image belongs to class-B;
$12.5 <=$ output value $<= 17.5$, test image belongs to class-C.
The decision making membership function through the range of all possible output values is given in Figure-3. Each membership function is a generalized bell shaped curve, which corresponds to each output classes; the overlapped region between two or more classes gives the possibility of existing of the particular pattern in all of the overlapped classes.

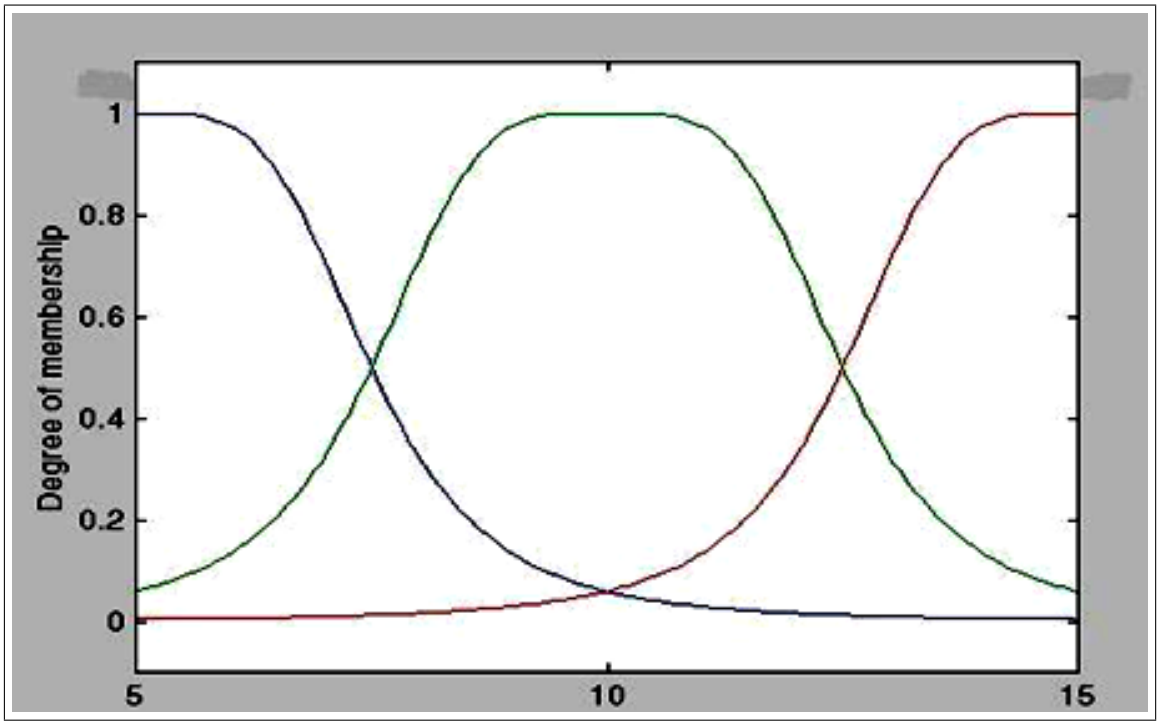


Figure 3: Output decision making membership function

3 Experimental Results

The experiment has been conducted with three distinct boundary shapes extracted from CT and MR images for section of human brain having space occupying lesions shown in Figure 4 belonging to class-A, class-B & class-C. Two membership functions are chosen for each input terminal of the network, to obtain the best possible classification result. The superiority of GA is investigated over the conventional FCM clustering technique to classify the noisy images.

3.1 Choice of String Length in GA Based Feature Subset Selection Problem

In genetic algorithm a particular string of length l contains 2^l search points. As a result, a population of size n contains some where between 2^l to $n * 2^l$ search points, depending upon the population diversity. Now among these large numbers of search points, only a few are processed in a useful manner. The reproduction, crossover and mutation operators determine the exponential growth or decay of important search points from generation to generation. It has been observed that GA with samples containing less number of bit strings which are shifted towards the enumerative search. With string length 20,

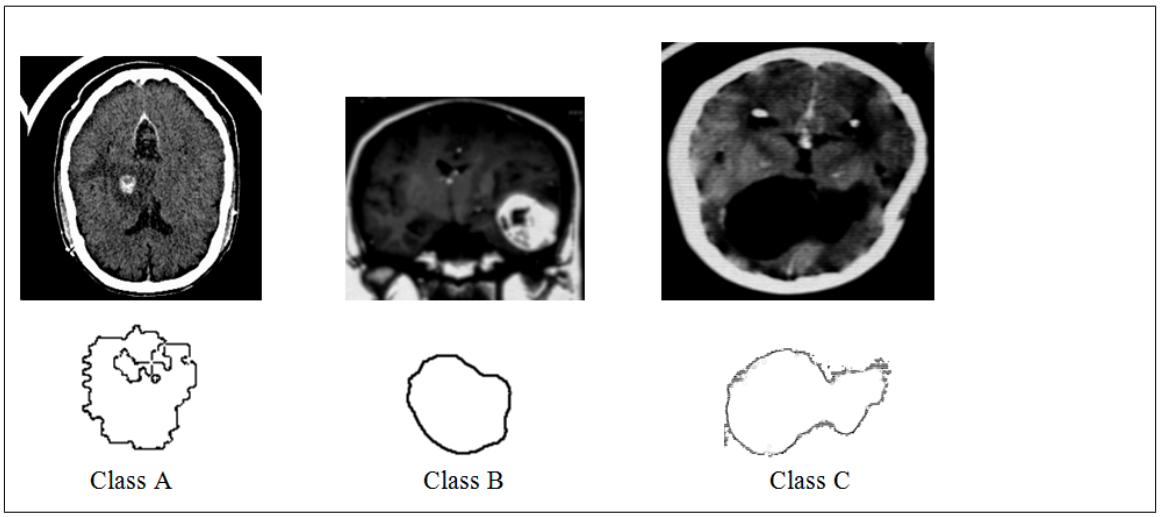


Figure 4: Boundary Features of tumors in human brain

there are a least $2^{20} = 1.04 * 10^6$ search points in the search space. Thus GA converges rapidly with the samples containing large string length. But too much increase of string length is not profitable for computational enumeration. Figure 5 shows an optimum sting length which is 20 and is acceptable for efficient feature subset selection. The variations of average and maximum values of objective function and the corresponding population size for each generation with different string lengths are shown below in Figure $5, 6, 7$ respectively.

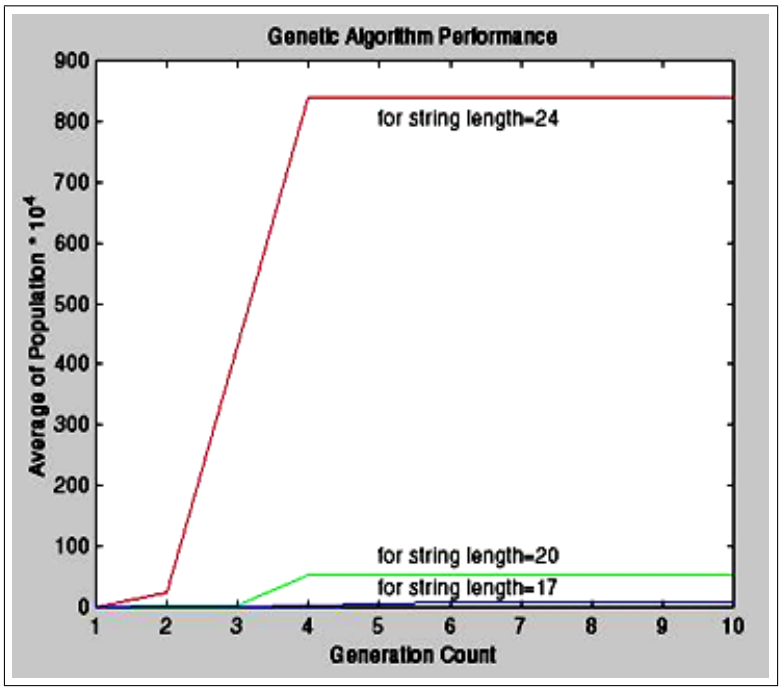


Figure 5: Variations of average value with different string length

It is also viewed from above results that GA with samples containing less number of bit string, shifted towards the enumerative search or random walk. But with string length 24 or more, there are at least $2^{24} = 1.68 * 10^7$ search points in the search space and random walk or enumeration would not be profitable. Thus GA converges rapidly with the samples containing large number of bit string.

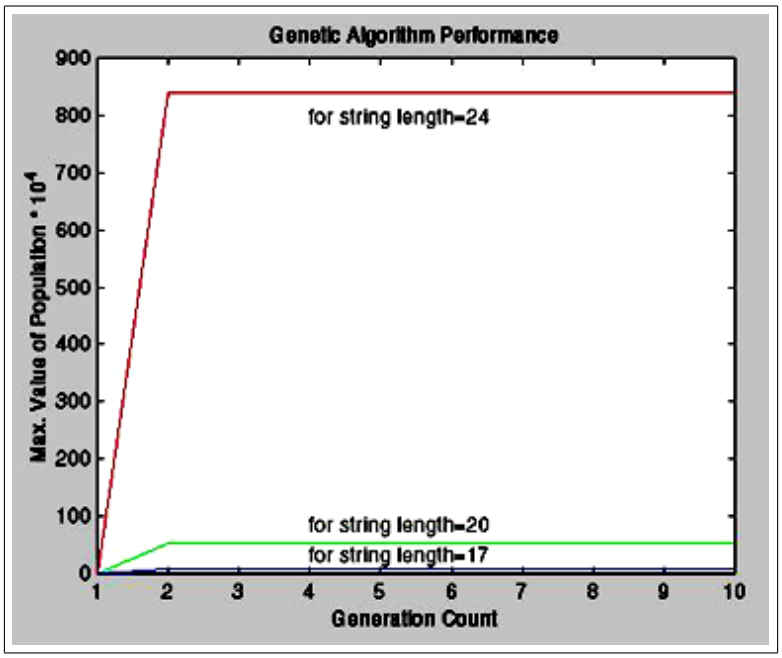


Figure 6: Variations of maximum value with different string length

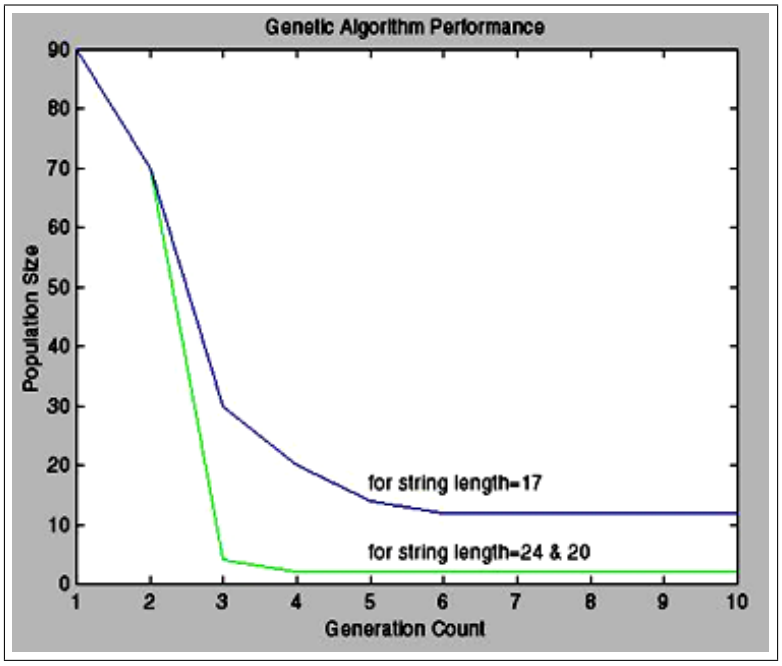


Figure 7: Variations of Population size with different String Length.

Table 1: Recognition of Distinct Noise Free Test Images Using GA and ANFIS Model

	Classification Rate		
Tested Image	Training Value	Tested Value	Decision
Boundary-1	5.0000	4.9496	Belongs to class-A.
Boundary-2	10.000	9.8162	Belongs to class-B.
Boundary-3	15.000	14.8400	Belongs to class-C.

Table 2: GA based

	Classification Rates			
Noise	Training Value	Tested Value	Classification	Decision
0.002	5.0000	5.5309	class-A	Correct
0.004	5.0000	4.5801	class-A	Correct
0.006	5.0000	5.9943	class-A	Correct
0.008	5.0000	6.6275	class-A	Correct
0.010	5.0000	5.0767	class-A	Correct
0.015	5.0000	3.0135	class-A	Correct
0.020	5.0000	8.6202	class-B	Misclassification

3.2 Results of Experiment to Recognize the Distinct Noise Free Test Images using GA Based Feature Subset Selection & ANFIS Model

ANFIS Sugeno fuzzy model is implemented to recognize the image boundary-2. The network with three significant input features and two optimum membership functions on each would result in $2^3 = 8$ fuzzy if-then rules and thus the input space is partitioned with 8 grids.

3.3 Comparative study of GA & FCM based Feature Subset Selection (FSS) Mode in presence of Noise

In the proposed model, the inputs of ANFIS network are GA based feature subset. This reduced feature subset helps to form a simple ANFIS classifier. Table-2 and Table-3 compare the classification rates of GA and FCM Based FSS model respectively for Image Boundary-1 in presence of Gaussian noise.

Table 3: FCM based

	Classification Rates			
Noise	Training Value	Tested Value	Classification	Decision
0.002	20.0000	39.6124	class-A	Correct
0.004	20.0000	83.5869	class-C	Incorrect
0.006	20.0000	33.3989	class-A	Correct
0.008	20.0000	96.9512	class-C	Incorrect
0.010	20.0000	34.6393	class-A	Correct
0.015	20.0000	92.7572	class-C	Incorrect
0.020	20.0000	91.0327	class-C	Incorrect

4 Discussions

Authors have presented a pattern recognition scheme by efficiently selecting the significant features and finally using adaptive neuro-fuzzy techniques for design of classifier. For efficient feature selection, two approaches like *Genetic Algorithm* and *Fuzzy Clustering* techniques have been implemented. Finally the classification of patterns has been achieved using adaptive neuro-fuzzy techniques. The aim of entire work is to implement the recognition scheme for classification of tumor lesions appearing in human brain as space occupying lesions identified by CT and MR images. The comparative study of GA and FCM based feature subset selection (FSS) reveals that there is a large possibility of misclassification if FCM is used for significant FSS in presence of noise. GA based FSS is resistant from noise up to a certain level and classification rate is improved for GA based FSS model. This is because, FCM has partitioned the large number shape descriptors such that the degree of association is strong for the descriptors within the same cluster and weak for the descriptors in different clusters. Genetic Algorithm (GA) searched the significant shape descriptors by applying the beauty of natural argument. Using three operators like reproduction, crossover and mutation, GA is capable to select significant feature subset.

Bibliography

[1] I. D. E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning. Reading, MA: Addison-Wesley, 1989.

[2] B. K. Fukunaga and R. R. Hayes, "Effects of sample size in classifier design," IEEE Trans. Pattern Anal. Mach. Intell., vol. 11, pp. 873-885, Aug. 1989.

[3] D'haeseleer, P. "Context preserving crossover in genetic programming"' Proc. of the 1994 IEEE World Congress on Computational Intelligence, vol. 1, pages 256-261, Orlando, FL, USA. IEEE Press, 1994.

[4] [4]. Burke, E., Gustafson, S., and Kendall, G. Diversity in genetic programming: An analysis of measures and correlation with fitness. IEEE Transactions on Evolutionary Computation, 8(1): pp. 47-62, 2004.

[5] J. Yang and V. Hanovar, "Feature subset selection using genetic algorithm", Journal of IEEE Intelligent Systems, vol. 13, pp. 44-49, 1998.

[6] S. S. Sanz, G.C .Valls, F. P. Cruz, J. S. Sanchis, C. B. Calzn, "Enhancing Genetic Feature Selection Through Restricted Search and Walsh Analysis", IEEE Trans. on Systems, Man, and Cybernetics, Vol. 34, No. 4, November 2004.

[7] P. Leray and P. Gallinari, "Feature selection with neural networks," Behaviormetrika, vol. 26, Jan. 1999.

[8] B. Hassibi and D. G. Stork, "Second order derivatives for network pruning: optimal brain surgeon," in Advances in Neural information Processing Systems, S. J. Hanson, J. D. Cowan, and C. L. Giles, Eds. San Mateo, CA: Morgan Kaufmann, 1993, vol. 5, pp. 164-171.

[9] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees, 3rd ed. London, U.K.: Chapman & Hall, 1984.

[10] T. E. Campos, I. Bloch, and R. M. Cesar Jr., "Feature selection based on fuzzy distances between clusters: first results on simulated data," Lecture Notes in Computer Science, vol. 20, no.13, pp. 186, 2001.

[11] E. Nabil; A. Badr; I. Farag; "An Immuno-Genetic Hybrid Algorithm", International Journal of Computers, Communications & Control, vol. IV, no. 4, ISSN 1841 - 9836; E-ISSN 1841-9844, 2009.

[12] Adlassnig, K. P., "Fuzzy neural network learning model for image recognition." Integrated Computer-Aided Engineering, pp. 43-55, 1982.

[13] Kim, J.S. and H. S. Cho, "A fuzzy logic and neural network approach to boundary detection for noisy images." Fuzzy Sets and Systems, pp. 141-159, 1994.

[14] Jang, J.-S.R., C.-T. Sun, E. Mizutani, "Neuro-Fuzzy and Soft Computing, A Computational Approach to Learning and Machine Intelligent" Pearson Education.

[15] C. Muńoz, F. Vargas, J. Bustos, M. Curilem, S. Salvo ; H. Miranda; "Fuzzy Logic in Genetic Regulatory Network Models", International Journal of Computers, Communications & Control, vol. IV, no. 4, ISSN 1841 - 9836; E-ISSN 1841 - 9844, 2009.

Mahua Bhattacharya, an Associate Professor of Indian Institute of Information Technology & Management, Gwalior, India is working in the area of medical image analysis more than a decade in various fields of bio - medical applications like multimodal medical image fusion and registration, mammographic image analysis, classification of tumor / cancer lesion in CNS, computational techniques for study of neuro- degeneracy in brain, study of bone degeneracy and erosion. She had her B.Tech and M.Tech degree and from the institute of Radio Physics and Electronics, University of Calcutta. She worked as a research scientist at Indian Statistical Institute, Calcutta from 1995 till 2000 Calcutta and got her Ph.D degree in the area of Multimodal Medical Image Processing and Analysis Used Knowledge Based Approach in 2001 She was recipient of Frank George award for the paper - *Cybernetic Approach To Medical Technology : Application To Cancer Screening And Other Diagnostics'* WOSC - The World Organization Of Systems & Cybernetics, UK. She has published more than 70 papers in international journals and conference proceedings and as book chapters.

Arpita Das is an Assistant Professor of Institute of Radio Physics & Electronics, University of Calcutta, India. She received her B.Tech. and M.Tech. degree in Radio Physics and Electronics, University of Calcutta, in 2004 and 2006, respectively. Presently she is pursuing her Ph.D. on 'Some Studies on Medical Image Processing Methods And Their Implementation'. She was a senior research fellow under CSIR. Her research interests include image processing, pattern recognition, soft computing approaches for biomedical applications

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 4, pp. 469-476

Hierarchical and Reweighting Cluster Kernels for Semi-Supervised Learning

Z. Bodó, L. Csató

Zalán Bodó, Lehel Csató
Department of Mathematics and Computer Science
Babeş–Bolyai University
Kogălniceanu 1, 400084 Cluj-Napoca, Romania
E-mail: {zbodo, lehel.csato}@cs.ubbcluj.ro

Abstract: Abstract: Recently semi-supervised methods gained increasing attention and many novel semi-supervised learning algorithms have been proposed. These methods exploit the information contained in the usually large unlabeled data set in order to improve classification or generalization performance. Using data-dependent kernels for kernel machines one can build semi-supervised classifiers by building the kernel in such a way that feature space dot products incorporate the structure of the data set. In this paper we propose two such methods: one using specific hierarchical clustering, and another kernel for reweighting an arbitrary base kernel taking into account the cluster structure of the data.
Keywords: Kernel methods, semi-supervised learning, clustering

1 Introduction

Extracting information from large data collections is an important research topic in mathematical modeling: it helps designing automated inference procedures with limited or no user intervention [9]. The resulting algorithms are used in various domains like bioinformatics or natural language processing, both involving the processing of large data sets. Data sets are usually labeled; this manual labeling is done before the automated information extraction procedure takes place. The limitation of the procedure is that the *total number* of items cannot be labeled. In this scenario the semi-supervised learning (SSL) develops methods that handle partially labeled data sets where only a small portion has labels, the rest of it is collected but unlabeled. Since unlabeled data is ubiquitous, in semi-supervised learning we jointly handle both the labeled and the unlabeled parts to improve the performance of the algorithm. In the following we only consider semi-supervised classification, *i.e.* those methods that assign single or multiple labels to a given input.

To use the unlabeled part of the data set, some assumptions have to be made [5]: (i) smoothness assumption, (ii) cluster assumption, (iii) manifold assumption. Most SSL methods are built on the top of the supervised algorithms by using these assumptions together with estimates of the input distribution. We use the input distribution to define a change in the input metric, leading to a modified distance between items. We study SSL methods comprising the following two steps: first we determine the new distance – dot product or kernel function – between the learning examples, and in the second step with a supervised method we obtain the decision function by using the new distance obtained in the first step.

In this paper we focus on methods that exploit the induced changes in distances and characterize this induced distance measure with kernels [3]. Kernel methods constitute a powerful tool to rewrite a linear algorithm into a non-linear one. They are based on a symmetric positive semi-definite (kernel) function, that is a dot product in a high-dimensional space [10]. The kernel in the first step of the generic SSL method mentioned above is *data-dependent*. Data-dependent kernels combine kernel algorithms and semi-supervised learning by providing a new representation for the examples that uses both the labeled

and the unlabeled parts of the data set. A formal definition of the data-dependent kernel is the following: if $D_1 \neq D_2$,

$$k(\mathbf{x}_1, \mathbf{x}_2; D_1) \eqslantless k(\mathbf{x}_1, \mathbf{x}_2; D_2)$$

where "$\eqslantless$" reads as "not necessarily equal" and the semicolons denote conditioning. It is important that these kernel functions are conditioned on the data sets, nevertheless we omit it in the following: it will be clear from the context if a kernel is data-dependent.

We propose two data-dependent kernels in this paper: (i) a kernel using the distances induced by hierarchically clustering the labeled and unlabeled data; (ii) a reweighting kernel based on the Hadamard product of some base kernel matrices.

The paper is structured as follows: Section 2 outlines the notations, in Section 3 the hierarchical and graph-based hierarchical cluster kernels for semi-supervised classification are described. In Section 4 we introduce the reweighting kernels based on data clustering and kernel combination. In Section 5 we present the experiments and results and in Section 6 we present the conclusions and discussions on the proposed kernels.

2 Notation

We denote with $D = \{(\mathbf{x}_i, y_i) \,|\, i = 1, 2, \ldots, \ell\} \cup \{\mathbf{x}_i \,|\, i = \ell + 1, \ldots, \ell + u\}$ the training data, with the first set being the labeled and the second the unlabeled data set. We further assume $\mathbf{x}_i \in X$ with X metric space, $y_i \in Y$, and the set of labels is of finite cardinality, *i.e.* $|Y| < \infty$. A key assumption is that the size of labeled data is much smaller than the available unlabeled part, *i.e.* $\ell \ll u$. In the paper N denotes the size of the entire training set, $N = \ell + u$. We use the scalar K to denote the number of clusters, where needed. Boldface lowercase letters denote vectors, boldface capitals are matrices, all other variables are scalars; $\mathbf{A}'$ denotes the transpose. For a matrix $\mathbf{A}$, A_{ij} is its element in the i-th row and j-column, and $\mathbf{A}_{i\cdot}$ and $\mathbf{A}_{\cdot j}$ denote the vectors corresponding to the i-th row and j-th column.

3 Hierarchical cluster kernels

In this section we introduce the proposed hierarchical cluster kernels. We propose the use of distances induced by different clustering algorithms instead of the original distance measure in the input space. If unlabeled data is added to the relatively small labeled data set, we expect that the *new* distance, obtained via clustering and the use of unlabeled data, induces a better representational space for classification. For clustering we use special hierarchical clustering techniques – the ones that result in ultrametric distance matrices – leading to positive semi-definite kernel matrices.

Our method is based on the connectivity kernel [8] and we extend on this kernel construction by involving the unlabeled data and allowing any hierarchical clustering method leading to ultrametric distance matrices. For data sets where the manifold assumption is expected to hold, we construct the hierarchical cluster kernel using distances induced by the k-NN and ε-NN data graphs.

3.1 Hierarchical clustering and ultrametricity

Hierarchical clustering builds a tree in successive steps, where the nodes of the tree represent nested partitions of the data, in contrary to partitional clustering methods, which result in a single partition. For the proposed hierarchical cluster kernel we use special agglomerative clustering methods. To fully specify a hierarchical clustering algorithm, cluster similarities have to be measured; these are called *linkage distances*. Based on the choice of the linkage distance measuring distances between clusters in agglomerative clustering, one can design a large variety of clustering methods.

$$\text{single linkage:} \quad D(C_1,C_2) = \min\{d(\mathbf{x}_1,\mathbf{x}_2) \mid \mathbf{x}_1 \in C_1, \mathbf{x}_2 \in C_2\} \tag{1}$$

$$\text{complete linkage:} \quad D(C_1,C_2) = \max\{d(\mathbf{x}_1,\mathbf{x}_2) \mid \mathbf{x}_1 \in C_1, \mathbf{x}_2 \in C_2\} \tag{2}$$

$$\text{average linkage:} \quad D(C_1,C_2) = \frac{1}{|C_1||C_2|} \sum_{\mathbf{x}_{1i} \in C_1} \sum_{\mathbf{x}_{2j} \in C_2} d(\mathbf{x}_{1i},\mathbf{x}_{2j}) \tag{3}$$

The linkage distances are based on $d(\mathbf{x}_1,\mathbf{x}_2)$, the *pointwise distance* in the input space that usually is the Euclidean distance $d(\mathbf{x}_1,\mathbf{x}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|_2$.

In this paper we experiment only with the three linkage distances presented above; for a detailed discussion of these and other distances see [7]. All these three methods lead to *ultrametric* hierarchical clustering. The property is used for constructing positive semi-definite kernels from clusters: suppose that we choose to merge three clusters, C_1, C_2 and C_3 in the following order: we first merge C_1 with C_2 resulting in C_{12}, and then we merge it with C_3. Now if

$$\left.\begin{array}{l} D(C_1,C_2) \leq D(C_1,C_3) \\ \text{and} \\ D(C_2,C_1) \leq D(C_2,C_3) \end{array}\right\} \quad \text{then} \quad D(C_1,C_2) \leq D(C_{12},C_3) \tag{4}$$

Based on an agglomerative clustering method that uses ultrametric linkage distance, we can define an ultrametric distance matrix, based on that a kernel function that can be used for a better representation.

3.2 The connectivity kernel

Our method of constructing hierarchical cluster kernels is based on [8]. The authors propose a two-step clustering: map the points to a new representational space based on the effective dissimilarities, and cluster them using the new representation.

The method is an approximation to pairwise clustering. To compute the effective dissimilarities used in pairwise clustering, the authors build a graph of the data; they assume that on the path between two points belonging to different clusters there will be an edge with large weight, representing the *weakest* link on the path. The effective dissimilarity will be represented by this value. They approximate the effective dissimilarities using a Kruskal-style algorithm [8].

Our method can be viewed as a generalization of the connectivity kernel, since if the ultrametric property is satisfied, we can use an arbitrary linkage distance when performing the agglomerative clustering. Moreover we propose to use the kernel in semi-supervised learning settings, when only a small portion of the data labels is known, and we propose a manifold-based extension of the kernel too.

3.3 Constructing the kernel

The hierarchical clustering results in a dendrogram, whose nodes are labeled with the distance between the clusters that were merged at the respective node. We build a distance matrix by taking the label attached to the lowest common ancestor of the points in the tree. In order to transform distances to dot products we use a method similar to multi-dimensional scaling (MDS) [2]:

$$\mathbf{K} = -\frac{1}{2}\mathbf{JMJ} \qquad \text{with} \qquad \mathbf{J} = \mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}'$$

where $\mathbf{M}$ contains the squared distances based on the dendrogram and $\mathbf{J}$ is the centering matrix built from the identity matrix $\mathbf{I}$ and the tensor product of the vector with all elements 1. The resulting matrix contains the dot products between a set of vectors $\{\mathbf{z}_i\}_{i=1}^N$ with squared Euclidean distances $\|\mathbf{z}_i - \mathbf{z}_j\|_2^2 = M_{ij}$ [8].

In what follows, we construct the cluster kernel using linkage distances from hierarchical clustering. Thus we map the points to a feature space where the pointwise distances are equal to the cluster distances in the input space. The steps are shown in Algorithm 1.

Algorithm 1 Hierarchical cluster kernel

1: Perform an agglomerative clustering with ultrametric linkage distances from Section 3.1 on the labeled and unlabeled data.
2: Define matrix $\mathbf{M}$ with $M_{ij} =$ linkage distances of $\mathbf{x}_i$ and $\mathbf{x}_j$; $M_{ii} = 0$.
3: Define the kernel matrix as $\mathbf{K} = -\frac{1}{2}\mathbf{JMJ}$.

The resulting kernel $\mathbf{K}$ is obviously data-dependent. We use the unlabeled data in the clustering step to determine *better* pointwise distances, leading to the kernel; we expect to obtain better similarities then using only the labeled part. It is important that in order to compute the kernel function for the test set we include them into the unlabeled set. This means that if the test point is unavailable at training time, the whole clustering process should be repeated, slowing down the classification. To efficiently compute the kernel for unseen points is left as a future realization and is discussed in Section 6.

3.4 Hierarchical cluster kernel with graph distances

In building the hierarchical cluster kernel, we only used the cluster assumption. Here we extend the above kernel to also exploit the manifold assumption, mentioned in Section 1, using a graph-based hierarchical cluster kernel. We approximate distances by using shortest paths based on k-NN or ε-NN graphs, similar to ISOMAP [11]. In this process we substitute the graph distances for the pointwise distances $d(\cdot,\cdot)$. The result is that the hierarchical clustering algorithm is preceded with the steps shown in Algorithm 2.

Algorithm 2 Graph-based hierarchical cluster kernel

-2: Determine the k-nearest neighbors or an ε-neighborhood of each point, and let distances to other points equal to ∞.
-1: Compute shortest paths for every pair of points – using for example Dijkstra's algorithm.
0: Use these distances for the pointwise distance in eqs. (1), (2), or (3).

We deliberately started the numbering from -2 to emphasize that these steps *precede* the algorithm from the previous subsection. We emphasize that these steps are optional: should only be used if the manifold assumption holds on the data set.

We use here the shortest path distances computed on the k-nearest neighbor or the ε-neighborhood graph of the data, thus – if the data lie on a low-dimensional manifold – approximating pointwise distances on this manifold.

The graph built from the k-nearest or the ε-neighborhoods may contain several disconnected components, for example if k or ε is too small. If this scenario happens we use a method similar to the one described in [13] for building a connected graph.

4 Reweighting kernels

Combining kernels to improve classification performance was thoroughly studied [10]. In the following – based on kernel combination – we propose three techniques to reweight a base kernel using the cluster assumption of semi-supervised learning.

We make use of the following properties: for any $\mathbf{K}_1$ and $\mathbf{K}_2$ positive semi-definite matrix and any positive scalar value $a > 0$, the following combinations are positive semi-definite matrices:

$$\mathbf{K}_1 + \mathbf{K}_2, \quad a\mathbf{K}_1, \quad \mathbf{K}_1 \odot \mathbf{K}_2,$$

where $\odot$ denotes the Hadamard, or direct product. We develop techniques that reweight the kernel matrix by exploiting the cluster structure of the training data. Thus, if two points are in the same cluster, their similarity obtains a high weight, while lying in different clusters induces a lower weight; the resulting kernel is called the *reweighting kernel*, or $k_{\mathrm{rw}}(\mathbf{x}_1, \mathbf{x}_2)$. The similarity weights are combined with the values of the base kernel $k_{\mathrm{b}}(\mathbf{x}_1, \mathbf{x}_2)$, thus forming the final kernel matrix. To sum up, the new cluster kernel is

$$k(\mathbf{x}_1, \mathbf{x}_2) = k_{\mathrm{rw}}(\mathbf{x}_1, \mathbf{x}_2)\, k_{\mathrm{b}}(\mathbf{x}_1, \mathbf{x}_2)$$

where $k_{\mathrm{rw}}(\cdot, \cdot)$ is the reweighting and $k_b(\cdot, \cdot)$ is the base kernel. In matrix form it can be written as

$$\mathbf{K} = \mathbf{K}_{\mathrm{rw}} \odot \mathbf{K}_{\mathrm{b}}$$

We are faced with two problems in the construction of the above cluster kernel: (i) the reweighting kernel must be positive semi-definite, (ii) the base kernel matrix has to be positive semi-definite and *positive*. The first requirement is obvious: it is needed to guarantee the positive semi-definiteness of the resulting kernel. The second condition is crucial, since for negative values in the base kernel matrix a quite different reweighting should be performed. To avoid complications due to negativity, we require a positive base kernel, $k_b(\mathbf{x}_1, \mathbf{x}_2) \geq 0$.

The bagged cluster kernel, proposed in [12], reweights the base kernel values by the probability that the points belong to the same cluster. For computing this probability the bagged cluster kernel uses k-means clustering, together with its property that the choice of the initial cluster centers highly affects the output of the algorithm. Assuming we have N data items and K clusters, the kernel is constructed by running k-means T times, each time with different initialization resulting in different cluserings. The resulting kernel is called the bagged kernel. The final cluster kernel is the Hadamard product of the base kernel and the bagged kernel.

Borrowing the underlying idea of the bagged cluster kernel in the following we develop reweigthing kernels based on various clustering algorithms.

4.1 Gaussian reweighting kernel

Suppose that we are given the output of a clustering algorithm, the cluster membership matrix $\mathbf{U}$ of size $K \times N$. We assume that two points belong to the same cluster(s) if their cluster membership vectors are similar or close to each other. We define similarity via the Gaussian kernel in the following way:

$$k_{\mathrm{rw}}(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\frac{\|\mathbf{U}_{\cdot \mathbf{x}_1} - \mathbf{U}_{\cdot \mathbf{x}_2}\|^2}{2\sigma^2}\right) \tag{5}$$

where $\mathbf{U}_{\cdot \mathbf{x}}$ denotes the cluster membership vector of point $\mathbf{x}$, *i.e.* the column of $\mathbf{x}$ in $\mathbf{U}$. We know that the resulting matrix is positive semi-definite [10] with each element between 0 and 1. In this case the parameter σ defines the amount of separation between similar and dissimilar points: if σ is large, the gap between the values expressing similarity and dissimilarity becomes smaller, while for smaller σ these values get farther from each other.

4.2 Dot product-based reweighting kernels

Another possibility of using the cluster membership vectors is to define the following reweighting kernel:

$$\mathbf{K}_{\mathrm{rw}} = \mathbf{U}'\mathbf{U} + \frac{\alpha}{N}\mathbf{1}\mathbf{1}' \tag{6}$$

where $\mathbf{U}$ denotes the cluster membership matrix and $\alpha \in [0, 1)$. The first term scores point similarity according to the cluster memberships, and the second term is used to avoid zero similarities: if two membership vectors are orthogonal, leading to a zero in the dot product matrix. We may assume that the obtained clustering is not too *confident*, *i.e.* we should use a small value α, the crisp cluster membership is thus alleviated with the term $(\alpha/N)\mathbf{1}\mathbf{1}'$.

Shoving that the reweighting kernel from equation (6) is positive semi-definite is straightforward: both the first and the second term is an external product, thus positive semi-definite, and owing to the properties defined in Section 4, results that the kernel is indeed positive semi-definite.

Another version of the kernel in (6) is

$$\mathbf{K}_{\text{rw}} = \beta\, \mathbf{U}'\mathbf{U} + \frac{1}{N}\mathbf{1}\mathbf{1}' \tag{7}$$

where $\beta \in (0, \infty)$. Here the kernel values for which the dot product matrix of cluster membership vectors correspond to zero, by $(1/N)\mathbf{1}\mathbf{1}'$, remain the same, however if the points lie in the same cluster $\beta\mathbf{U}'\mathbf{U}$ gives a weight greater than zero, thus this kernel value will be increased.

The above equations could clearly be merged, but we left them as separate reweighting kernels to differentiate between the underlying ideas.

5 Experiments and results

In this section we present the results obtained using our cluster kernels, and we compare it to other data-dependent kernels. For learning we used support vector machines (SVMs) [3], namely the LIBSVM (version 2.85) implementation [4]. The data sets used for evaluating the kernels were the following: USPS, Digit1, COIL2, Text. The detailed descriptions of these sets can be found in [5]. Each data set has two variations: one with 10 and one with 100 labeled data; furthermore each data set contains 12 labeled/unlabeled splits of the data. We used only the first split from each set. The columns having labels 10 and 100 in the tables showing the obtained result indicate which version of the data set was used, *i.e.* they show the number of labeled and unlabeled examples used.

Table 1 shows the accuracy results obtained using different kernels. We used accuracy as the evaluation measure, and the results are given in percentage. For each data set we indicated the best, second best and third best results obtained.

The first two rows contain the *baseline* results obtained with linear and Gaussian kernels. Principally we wanted to improve on these results. The hyperparameter for the Gaussian kernel was set using a cross-validation procedure.

The following 7 rows show the results obtained using the ISOMAP kernel [11], the neighborhood kernel [12], the bagged cluster kernel [12], the multi-type cluster kernel with different transfer functions [6] and Laplacian SVM [1], respectively.

The next 15 rows – below the second horizontal line – show the results obtained using our kernels. HCK and gHCK denote the hierarchical cluster and graph-based hierarchical cluster kernels from Section 3, respectively. RCK1, RCK2 and RCK3 denote the reweighting cluster kernels defined in equations (5), (6) and (7), respectively. Here we experimented with three clustering techniques: k-means, hierarchical and spectral clustering.

Because of lack of space we omitted the description of setting the parameters.

6 Discussion

As the results show we obtained good results with the proposed hierarchical and reweighting cluster kernels, and in many cases the results provided by our kernels are very close to the best results, as shown

	USPS		Digit1		COIL2		Text	
	10	100	10	100	10	100	10	100
linear	72.82	86.43	81.07	90.86	60.74	80.43	58.26	67.86
Gaussian	80.07	89.71	56.11	93.86	57.38	82.50	59.06	56.43
ISOMAP	85.10	86.71	**94.43**	**97.43**	62.62	80.64	59.80	72.43
neighborhood	76.31	94.14	87.11	94.21	64.43	84.43	51.68	62.79
bagged	**87.38**	92.79	**93.29**	**96.93**	**71.28**	85.57	63.29	66.14
multi-type, step	80.07	92.86	91.01	91.29	55.77	84.86	53.56	**74.79**
multi-type, linear step	80.07	92.86	91.01	91.36	55.77	84.86	53.02	**75.29**
multi-type, polynomial	80.07	80.29	48.86	65.07	54.23	82.29	50.60	56.71
LapSVM, Gaussian	81.95	**95.93**	84.50	**97.64**	**76.64**	**97.71**	**63.42**	62.50
HCK, single	80.07	81.79	48.86	70.21	67.85	**96.00**	**66.78**	**73.14**
HCK, complete	82.01	89.50	60.67	89.71	55.64	86.36	50.27	49.57
HCK, average	81.48	92.86	71.75	93.79	68.05	91.71	**64.63**	50.14
gHCK, single	80.07	81.79	48.86	70.21	60.60	**93.86**	**66.78**	**73.14**
gHCK, complete	**88.26**	**95.64**	75.50	93.71	**68.52**	88.79	56.17	67.71
gHCK, average	**89.26**	**95.64**	**94.70**	95.21	60.54	90.64	47.32	66.86
RCK1, k-means	84.45	92.98	83.14	94.28	58.55	83.76	–	–
RCK1, hierarchical	86.17	95.29	89.06	94.94	62.08	85.93	62.35	68.07
RCK1, spectral	81.43	90.87	88.32	95.20	58.22	83.83	63.26	66.93
RCK2, k-means	83.86	92.45	84.58	94.08	58.95	83.76	–	–
RCK2, hierarchical	86.11	**95.50**	89.06	95.29	62.08	85.64	61.28	71.14
RCK2, spectral	81.63	91.39	88.32	94.64	58.03	83.37	61.50	70.07
RCK3, k-means	83.66	92.59	84.13	92.96	58.60	83.28	–	–
RCK3, hierarchical	84.97	95.29	89.06	94.57	62.95	86.07	59.13	71.21
RCK3, spectral	81.16	91.56	88.32	94.73	55.83	83.20	59.26	71.00

Table 1: Accuracy results using different kernels. The results are given in percentage. For each data set the best three results were formatted in boldface.

in Table 1. Individually LapSVM outperformed every other method, possibly because of the careful selection of its parameters, but also because it is a very powerful technique.

Thus the results show that the proposed kernels for semi-supervised classification can be used for different types of data sets, and they provide better performances compared to simple, data-independent kernels, *e.g.* the Gaussian kernel. Moreover with data-dependent kernels any supervised kernel method can be easily turned into a semi-supervised method, without changing the underlying learning algorithm.

In order to compute the kernel for the test points one needs to include these points in the unlabeled data set. That is one can say that the methods resemble transductive learning, where the decision function is computed only on the points in question. Thus if a new point arrives the whole process must be repeated. To overcome this costly process approximation methods could be implied, but this is left as a future work. We also plan to develop methods or heuristics for automatically choosing the parameters of the proposed kernels.

Acknowledgments The authors acknowledge the partial support of the Romanian Ministry of Education and Research via grant PNII 11-039/2007.

Bibliography

[1] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[2] Ingwer Borg and Patrick J. F. Groenen. *Modern Multidimensional Scaling, 2nd edition*. Springer-Verlag, New York, 2005.

[3] B. E. Boser, I. Guyon, and V. N. Vapnik. A Training Algorithm for Optimal Margin Classifiers. *Computational Learning Theory*, 5:144–152, 1992.

[4] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001.

[5] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, September 2006.

[6] Olivier Chapelle, Jason Weston, and Bernhard Schölkopf. Cluster Kernels for Semi-Supervised Learning. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *NIPS*, pages 585–592. MIT Press, 2002.

[7] Richard Duda, Peter Hart, and David Stork. *Pattern Classification*. John Wiley and Sons, 2001. 0-471-05669-3.

[8] Bernd Fischer, Volker Roth, and Joachim M. Buhmann. Clustering with the Connectivity Kernel. In Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf, editors, *NIPS*. MIT Press, 2003.

[9] Imre J. Rudas and János Fodor. Intelligent systems. *Int. J. of Computers, Communication & Control*, III(Suppl. issue: Proceedings of ICCCC 2008):132–138, 2008.

[10] B. Schölkopf and A. J. Smola. *Learning with Kernels*. The MIT Press, Cambridge, MA, 2002.

[11] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, December 2000.

[12] Jason Weston, Christina Leslie, Eugene Ie, and William Stafford Noble. Semi-Supervised Protein Classification Using Cluster Kernels. In Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors, *Semi-Supervised Learning*, chapter 19, pages 343–360. MIT Press, 2006.

[13] Quan Yong and Yang Jie. Geodesic Distance for Support Vector Machines. *Acta Automatica Sinica*, 31(2):202–208, 2005.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 4, pp. 477-482

The Avatar in the Context of Intelligent Social Semantic Web

A. Braşoveanu, M. Nagy, O. Mateuţ-Petrişor, R. Urziceanu

Adrian Braşoveanu
Lucian Blaga Univeristy of Sibiu, Romania
E-mail: adrian.brasoveanu@gmail.com

Mariana Nagy
Aurel Vlaicu University of Arad, Romania
E-mail: mnagy62@yahoo.com

Oana Mateuţ-Petrişor, Ramona Urziceanu
Agora University, Oradea and R&D Agora Ltd.
Cercetare Dezvoltare Agora Oradea, Romania
E-mail: ganearamona2002@yahoo.com, oana_mateut@yahoo.com

Abstract: When the first articles about the Semantic Web (SW) appeared, there were hardly any signals that the next revolution would be related to social networking. Social networking services (SNS) have grown after MySpace, LinkedIn and Facebook were launched in 2003-2004 and combine text, images, movies, music, animations and all sorts of lists to create personal presentation pages for users, means to connect to real or virtual friends from all over the world and recommendations based on trust.
The rise of the Social Semantic Web and the convergence of different media to create rich experiences is one of the most interesting paradigm shift in the last decades because the probable effect of this movement is the fact that in one day virtual meetings will become legitimate in all aspects of our daily lives (if they are not already). The most important question in this context (the one that we try to answer in this paper) is related to how people will try to shape and use their avatars. In order to understand this, we will study the links between multimodal ontologies, affective interfaces, social data portability and other recent findings.
This paper starts with a survey of the current literature of the field, examines some social semantic web mechanisms that changed the way we think about SNs and in the end discusses some methods of connecting emotions with the social semantic web which pose some interesting questions related to the use of avatars. Between conclusions, one of the most interesting is the one that states that the use of affective interfaces adds value to the multimodal ontologies, while another suggests that the avatar must be a mediator between different technologies.
Keywords: avatar, Semantic Web (SW), social networking services (SNS), affective interfaces, Human-Computer Interaction (HCI).

1 Introduction

The semantic web (SW) has evolved into a technology we use on daily basis, sometimes without even being aware of this, but dreams like those described by World Wide Web creator Timothy Berners-Lee and his collaborators in the 2001 article [4] are still not common place. This is often the case when new technologies are presented to the general public while still in their infancy. The original article has been revised 5 years later [5] and Berners-Lee admits that we are still very far from his original vision of agents replacing humans for several tasks like buying tickets or making appointments to the doctor.

Almost a decade later, the web has changed and the information is no longer presented through classic text and pictures. Social networking services (SNS) like Facebook, LinkedIn, MySpace combine text, images, movies, music, animations and all sorts of lists to create personal presentation pages for users, different means to connect to real or virtual friends from all over the world and recommend things based on trust. These days even the way we make a search is going through a paradigm shift, because it is widely believed that the recommendations of a

circle of trusted friends are more valuable than those of a search engine. This is clearly a human response to the strategies of increasing the ranks of the pages in search engines.

As it was stated by Berners-Lee, the original vision of the Semantic Web was to "enable machines to comprehend semantic documents and data, not human speech and writings" [4]. Because of the chaotic developments of the last decade, machines will not only need to comprehend human speech and writings, but will also need to be capable to search through videos almost in the same way we do. Such capabilities go well beyond the initial meaning of the Semantic Web, imply reasoning and emotions and since they are similar to the way we think, allow us to use the term "Intelligent" before the terms "Social Semantic Web".

In this paper we will try to see how the developments in different areas like HCI, SNS or SW reshape old concepts on the fly. We have chosen the avatar in order to express some of our thoughts.

2 Rationale and Approach: Why the Avatar?

To explain the rationale behind choosing the avatar for presenting some of our ideas, we have to examine a brief history of the links between the Semantic Web and social networking. We will also need to look at the trends from these areas and HCI.

Constraints force us to think creatively, is the mantra of the agile development community (and of Ruby on Rails in particular), but also the phrase that defines the last decade in the IT industry. When we look back to the year 2001, we do not see WTC but rather a fragile industry which tried to recover from the "dot com bubble". In this climate of uncertainty when the seminal article about Semantic Web by Berners-Lee and his research group [4] was published it sparked a lot of debates and started a series of innovations which did not stop to date. In those days it was almost impossible to predict that any company related to the IT would be sold for half a billion or more, given the fact that many companies failed to bring cash to the investors. Few years later, when MySpace was sold for a considerable amount to NewsCorp and Google acquired YouTube, everybody understood that something changed in the world of IT. Somewhere between 2002 and 2005, without knowing it, the world has become social. Historically social networks were viewed as the preserve of the rich and even as a sort of social divide between the rich and the poor, but with the rise of the social networking services they became useful for everyone [26] [27]. The user centric social network, where everything is a link from or to a friend, is necessary to everyone who tries to find a job or his old friends from college these days. Social networking services added new layers not only to the social interaction but to the Semantic Web as well. Some ideas like Mika's community-based ontology extraction from Web pages [21] would have never emerged without the rising of SNS. The field of ontology was connected with IT for several decades, but the first definition dates back to Gruber's 1993 trial, an ontology being an "explicit specification of a conceptualization" [14]. This definition has been revised several times by different authors [15], but all these revisions are still based on [14]. Between 2001 and 2007 a lot of ontologies have been created with the purpose of connecting different SNS or as extensions to applications that were built around data extraction from web applications [18] [22] [24]. Rich visualization techniques [22] like those generally used in research appeared on the Web with the huge success of the Adobe Flash technology (1998-2005) and the introduction of Ajax and tag clouds (2004-2005). The use of labels for annotating different pieces of information has created a new field for knowledge representation called folksonomies [26]. The use of folksonomies is often linked with the use of ontologies since both concepts aim to offer a way to retrieve information. The folksonomy will give us some ideas about the most valuable words for a certain group of users, while the ontology will also try to model the relations between different topics (as we can see in Figure 1). The various methods of bridging folksonomies and ontologies to enable better knowledge representation are presented in [18].

The links between social networking and Semantic Web are discussed in Berners-Lee revision of [4] in the 2006^{th} article [5] and in [6], while the history of the SNS is examined in [8]. When it comes to SNS we usually agree with the chronology proposed by Boyd in [8], but when we deal with bridging SNS and SW we propose a simple timeline:

- 2001 - 2004: The first attempts to link SW and SNS;
- 2004 - 2007: The explosion of SNS and the first important results in bridging SW and SNS like Katrina PeopleFinder [21];
- 2008 - Present Day: SNS are now present in all aspects of our lives and research is focused on rather more advanced topics like affective interfaces.

Perhaps the most interesting conclusion from the early years of research regarding the links between SNS and the SW (2001-2007) belongs to Mika from his most cited article from [21]:"It seems that ontologies are us: inseparable from the context of the community in which they are created and used". The research in recent years

Figure 1: a) FOAF Ontology visualization with Welkin; b) tag cloud for Facebook.

(2008 - 2010) has rendered Mika's thesis as true. Some interesting problems related to social networking (similarity problems usually phrased like "find all users which share certain interests"; semantic concept clustering and many others) have been modeled on various datasets (from small predefined datasets selected for special topics to large online datasets), but the emergence of the semantic web tools on large scale, mostly in the last six years (in the same time with the sudden growth of interest in the area of SNS), enables us to look at them from new perspectives.

Our approach to the subject consists in making connections between different recent findings in the fields of SNS and SW. The most interesting trends in social software (2008-2009) we investigated are social data portability [7] [23], live social semantics [1], mass interpersonal persuasion also known as MIP [15], and the growth of interest surrounding affective interfaces [10] [19].

Social data portability is a classic problem in the field of social software since its inception. It might have started as something which was interesting only for people with large social networks in real life, but as soon as the companies discovered that SNS had the potential of attracting clients with clear profiles, something that TV or written media does not guarantee, it became the norm to be present on all the major SNS. The growing number of social networking sites in different niches is a problem for the users or companies interested in more than one field of knowledge who want to share their data. Porting data between different SNS or platforms like Facebook and OpenSocial or defining standards for presenting the articles from magazines or blogs is the work of researchers like Uldis Bojars and John G. Breslin from DERI, Ireland, authors of the SIOC Ontology [7].

Live social semantics is a concept proposed by a team of researchers from UK, France and Italy lead by Harith Alani [1]. The main idea is to "integrate data from the semantic web, online social networks, and a real-world contact sensing platform" [1]. Their system which integrated the Social Semantic Web mechanisms with the real world was tested at ESWC 09. Live Social Semantics is something that benefits when being linked with Social Data Portability, because in the same room we will have people connected to more than one SNS in most of the cases.

Mass interpersonal persuasion or MIP is the generic name given to those techniques of persuading millions of people to join to a certain initiative [12]. Before Facebook gathering millions of people was not an easy task, therefore it is not a surprise that in this case the theory appeared only after several applications succeeded. Its creator, BJ Fogg is the same man who introduced captology and coined the term persuasive technology [13]. The main idea behind MIP is that we should design new systems with MIP and the huge social graph in mind if we want to change something in this world. When it comes to MIP, the technology, the topic or the creator's initial intent does not matter much. What is important is that MIP gives to a creator the possibility of reaching a wider than expected audience. Fogg stated that: "we are at the start of a revolution in how individuals and cultures make decisions and take action" [13].

Affective interfaces have been proposed since 1980s but recently there has been a renowned interest. Beyond Fogg's PhD work and his seminal book [13], today affective computing is sometimes related to the Semantic Web. Between the most interesting concepts is the use of multimodal ontologies for describing emotions [10] [19]. The use of affective interfaces adds value to the multimodal ontologies because they improve communicability. Other interesting topic in the field is the way in which emotional experiences from SNS affect us in ordinary life as presented in [22] and [24]. While work in 3D graphics is not necessarily connected with the Semantic Web, when it comes to affective interfaces, the research related to the animation of the emotional facial expressions [25] can very easily be combined with the ontologies for describing emotions [10] [19]. The rise of the Social Semantic

Web and the convergence of different media to create rich experiences is one of the most interesting paradigm shifts in the last decades. One of the effects of this movement is the fact that in one day virtual meetings will become legitimate in all aspects of our daily lives (if they are not already). In such a context the role of the avatar is to replace a human being, as it was supposed to be, but the trends we have examined also suggest new approaches to the concept of the avatar and will also enable us to get closer to the visions described in [4].

3 The Avatar in the Intelligent Social Semantic Web

For the purpose of this paper we use the meaning of the term avatar which represents an agent that is a double for a real person, a double that takes care of our social self, its virtual ego [20]. It can very well be even a pseudo-avatar, not necessarily as it is viewed in Berners-Lee's paper [4]. It can represent a person, an organization or a fictional character that needs to be in the social space and it can also have a graphical representation be it 2D or 3D.

The first question that comes in mind is clear: What is the purpose of the avatar in the Intelligent Social Semantic Web at which all of the concepts presented in the previous section are aimed toward? Are there any links between those concepts?

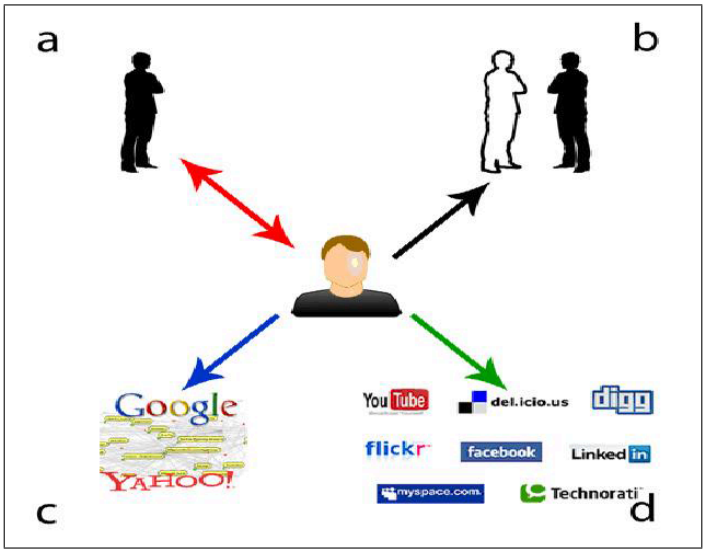


Figure 2: Some of the functions of the avatar in the Intelligent Social Semantic Web: a) maintain communication with its human counterpart (or organization); b) act as replacement for the human in virtual meetings; c) use the web, ontologies and other sources to find out news about the master's fields of interest; d) update the social network or the sites of its human counterpart.

One link was already shown in the previous section by connecting Social Data Portability and the Live Social Semantics. Another one is the idea that only emotional agents are believable [13], proposed by Fogg. But why would we need an avatar that is an emotional being? Even though the work presented in [11] [12] [16] [18] [25] is good for avatars that represent human beings, which are emotional in the real world, it is debatable if the same can be applied to organizations. Advertising is often misleading because a firm only uses emotions to sell products - to produce emotional reactions which can lead to the decision of choosing a certain brand - not because it really has emotions. People that work for a certain organization do have emotions, but the organization itself does not. There should be clear differences between the avatars that represent organizations and the avatars that represent human beings, but these are not in our sights for the moment. An avatar that represents an organization uses SNS to achieve MIP. Without MIP and the targeted advertising social networking has no value for organizations. The main philosophical problem that arises when the avatars try to use MIP is related to trust and it is expressed very well in [3]: "Whereas trust is generic to human communication and implies evaluative aspects, social presence is aiming at mediated communication and is more descriptive by nature".

We can build different associations between the research areas we have mentioned in the previous section, but it should be enough to analyze the parts of the expression from the title (Intelligent Social Semantic Web) to

understand the purpose of the avatar in this medium. SNS represent the Social part of the equation, while the SW is represented through ontologies and their applications, so it should be clear that the avatar should be a part of the Intelligence. In fact it is in the same time a part of the Intelligent side of the equation, as well as of the Social part (if we limit ourselves to the meaning of the avatar in the current SNS like Facebook, MySpace, LinkedIn).

For each major field of interest of a human person or organization several ontologies already exist or will be developed. The main problem that an avatar will face will be to wisely choose those ontologies or even perform ontology matching [11] [16] and use them to extract the meaningful data from the web (like in [28]) or create content that could help us (humans or organizations) to fulfill our objectives. It has to work for us when we sleep and alert us when something critical for our activity happens. The avatar will be credible only if it will be emotional, because it's not easy to wake up a man at 3:00 am and tell him that in other part of the world something happened and it will change his life. In any other way it would be impossible to have any impact in an open dynamic environment [2]. We might also need to change the way we design socio-technical systems [9] in order to enable the avatars to automate different tasks. When we will have this, the vision from [4] will be closer to us than ever, if not reality [17].

4 Conclusions and Future Work

Predicting the future is not an easy task as we have seen. Any technology needs several iterations before achieving its goal so we should not be surprised that it will take some time until the results of our work will be implemented and validated.

The avatar of the future will have some difficult tasks to solve (like choosing the proper ontologies) if we are to benefit from its use. It will also need to have emotions if we want it to be believable because the use of affective interfaces improves communicability. The problem of differentiating between the avatars of the real persons and the avatars of the organizations will remain an open problem until the use of avatars will be the subject of standards committees or international law.

The future work will consider implementing new mechanisms for linking the multimodal ontologies and affective interfaces with recent research in Semantic Web and HCI in a 3 years interval (during the PhD studies of the first author). The objectives are to be fulfilled involving European teams of researchers interested in this kind of projects.

Acknowledgements

This work was partially supported by the strategic grant POSDRUI88I1.5ISI60370 (2009) on "Doctoral Scholarships" of the Ministry of Labour, Family and Social Protection, Romania, co- financed by the European Social Fund - Investing in People.

Bibliography

[1] H. Alani, M. Szomszor, C. Cattuto, W. Van den Broeck, G. Correndo, A. Barrat. *Live Social Semantics*. In: 8th International Semantic Web Conference (ISWC), October 2009, US, 2009.

[2] I. Dzitac, B.E. Bărbat. *Artificial Intelligence + Distributed Systems = Agents*. International Journal of Computers, Communications & Control, IV, 1, 17-26, (http:llwww.britannica.comlbpsl additionalcontentl18l36182542lArtificial-Intelligence-Distributed-Systems-Agents), 2009.

[3] G. Bente, S. Ruggenberg, N. C. Kramer, F. Eschenburg. *Avatar-Mediated Networking: Increasing Social Presence and Interpersonal Trust in Net-Based Collaborations*. Human Communication Research 34 (2008) 287-318.

[4] T. Berners-Lee, J. Hendler, O. Lassila. *The Semantic Web*. Scientific American, May 2001, pp. 34-43.

[5] N. Shadbolt, W. Hall, T. Berners-Lee. *The Semantic Web revisited*. IEEE Intelligent Systems, pages 96- 101, MaylJune 2006.

[6] T. Berners-Lee, W. Hall, J.A. Hendler, K. O'Hara, N. Shadbolt, D.J. Weitzner. *A Framework for Web Science*. Foundations and Trends in Web Science, 1 (1), pages 1-130, 2006.

[7] U. Bojārs, A. Passant, J.G. Breslin, S. Decker. *Social Networks and Data Portability Using Semantic Web Technologies*. The 2nd Workshop on Social Aspects of the Web (SAW 2008) at the 11th International Conference on Business Information Systems (BIS 2008), Innsbruck, Austria, May 2008.

[8] D. M. Boyd, N. B. Ellison. *Social network sites: Definition, history, and scholarship*. In Journal of Computer-Mediated Communication, 13(1). http:lljcmc.indiana.edulvol13lissue1lboyd.ellison.html, 2007.

[9] V. Bryl, P. Giorgini, and J. Mylopoulos. *Designing socio-technical systems: From stakeholder goals to social networks*. Requirements Engineering, 14(1):47-70, 2009.

[10] I. Cearreta, J. M. Lopez, N. Garay-Vitoria. *Modelling multimodal context-aware affective interaction.* Proceedings of the Doctoral Consortium of the Second international conference on ACII'07. Lisbon, Portugal. Pages 57-64, 2007.

[11] J. Euzenat, P. Shvaiko. *Ontology Matching*, Springer, 2007

[12] B.J. Fogg. *Persuasive Technology*. Morgan Kaufmann, San Francisco, 2003.

[13] B.J. Fogg. *Mass interpersonal persuasion:An early view of a new phenomenon*. In H.Oinas- Kukkonen et al. (Eds.). Persuasive 2008, LNCS 5033 (pp.23-34). New York, Springer, 2008.

[14] T. R. Gruber. *A Translation Approach to Portable Ontologies*. Knowledge Acquisition, 5(2):199- 220, 1993.

[15] N. Guarino, D. Oberle, S. Staab. *What is an Ontology?* In S. Staab and R. Studer (eds.), Handbook on Ontologies, Second Edition. International handbooks on information systems. Springer Verlag: 1-17, 2009.

[16] Harth, S. Kinsella, S. Decker. *Using Naming Authority to Rank Data and Ontologies for Web Search*. In Proc. International Semantic Web Conference, ISWC'09, Washington, USA, October 2009, 2009

[17] D. J. Lewis. *Intelligent agents and the Semantic Web. Developing an intelligent Web*. Retrieved from http:llwww.ibm.comldeveloperworkslwebllibrarylwa-intelligentagel. 2008. Accessed: December 2009.

[18] F. Limpens, F.Gandon, and M. Buffa. *Linking folksonomies and ontologies for supporting knowledge sharing: a state of the art*. Technical report, EU Project, ISICIL, 2009.

[19] J. M. Lopez, R. Gil, R. Garcia, I. Cearreta, N. Garay. *Towards an Ontology for Describing Emotions*. WSKS (1) 2008: 96-104.

[20] P. Messinger, X. Ge, E. Stroulia, K. Lyons, K. Smirnov, M. Bone. *On the relationship between my avatar and myself*. Journal of Virtual Worlds Research 1 (2). http:lljournals.tdl.orgljvwrlarticlelviewl3521. (Accessed December 2009)

[21] P. Mika. *Social Networks and The Semantic Web*, Springer, 2007

[22] D. Petrelli, S. Mazumdar, A-S. Dadzie, F. Ciravegna. *Multivisualization and Dynamic Query for Effective Exploration of Semantic Data*. In Proc. International Semantic Web Conference, ISWC'09, Washington, USA, October 2009, 2009.

[23] L. Răzmeriţa, M. Jusevičius, Rokas Firantas. *New Generation of Social Networks Based on Semantic Web Technologies: the Importance of Social Data Portability*. In: Workshop on Adaptation and Personalization for Web 2.0, UMAP'09, June 22-26, 2009.

[24] C. Sas, A. Dix, J. Hart, S. Ronghui. *Emotional Experience on Facebook Site*. In: CHI '09: CHI '09 Extended Abstracts on Human factors in Computing Systems, 4-9 April 2009, Boston, MA.

[25] R.J.S. Sloan, M. Cook, B. Robinson. *Considerations for believable emotional facial expression animation*. 2nd International Conference on Visualization, Barcelona, Spain, 2009.

[26] T. VanDerWall. *Folksonomy Coinage and Definition*. 2007. Retrieved from http:llvanderwal.netlfolksonomy.html. Accessed: December 2009.

[27] M. Webb. 2004. *On Social Software*. http:llinterconnected.orglhomel2004l04l28l.

[28] S.Y. Yang. *Developing an Ontological FAQ System with FAQ Processing and Ranking Techniques for Ubiquitous Services*. Proc. of The First IEEE International Conference on Ubi-media Computing, Lanzhou, China, 2008, pp. 541-546.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 4, pp. 483-489

Stream Ciphers Analysis Methods

D. Bucerzan, M. Crăciun, V. Chiş, C. Raţiu

Dominic Bucerzan, Mihaela Crăciun, Violeta Chiş
"Aurel Vlaicu" University of Arad
Faculty of Exact Sciences
Department of Mathematics-Informatics
România, 310330 Arad, 2 Elena Drăgoi
E-mail: dominic@bbcomputer.ro, qbt@rdslink.ro, viochis@yahoo.com

Crina Raţiu
DARAMEC srl, Arad
România, Sofronea FN
E-mail: ratiu__anina@yahoo.com

Abstract: The purpose of this paper is to present and to discuss analysis methods applied in symmetric cryptography, especially on stream ciphers. The tests were made on some algorithms and also on the personal symmetric cryptographic algorithm, HENKOS, based on a pseudorandom number generator. The test confirms that the algorithm appears to be secure and fast. The paper describes first the main parts of the cryptosystem, its implementation and different analysis methods. The code is written in the C/C++ language. The software application and the tests applied were processed on a PC computer. The quality analysis presents the results of many classical statistical tests, comparing some algorithms based especially on pseudo random number generators. The tests use standard sequence of 12.5 MB resulted from some test generators. The main part of the work presents selected results for the most important statistical tests like: FIPS 1401, FIPS 1402 , ENT tests, Diehard battery of tests, NIST Statistical Test Suite. The final question is: are these tests enough to certifie the quality of a tested algorithm?
Keywords: stream cipher, synchronous stream cipher, pseudorandom number generator (PRNG), performance analysis, statistical tests.

1 Introduction

Stream ciphers are an important class of encryption algorithms. They encrypt individual characters (usually binary digits) of a plaintext message one at a time, using an encryption transformation which varies with time. Various design methods where proposed for stream ciphers and the specialists proposed many analysis methods. However, the reality is that in the literature we can find relatively few fully-specified stream cipher algorithms. One possible explanation can be the fact that many stream ciphers used in practice tend to be proprietary and confidential.

A stream cipher generates what is called a keystream (a sequence of bits used as a key). Encryption is accomplished by a simple operation combining the keystream with the plaintext, usually with the bitwise XOR operation. Stream ciphers can be either symmetric-key or public-key. The focus of this chapter is symmetric-key stream ciphers. A stream cipher generates successive elements of the keystream based on an internal state. This state is updated in essentially two ways: if the state changes independently of the plaintext or ciphertext messages, the cipher is classified as a synchronous stream cipher. By contrast, self-synchronizing stream ciphers update their state based on previous ciphertext digits.

For the synchronous stream ciphers some properties are mandatory.

(i) synchronization requirements. The sender and the receiver must be synchronized – using the same key and operating at the same position (state) within that key. If synchronization is lost due to ciphertext digits being inserted or deleted during transmission, then decryption fails and can only be restored through additional techniques for re-synchronization.

(ii) no error propagation. A ciphertext digit that is modified (but not deleted) during transmission does not affect the decryption of other ciphertext digits.

(iii) active attacks problem. As a consequence of the synchronization requirement, the insertion, deletion, or replay of ciphertext digits during an attack causes loss of synchronization, and offer the possibility to be detected by the attacker. An active attack offers the possibility to make changes to selected ciphertext digits, and find out what affect these changes have on the plaintext. This conclusion proves that the data origin authentication and data integrity must be assured by additional mechanisms.

In a synchronous stream cipher a stream of pseudo-random digits is generated independently of the plaintext and ciphertext messages, and then combined with the plaintext (to encrypt) or the ciphertext (to decrypt). In the most common form, binary digits (bits) are used, and the keystream is combined with the plaintext using the exclusive or operation (XOR). This is called a binary additive stream cipher.

Another approach uses several of the previous N ciphertext digits to compute the keystream. Such schemes are known as self-synchronizing stream ciphers or asynchronous stream ciphers. The idea of self-synchronization has the advantage that the receiver will automatically synchronize with the keystream generator after receiving N ciphertext digits, making it easier to recover if digits are dropped or added to the message stream. Most stream cipher designs are for synchronous stream ciphers. For further details see [5].

2 Design of a Stream Cipher

The design of a new stream cipher involves some important goals:

– To deduce the internal state from the result should be impossible,

– There should be no short cycles,

– It should be cryptographically secure,

– It should be easy to implement,

– The code should be optimized for speed,

– To create as much confusion and diffusion as possible.

I tried to achieve the same goals in designing a new stream cipher named HENKOS (see [1], [2]). This cryptosystem is a symmetric synchronous stream cipher encryption system designed for a software implementation.

After many efforts, here are the results:

– An easy to implement algorithm,

– A cryptographically secure algorithm (proven by statistical tests),

– A very fast algorithm: a 50 megabytes file is encrypted / decrypted in less then one second,

– A fast pseudorandom number generator: a 12,500,000 byte stream needs 0.30 sec for C/C++ code.

This cryptosystem uses a binary additive stream cipher and two types of keys:

– a short-term key named data key (*DK*) with a fixed length of 1024 bytes that is an input in the keystream generator. This key can be generated with a PRNG (not necessarily a cryptographic secure PRNG) or can be an ordinary file, if PRNG is not available.

– a long-term key named master key (*MK*) with a fixed length, which contains 1024 numbers, used to mix the data key and the internal state of the keystream generator. This key must be generated with a true RNG (hardware).

– If during the transmission an attacker intercepts the encrypted data, it is not possible to decrypt the ciphertext correctly without having the master key, because there is a very large number of possible combinations of decrypted ciphertext.

Every attempt to find the master key produces a different plaintext, including the one with the same numbers but the changed order of the numbers in the key affects the decryption process.

2.1 Index keys generation

In this section of the algorithm the master key *MK* is transformed into two index keys *MKS* and *MKT* in two steps. Two functions **Sum** and **Inv** are used: the first one is an additive function and the second one produced a sort of symmetrical figures of number transformation.

The function **Sum** is **Sum** : $\{1,2,\ldots,1024\} \rightarrow \{1,2,\ldots,1024\}$,

$$\mathbf{Sum}(i;MK) = \sum_{j=0}^{i} MK(j) \; modulo \; 1024 \; .$$

The function **Inv** is **Inv** : $\{1,2,\ldots,1024\} \rightarrow \{1,2,\ldots,1024\}$, $\mathbf{Inv}(i) = i^* \; modulo \; 1024$, where i^* is the number obtained by writing the digits of the number i in reverse order. The index keys *MKS* and *MKT* are:

$$\begin{aligned} Step\ 1:\ MKS(i) &= \mathbf{Sum}(i;MK),\ i \in \{1,2,\ldots,1024\}\ , & (1)\\ Step\ 2:\ MKT(i) &= \mathbf{Inv}\big(MKS(i)\big),\ i \in \{1,2,\ldots,1024\}\ . & (2) \end{aligned}$$

The transformation has two targets:

- Not to use the original *MK* key directly in the process,
- To create confusion and diffusion for master key.

Keystream generation transform the *DK* key to obtain the real K key for encryption using two functions: the first one is the essential function in this algorithm the "switch function" **Sw** and the second function **Ad** is an additive one:

$$(\mathbf{Sw}) : DK(j) \leftrightarrow DK(k),\ \text{where } j = MKS(i) \text{ and } k = MKT(i) \text{ for } i \in \{1,2,\ldots,1024\}\ , \quad (3)$$

$$(\mathbf{Ad}) : DK(i) = DK(i) + DK(i+1) \; modulo \; 256,\ i \in \{1,2,\ldots,1023\} \text{ and } DK(1024) = DK(1024) + DK(1)\ . \quad (4)$$

These functions create a totally changed image of the data key *DK*. After these two transformations we obtain DK1; the new key is the input for transformations (3) and (4) and the process will be repeated 64 times:

$$DK \rightarrow DK1 \rightarrow DK2 \rightarrow \ldots \rightarrow DK63 \rightarrow DK64\ . \quad (5)$$

To obtain the keystream bytes $K(i)$ of the final K key, the last operation is:

$$K(i) = \big(DK64(i) + DK64(i+1)\big) \oplus DK64(i) \text{ for } i \in \{1,2,\ldots,1023\};$$
$$K(1024) = \big(DK64(1024) + DK64(1)\big) \oplus DK64(1024)\ . \quad (6)$$

The encryption / decryption process will transform a plain text P of 1024 bytes into a cipher text C of 1024 bytes by using the encryption key K and the function $\oplus$:

$$C(i) = P(i) \oplus K(i)\ .$$

For every stream of 1024 bytes of plain text, another K key wil be used. The new encryption K key will be obtained by the algorithm with the last values of DK as input and the operations described in (4), (5) and (6) will be effectuated one time. This sequence will run until the plain text is finished for one session. Remarks:

- The confusion and the diffusion of the bits are given specially from (3) and (4),
- The data key for the next session we have generated with the same algorithm.

3 Quality analysis

The quality of a stream cipher is measured performing statistical tests. These tests FIPS 140–1, FIPS 140–2, ENT tests, Diehard battery, NIST Statistical Test Suite (a statistical test suite for testing the pseudo–random number generators used in cryptographic applications) were performed on large ciphertext samples of 12.5 megabytes.

3.1 FIPS 140–1/FIPS140–2 Test

FIPS statistical tests contain the Monobit Test, the Poker Test, the Runs Test and the Long Run Test. The following tests are based on performing a pass/fail statistical test on 5000 sequences of 2500 bytes each. In my results the well known generators SHA–1 and CCG together with the new HENKOS pass FIPS 140–1 in proportion of 100%. SHA–1 passes FIPS 140–2 in proportion of 99.6%, CCG in proportion of 99.4% and HENKOS in proportion of 99.64%. For these statistical tests, even if the generators present good statistical properties this isn't a guarantee that the algorithm is good for cryptographic purposes.

3.2 DIEHARD Statistical Tests

The next set of tests was designed to identify weaknesses in many common non–cryptographic PRNG algorithms. These tests analyze a single large file from the output of the generator of 11 megabytes or more (see [3]). The battery of tests include: birthday spacing test, overlapping 5–permutation test, binary rank test 31×31, binary rank test 32×32, binary rank test 6×8, bitstream test, opso, oqso and DNA tests, count–the–1's test on a stream of bytes, parking lot test, minimum distance test, 3Dspheres test, etc.

Majority DIEHARD tests return a $p-$value, which should be uniform on [0,1) if the input file contains truly independent random bits. Those $p-$values are obtained by $p = F(X)$, where F is the assumed distribution of the sample random variable X. When a bit stream really fails, it get $p-$values of 0 or 1 (or close to 0 or 1) to six or more places.

For SHA–1 generator we have 2 $p-$values very close to 1, and for Cubic Congruent Generator we have 28 $p-$values near to 1. For HENKOS we don't have $p-$values close to 0 or 1.

3.3 ENT Tests

ENT applies various tests to a sequence of bytes stored in files and reports the results. The program can be used to evaluate pseudorandom number generators for encryption and compression algorithms. It calculates entropy, optimum compression, chi–square distribution, arithmetic mean, Monte Carlo value for *pi* and serial correlation coefficient. HENKOS obtained good results.

3.4 NIST Statistical Test Suite

The package includes statistical tests for: frequency, block frequency, cumulative sums, runs, long runs, Marsaglia's rank, spectral (based on the Discrete Fourier Transform), nonoverlapping template matching, overlapping template matching's, Maurer's universal statistical, approximate entropy, random excursions (due to Baron and Rukhin), Lempel–Ziv complexity, linear complexity, and serial.

The NIST framework, like many tests, is based on hypothesis testing.

– State your null hypothesis. Assume that the binary sequence is random.

– Compute a sequence test statistically. Testing is carried out at the bit level.

– Compute the $p-$value which must be less than 0.01, otherwise, failure is declared.

The Cubic Congruential Generator fails Frequency, Cumulative Sums, Runs, Aperiodic Template at 11 from 284 templates, Approximate Entropy test, serial and Lempel–Ziv test and Micali generator has 3 fails at Aperiodic Template. HENKOS and BBS pass all this tests.

4 Security Analisys

What is a secure stream cipher? That is a question with no definitive answer, but I can make some assumption on that subject. The entire test package presented here can eventually reveal weaknesses but, even if the ciphers pass with good results, that is not a guarantee of its security. That does not make it fail proof.

Cryptographers consider that there are two main conditions for the security of a stream cipher with a $k-$bit key.

- The attacker should not be able to predict future keystream generated by the cipher in any conditions: recovering the secret key, recovering the internal state of the cipher at some point, or otherwise. The attacker can obviously test all possible secret keys, so the complexity of a brute force attack (requiring at most $2k$ executions of the algorithm) gives a performance baseline to which any alleged attack should be compared to.

- The attacker should not be able to recover the cipher's key or internal state from the keystream. Cryptographers also demand that the keystream be free of even subtitle biases that would let attackers distinguish a stream from random noise, and free of detectable relationships between keystreams that correspond to related keys or related nonce. This should be true for all keys (there should be no weak keys), and true even if the attacker can know or choose some plaintext or ciphertext.

Like other attacks in cryptography, stream cipher attacks can be certificational, meaning they aren't necessarily practical ways to break the cipher but they indicate that the cipher might have weaknesses.

Securely using a secure synchronous stream cipher requires that one never uses the same keystream twice; that generally means that a different nonce or key must be supplied to each invocation of the cipher. Application designers must also recognize that most stream ciphers don't provide authenticity, only privacy: encrypted messages may still have been modified in transit.

Stream Cipher	Creation Date	Speed (cycles/ byte)	bits			Attack	
			Key Length	Init. vector	Internal State	Best Known	Comp. Compl.[1]
$A5/1-2$	1989	Voice	54	114	2^6?	Active	2^{40}
FISH	1993	Quite Fast	Huge	?	?	K-p A[2]	2^{11}
Grain	≤ 2004	Fast	80	2^6	160	Key D[3]	2^{43}
HC-256	≤ 2004	2^2?	2^8	2^8	2^{15}	?	?
HENKOS	2005	7.8	2^8-2^{10}	2^8	2^{10}	?	2^{1024}
ISAAC	1996	$2.38-4.69$	$40-2^8$	N/A[4]	8288	2006 WIS[5]	5×10^{1240}
PANAMA	1998	2	2^8	2^7?	1216?	2001 HC[6]	2^{82}
Rabbit	2003	$3.7-9.7$	2^7	2^6	2^9	2006 N/A	2006 N/A
RC4	1987	Impressive	$40-2^8$	2^3	2064	Key D	$2^{13}-2^{33}$
SEAL	1997	Very Fast	?	2^5?	?	?	?
SOBER-128	2003	?	$\geq 2^7$	?	?	Mes. Forge	2^6
Trivium	≤ 2004	2^2-2^3	80	80	288	Brute Force	2^{135}

Table 1: Comparison of some well known stream ciphers

5 Performances Analysis

5.1 Testing Platform

For testing we select only one platform among many other possible platforms, on the criteria of disponibility and reproductibility of measurements. There are also results, reported in papers, that make possible only a relative comparison between the performances of different algorithms.

5.2 Performances Measurement

The performance algorithms are mesuread by a special program, written in Visual C and based on reading the processor clock before and after the calls of the main phases implementing functions, using the routines get_start_time and get_stop_time written in ASM. The result is calculated as the difference between the two clock readings, minus the additional time consumed with the calls of the clock reading routines.

In the HENKOS case the CPU time for K keys generation is 7.8 cycles/byte, and the encryption process performs 181 megabytes/second. Comparing it to well–known algorithms and the algorithms tested in the Estream Project like CryptMT, Dragon or Salsa20, which are among the top algorithms in the final list, my results are good enough (for details see [6] and [4]).

5.3 Implementation details

The performance was measured reading the processor clock cycles using the RDTSC instruction.

5.4 Comparison of performances

In the real life, in very few cases the authors reveals all the details and the performances of the ciphers. In table 1 there are some results for some well known stream ciphers, [8].

6 Conclusions

There are a lot of stream ciphers used in cryptography because of the speed, but in this case nobody tried a standardisation like in block cipher area. The European Union–based NESSIE project [7], which was aimed at evaluating the security of various cryptographic primitives, did not recommend any stream ciphers in their report.

The performances and quality analysis on cryptographic stream ciphers algorithms are an ambitious goal for all the designers of algorithms. In majority of cases there is no proof of the behaviour of the new cipher, but it's possible to verify the quality by performing statistical tests, and also to measure the performances of implementation and the speed by software means.

In the future, new stream ciphers will appear so that new methods for analysis will be a permanent preoccupation for the cryptographic community.

Bibliography

[1] Bucerzan D. and Gheorghiţă M., HENKOS – *A New Stream Cipher: Performance Analysis*, WARTACRYPT '04 The $_4{}^{th}$ Central European Conference on Cryptology, Bedlewo, Poland, July 2004.

[2] Bucerzan D., *A Cryptographic Algorithm Based on a Pseudorandom Number Generator*, SYNASC'08, Timişoara, October 2008.

[3] Marsaglia G., Diehard Statistical Tests, `http://stat.fsu.edu/pub/diehard/`

[4] Matsumoto M., Saito M., Nishimura T. and Hagita M., CRYPTMT Stream Cipher Version 3, eSTREAM project, `http://www.ecrypt.eu.org/stream/`

[5] Schneier B., *Applied Cryptography*, J. Wiley & Sons Inc, (second edition), 1996.

[6] ***, eSTREAM, `http://www.ecrypt.eu.org/stream/`

[7] ***, NESSIE European Proiect, `http://www.cosic.esat.kuleuven.be/nessie/`

[8] ***, `http://www.answers.com/topic/stream-cipher`

Dominic Bucerzan (b. May 17, 1956) received his M. Sc. in Information Technology from "Aurel Vlaicu" University of Arad, Romania and a PhD in Economic Cybernetics from the "Bucharest Academy of Economic Studies" (2005), with a paper in the field of Information Security. Currently he works as a lecturer in informatics at the Department of Mathematics-Informatics, Faculty of Exact Sciences, "Aurel Vlaicu" University of Arad, România. His current research interests include aspects of IT Security and Cryptography. He is author or co-author of 4 books and more than 45 papers and participated in 35 conferences and workshops.

Mihaela Crăciun (b. March 10, 1972) received her Master of Science in Information Technology from "Aurel Vlaicu" University of Arad, România. At present she is a candidate for a PhD in Computer Science at "Politehnica" University of Timişoara, România. Her current research is focused on Decision in Enterprise Analysis. She published articles in her field of interest.

[1] Computational Complexity
[2] Known-plaintext Attack
[3] Key Derivation
[4] Not Available
[5] Weak Internal State
[6] Hash Collisions

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 4, pp. 490-505

Implementation of the Timetable Problem Using Self-assembly of DNA Tiles

Z. Cheng, Z. Chen, Y. Huang, X. Zhang, J. Xu

Zhen Cheng
College of Computer Science and Technology
Zhejiang University of Technology
288 Liuhe Road, Hangzhou, P.R. China
Email: chengzhen0716@163.com

Zhihua Chen, Yufang Huang, Xuncai Zhang
Department of Control Science and Engineering
Huazhong University of Science and Technology
1037 Luoyu Road, Wuhan, P.R.China

Jin Xu
School of Electronics Engineering and Computer Science
Peking University
No.5 Yiheyuan Road Haidian District, Beijing, P.R.China
E-mail: jxu@pku.edu.cn

Abstract: DNA self-assembly is a promising paradigm for nanotechnology. Recently, many researches demonstrate that computation by self-assembly of DNA tiles may be scalable. In this paper, we show how the tile self-assembly process can be used for implementing the timetable problem. First the timetable problem can be converted into the graph edge coloring problem with some constraints, then we give the tile self-assembly model by constructing three small systems including nondeterministic assigning system, copy system and detection system to perform the graph edge coloring problem, thus the algorithm is proposed which can be successfully solved the timetable problem with the computation time complexity of $\Theta(mn)$, parallely and at very low cost.
Keywords: timetable, self-assembly, graph edge coloring, DNA tiles

1 Introduction

Since Adleman [1] demonstrated the use of recombinant DNA techniques for solving a small combinational search problem, the field of DNA-based computing has experienced a flowering growth and leaves us with a rich legacy. DNA computing [2, 3] potentially provides a degree of parallelism and high density storage far beyond that of conventional silicon-based computers.

DNA tile self-assembly is an important method of molecular computation and it is also a crucial process by which objects autonomously assemble into complexes [4]. This phenomenon is common in nature and yet is poorly understood from mathematical and programming perspectives. It is believed that self-assembly technology will ultimately permit the precise fabrication of complex nanostructures. The DNA nanotechnology was initiated by Seeman [5] who proposed self-assembled nanostructures made of DNA molecules, and the key of this technology is immobilization of Holliday junction (crossover) to make well-defined DNA structures. Seeman [6] also utilized one of such structures called DX (double crossover) tile to realize a patterned lattice made of these tiles, which is used to construct not only simple pattern such as periodic stripes or barcodes, but also the complex algorithmic pattern. Winfree [7] proposed 2D self-assembly process and showed that computation by self-assembly is Turing-universal. Eng [8] demonstrated that self-assembly of linear, hairpin, and branched DNA molecules can generate

regular, bilinear, and context-free languages, respectively. Researchers have used DNA tile algorithmic self-assembly to create crystals with patterns of binary counters [9, 10] and Sierpinski triangles [11], which can be used to implement arbitrary circuit [12]. But those crystals are deterministic, generating nondeterministic crystals may hold the power to solve complex problems quickly.

Because of the complex and special structure of DNA tiles, tile self-assembly is theoretically an efficient method of executing parallel computation where information is encoded in DNA tiles and a large number of tiles can be self-assembled via sticky-end associations. Mao et al. [13] experimentally implemented the first algorithmic DNA tile self-assembly which performed a logical computation (cumulative XOR), however that study only executed two computations on fixed inputs. For the application in arithmetic, Brun [14] proposed and studied theoretically the systems that computed the sums and products of two numbers using the DNA tile self-assembly model, which enough revealed that DNA tile self-assembly had the basic computational ability; For the complex application in combinational problems, tile self-assembly has been proposed as a way to cope with huge combinational NP-complete problems, such as solving the satisfiability problem [15] by using 2D DNA self-assembly tiles, nondeterministically factoring numbers [16], deciding a system of subset sum problem [17]. But generally, the scale is limited to only moderate size problem at best, which further explores the power of computing using DNA tile self-assembly. Furthermore, this model can also be used in the cryptography. XOR computation on pairs of bits can be used for executing a one-time pad cryptosystem that provides theoretically unbreakable security [18].

It is well known that timetable problems [19] are very difficult and time consuming to solve, especially when dealing with large instances. The timetable problem is a combinatorial problem [20] consisting in finding an assignment of a fixed number of teachers to a fixed number of hours in a week, in such a way that a large number of given constraints are satisfied. And it is also known in general to be NP-complete [21]. For the most important, the timetable problems are subject to many strict constraints that are usually divided into two categories: “hard" and “soft" [22]. Hard constraints are rigidly enforced and have to be satisfied for the timetable problem. Soft constraints are those that are desirable but not absolutely essential. So it is difficult to generate a satisfactory solution within a short time. In order to avoid the disadvantage of their exponential computation complexity, here we mainly focus on the timetable problem based on DNA tile self-assembly, which is a kind of better technique and the model can successfully perform the problem with the operation time complexity of $\Theta(mn)$, parallely and at very low cost.

The rest of this paper is structured as follows: Section 2 describes the mechanism of self-assembly based on the DNA tiles in detail. Section 3 shows the process of performing the timetable problem by self-assembling. The conclusion will summarize the contribution of our work.

2 Algorithmic DNA tile self-assembly

Algorithmic DNA self-assembly is both a form of nanotechnology and a model of DNA computing. As a nanotechnology, the aim of algorithmic DNA self-assembly is to design tiles with carefully choosing glue types on their sides. Two tiles are said to be of different types if their sides have different glue types. Useful tile types are nontrivial to design but relatively easy to duplicate in large quantity. A key design challenge for algorithmic DNA tile self-assembly is to use only a small number of different tile types to assemble a target nanostructure to complete the corresponding computation.

2.1 Models for algorithmic DNA tile self-assembly

The tile assembly model extends the theory of Wang tilings [23] of the plane by adding a natural mechanism for growth. As a computational model, algorithmic DNA self-assembly encodes the input of a computational problem into DNA patterns and then manipulates these patterns to produce new DNA

patterns that encode the desired output of the computational problem. Informally, the model consists of a set of four sided Wang tiles whose sides are each associated with a type of glue. The bonding strength between any two glues is determined by a glue function. A special tile in the tile set is denoted as the seed tile. Assembly takes place by starting with the seed tile and attaching copies of tiles from the tile set one by one to the growing seed configuration whenever the total strength of attraction from the glue function meets or exceeds a fixed parameter called the temperature. Generally, the tile set and the seed configuration should be constructed before the biological operations together with the suitable temperature.

In addition, the tile assembly model [24] is a formal model of crystal growth. It was designed to model self-assembly of molecules such as DNA. Rothemund and Winfree [25] defined the abstract tile assembly model, which provides a rigorous framework for analyzing algorithmic self-assembly. Here, we mainly use the abstract tile assembly model to solve the timetable problem. Intuitively, the model has tiles or squares that stick or don't stick together based on various binding domain on their four sides. Figure 1 gives the structures of DNA tiles, mainly including the TAO and TAE tiles. Figure 1(a) describes the structure of TAO tile. Figure 1(b) shows the three TAO tiles joining diagonally. The TAE tiles and the corresponding abstract tiles can be seen in (c), (d) and (e). Figure 1(f) gives the structures which are assembled to form a compact lattice.

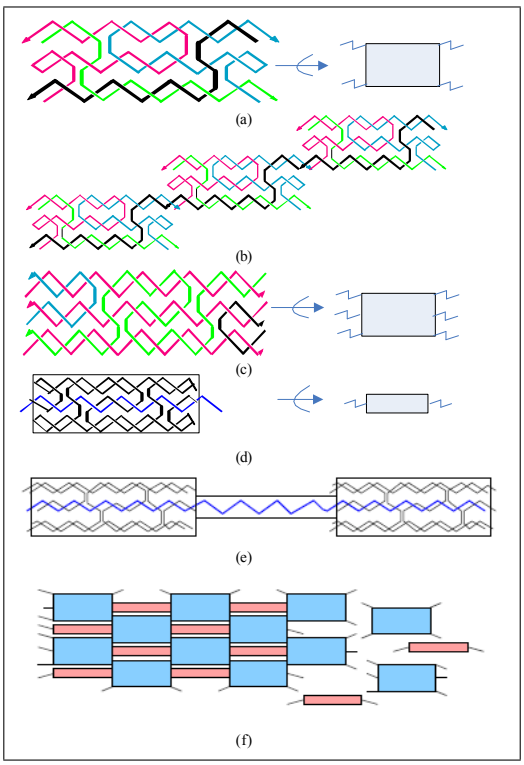


Figure 1: DNA Tiles. (a)Structure of TAO tile. (b)Three TAO tiles join diagonally. (c)Structure of TAE tile. (d)TX tile. (e)Two TAE tiles join. (f)The two types of tiles can assemble to form a compact lattice structure.

2.2 Computation by DNA tile self-assembly

Computation by self-assembly is the spontaneous self-ordering of substructures into superstructures driven by annealing of Watson-Crick base-pairing DNA sequences. Computation by DNA tile self-assembly entails the building up of superstructures from starting units such that the assembly process

itself performs the actual computation.

DNA tile self-assembly is also a highly parallel process, where many copies of different molecules bind simultaneously to form intermediate complexes. One might be seeking to construct many copies of the same complexes at the same time, as in the assembly of periodic 1D or 2D arrays; Alternatively, one might wish to assemble in parallel different molecules, as in DNA-based computation, where different assemblies are sought to test out the combinatorics of the problem.

A sequential or deterministic process of DNA tile self-assembly has three highly parallel instruction steps [4]. The first one is molecular recognition: elementary molecules selectively bind to others. The second is growth: elementary molecules or intermediate assemblies are the building blocks that bind to each other following a sequential or hierarchical assembly. The cooperativity and non-linear behavior often characterize this process. The third way is termination: a built-in halting feature is required to specify the completion of the assembly. In practice, their growth is interrupted by physical and/or environmental constraints. DNA tile self-assembly is a time-dependent process and because of this, temporal information and kinetic control may play a role in the process before thermodynamic stability is reached.

3 Implementing the timetable problem based on DNA tile self-assembly

In this section, we first give the definition of the timetable problem, then we mainly show the algorithm for solving the timetable problem based on DNA tile self-assembly, and concretely introduce the process how they can perform this problem. Finally, examples of success and failure in the tile attachments are given to demonstrate the reasonability and validity of the algorithm.

3.1 The timetable problem

The typical timetable problem consists in assigning a set of activities/actions/events (e.g. work shifts, duties, classes) to a set of resources (e.g. physicians, teachers, rooms) and time periods, fulfilling a set of constraints of various types. Constraints stem from both nature of timetable problems and specificity of the institution involved. In other words, timetable or planning is a process of putting in a sequence or partial order a set of events to satisfy temporal and resource constraints required to achieve a certain goal, and is sometimes confused with scheduling, which is the process of assigning events to resources over time to fulfill certain performance constraints. However, many scientists consider scheduling as a special case of timetable and vice versa [26].

In this paper, we solve a special kind of timetable problem which is the coursetable problem [27]. The problem consists in scheduling courses for a set of courses in a university, taught by available teachers in a given period composing a number of weeks, and in available classrooms. Although the constraints of timetable problem vary from case to case, one can classify all constraints into hard constraints and soft constraints. Hard constraints must be strictly satisfied because any timetable that violates just one will become useless. A timetable that violates some soft constraints can still be usable although it may cause some inconvenience to the users. It is often very difficult to satisfy all the soft constraints in a real life. Some concrete definitions of hard constraints and soft constraints in a coursetable problem will be given as follows [28]. Some examples of hard constraints are:

HC0 - A teacher can only teach in a single place at a time.

HC1 - A teacher can only give one course at a time.

HC2 - A room can only host one course at a time.

HC3 - A student can only attend one course at a time.

HC4 - Room capacities must be respected.

HC5 - No more than a teacher is scheduled to teach in a room each time.

HC6 - Each subject is scheduled in a proper room (for example, a laboratory needs a proper equipment).

HC7 - Every teacher must have scheduled all his hours.

HC8 - Every student must have scheduled all his hours.

We denote the fact that some conditions can be deduced from other constraints. For example, HC0, HC2→ HC1. Some examples of soft constraints are:

SC0 - The courses should be scheduled in the morning and the seminaries and laboratories in the afternoon.

SC1- Some courses are scheduled with a prior consideration.

SC2 - As much as possible the preferences of the teachers and the ones of the students should be respected.

3.2 Solving the timetable problem based on DNA tile self-assembly

Here, we mainly introduce the algorithm for implementing the timetable problem based on DNA tile self-assembly. First, the timetable problem can be converted into the graph edge coloring problem, then the tile self-assembly model is used to solve the graph edge coloring problem with some constraints including mainly constructing three small systems which are nondeterministic assigning system, copy system and detection system, thus the timetable problem can be successfully carried out. Examples can be given to indicate how the tile self-assembly model performs in this problem.

The graph edge coloring problem

Let G be an undirected graph where V is the set of vertices and E is the set of edges. Mathematically, an assignment of colors to the edges of a graph G(one color to each edge so that adjacent edges are assigned different colors) is called a coloring of G. Edges with a same color define a color class. A k-coloring of G is proper if incident edges have different colors; that is, if each color class is a matching, otherwise conflicts happen. A coloring with at least one conflict is called an infeasible coloring. A graph is k-edge-colorable if it has a proper k-edge-coloring. For the given coloring of a graph G, a set consisting of all those edges assigned the same color is referred to as a color class.

In this study, the timetable problem can be converted into the graph edge coloring problem. First, a complete bipartite graph, denoted as $K_{m,n}$, is a graph consisting of two sets of vertices, one with m vertices and the other with n vertices. There is exactly one edge from each vertex in the one set to each vertex in the other set. There are no edges between vertices within a set. Then we give the bipartite graph from the arrangement matrix of the timetable problem. Second, the hard and soft constraints can be considered as the constraints of the graph edge coloring problem. According to the graph theory, the feasible solutions of the edge colorings are the arrangements of courses in the timetable problem. Here, we mainly propose non-deterministic algorithm to solve the graph edge coloring problem by using the massive parallelism possible in DNA tile self- assembly, thus the timetable problem with some given constrains can be successfully solved.

In the process of implementing the tile self-assembly systems, many assemblies happen in parallel by creating billions of billions of copies of the participating DNA tiles, so this is simulated by an exponential number of DNA assemblies which can be converted into the space occupied by the DNA molecules, thus we expect that the procedure will run in parallel on all possible colorings. In this case, there are many possible valid tilings, any or all of which may be produced. When tiles are implemented by real molecules, one would expect a set of tiles to nondeterministically generate a combinatorial library of input assemblies, and then a deterministic set of rule tiles could evaluate each input assembly to determine whether it represents the desired answer.

The nondeterministic assigning system

Non-determinism implies that at some steps the algorithm makes a non-deterministic choice. Of course, there are many differences between the deterministic computation and nondeterministic computation. In terms of deterministic computation, it can be defined as a tile system to produce a unique final seed configuration if for all sequences of tile attachments, all possible final configurations are identical. Comparing to the deterministic computation, the nondeterministic computation is a system in which different sequences of tile attachments can attach different tiles in the same position. Intuitively, a system nondeterministically computes a function if at least one of the possible sequences of tile attachments produces a final configuration, which contains the computation results. Furthermore, in many implementations of the tile assembly model that would simulate all the nondeterministic executions at once, it is useful to be able to identify which executions succeed and fail in a way that allows selecting only the successful ones.

The nondeterministic assigning system can give a color set to the edges of the graph. First, the edges of the given graph can be labeled as "$e_1, e_2, \cdots, e_m$", the vertices can be noted as $X_i (1 \leq i \leq n)$. Here, m, n is the number of edges and vertices of the graph respectively. Each edge in the graph can be nondeterministeically obtained one color. The same edge connecting different vertices should share the same color. If there is only one edge which is adjacent to the vertex, the information about the vertex and the edge needn't be arranged on the rightmost column in the seed configuration, but should be assigned with one color at the bottom of the seed configuration of the nondeterministic assigning system.

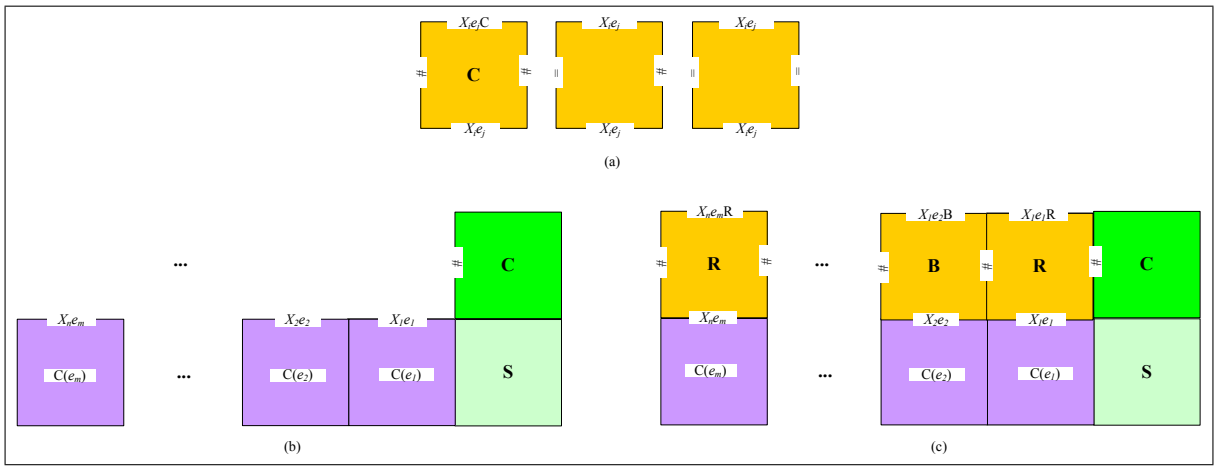


Figure 2: The framework of the nondeterministic assigning system. (a) The basic tile types of this system. (b) The seed configuration of the system. (c) An example of the nondeterministic assigning system.

The color set of the edges can be nonderministically generated by the tile self-assembly configuration. Here, suppose the edge e_j is on the vertex X_i which is labeled as X_ie_j. $C(e_j)$ is denoted as the edge e_j with the color "C". The same edge on the remainder vertices can pass the information from the bottom to the upper in the tile and can obtain the same color. The basic tile types of this system can be shown in Figure 2(a). Figure 2(b) shows the seed configuration of the system. Figure 2(c) is an example of the nondeterministic assigning system which can assign a color set to the edges.

The copy system

The copy system mainly carries out three functions by designing three basic tile types which can be shown as follows in Figure 3. Here, X_ie_j denotes the edge e_j is adjacent to the vertex $X_i (1 \leq i \leq n)$. "C" is the color of the edge e_j on the vertex X_i.

The first function is to pass the color of the edge given by the nondeterministic assigning system to the same edge on different vertices, so it can make sure that the same edge adjoining different vertices has the same color. The tile type is labeled with blue. If the edge is adjacent to different vertices, "$X_i e_j C*$" and $X_i e_j$ should be passed to the left and the upper of the tile respectively. Otherwise, the color "C" can be passed to the edge $X_i e_j$. At the same time, if the input at the bottom of the tile includes the information of the colors, the outputs are only needed to copy the information of the inputs from left to right, and synchronously from bottom to upper in the tile.

The second is to copy the possible colorings generated by the nondeterministic assigning system from the bottom of the seed configuration to the uppermost of the self-assembly complexes. After the edges sharing the common vertex have been checked the colorings by the detection system, there would be a condition that the edges adjoining different vertices $(k \neq i)$ and with different colors $(C_1 \neq C_2)$ will meet together, but they have no need to be checked the coloring and also should be passed to the left tiles which is shown in the second tile type with the color turquoise.

The third is to copy the information with the edges which are adjacent to the vertices on the rightmost column to the left tiles, so the detection system can check up whether colorings of the edges sharing the common vertex are feasible, and synchronously, they also should be passed to the upper in the tile. Here, $X_i e_j$ is the edge which has at least two adjacent edges sharing a common vertex. Once a vertex has l adjacent edges, $(l-1)$ edges with the labels smaller then should be arranged on the rightmost column in the seed configuration of the problem, but all the edges should be at the bottom of the seed configuration. The tile type can be shown as follows with the color rosiness.

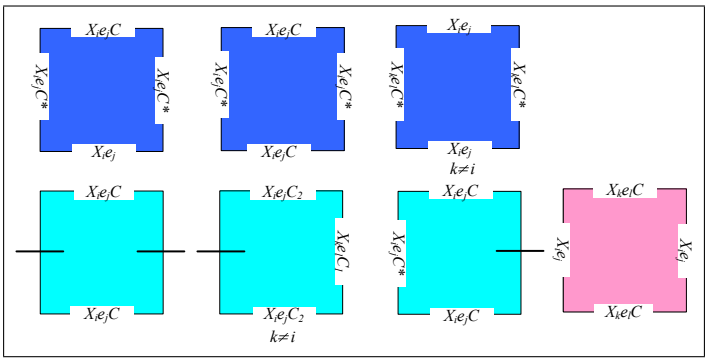


Figure 3: The basic tile types of the copy system

The detection system

The key to the detection system is to make one system implement the checking operations. If the adjacent edges sharing the common vertex have different colors, the feasible coloring of the edges for the vertex has been completed with the symbol "Ok". Once the edges sharing one common vertex have at least two same colors, the self-assembly complexes will stop to grow with the information "No tile can match" and the coloring is not feasible. Here, the coloring of the adjacent edges for each vertex should be satisfied with the same constraints. For the timetable problem, the hard and soft constraints can be described by the constraints of the edges coloring for the corresponding graph.

When the edges which are adjacent to the vertices on the rightmost column and the coloring of each edge from the copy system are passed to the detection system, it can check up whether the coloring is feasible or not. If the comparison result at this step is "Ok", it should be passed to the left tiles to continuously make the next color checking until the edges don't share the same vertex, then it doesn't check the coloring with the left tiles any more, and synchronously the colors of the edges at the bottom of the seed configuration in the nondeterministic assigning system should be passed to the higher layers.

Here, e_k and $e_j(k \neq j)$ are the adjacent edges on the common vertex X_i, and they have different colors, so the comparison result is "Ok". For the second tile type, the edge e_j which is adjacent to the vertex labeled as X_i meets the color "C", then it can pass "$X_i e_j C$" to the right in the tile, and the value of the tile is the color "C" of the edge e_j.

If the result is "No tile can match", the self-assembly complexes can't grow any more and the input colors of the edges are not the feasible solutions of the graph edge coloring problem. The formula of the detection system can be described as follows:

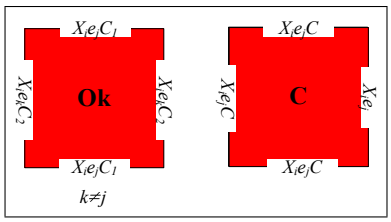


Figure 4: The basic tile types of the detection system

Here, this system will use the L-configuration to encode inputs, and produce its output on the top row of an almost complete rectangle. Therefore, systems could chain results together. The input structure encodes the edges colorings of the graph on the bottom row and encodes the edges and their adjacent vertices on the rightmost column. The output tiles needs three different kinds of tiles as follows in Figure 5. The pink tile shows the edge e_j with the color "C" adjoining the vertex X_i which is the feasible coloring for the graph. Only if all the edges adjoining different vertices have feasible colorings, the result is "Success", and the feasible solution can't be obtained otherwise.

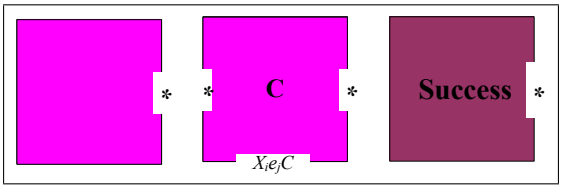


Figure 5: The output tiles of the timetable problem

Here, we design the algorithm to implement the timetable problem as following steps:

Step 1: Convert the timetable problem into the graph edge coloring.

Step 2: Generate all possible input combinational colors of the edges to the given graph for the timetable problem with some constraints.

Step 3: According to the rules for dealing with the constraints using the massive parallelism of DNA self-assembly to check up whether all the possible inputs are the feasible colorings for the edges in the graph. In this process, the copy system can pass the color of each edge in the graph from the bottom of the seed configuration to the higher layers, and synchronously copy the information of edges which are adjacent to the vertices, then the detection system can judge whether the colorings of the edges sharing the common vertices are feasible or not.

Step 4: Reject all infeasible solutions according to constraints of the edge coloring and reserve all feasible solutions, therefore we can obtain feasible solutions of the graph edge coloring problem which are also the solutions of the timetable problem.

Step 5: Reading the operation result is done by the reporter strand method. Under certain biological operations, we can obtain all the result strands which run through all the results of the feasible colors of the edges for the given graph. Each report strand records the result of one feasible input. The strands can be amplified by polymerase chain reaction using the primers to ligate each end of the long reporter strand. Then through gel electrophoresis and DNA sequencing, we can read out the result strands of

different lengths representing the information with the feasible coloring results. Finally, we can easily get the feasible solutions of the timetable problem.

The actual implementation detail is not discussed here since they fall outside of the scope of this paper. However, we believe that we make no arbitrary hypotheses. In fact, our work is based on the achievements that come with DNA tiling computation in general.

Examples of the timetable problem

Here, we take an example of timetable problem to verify the validity of our method. Suppose there are three teachers X_1, X_2, X_3, and four classes Y_1, Y_2, Y_3, Y_4. The arrangement matrix of the courses is shown as follows:

$$P = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \end{array} \begin{array}{c} \begin{array}{cccc} Y_1 & Y_2 & Y_3 & Y_4 \end{array} \\ \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

The timetable problem has the hard constraints are:

HC0 - A teacher can only teach in a single place at a time.

HC1 - A teacher can only give one course at a time.

HC2 - A room can only host one course at a time.

HC3 - A student can only attend one course at a time.

HC4 - No more than a teacher is scheduled to teach in a room each time.

HC5 - The teacher X_3 should give a course to the class Y_2, which is arranged in the second period in the morning time.

HC6 - There are enough rooms for the courses where the students attend. Some soft constraints are:

SC0 - Some courses are scheduled with a prior consideration.

SC1 - As much as possible the preferences of the teachers and the ones of the students should be respected.

SC2 - If possible, the order of the courses classes taken Y_j are more earlier than Y_{j+1}.

First, we should convert the timetable problem into the graph edge coloring problem with some constraints. The bipartite graph from the arrangement matrix of the timetable problem can be shown in Figure 6. All the edges of the graph are "e_1, e_2, e_3, e_4, e_5, e_6, e_7" and the vertices are "X_1, X_2, X_3, Y_1, Y_2, Y_3, Y_4".

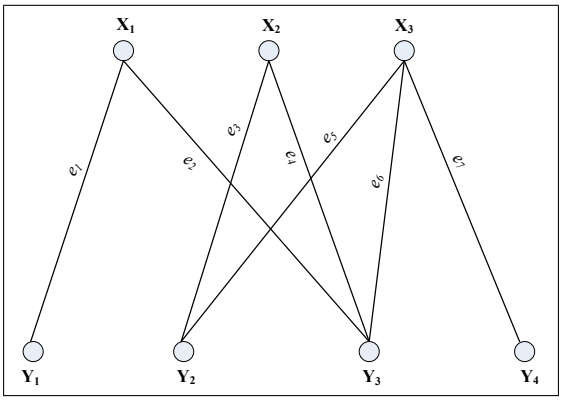


Figure 6: The bipartite graph from the arrangement matrix of the timetable problem

Second, according to the method introduced above, we need construct the basic tile types in each of the three small systems and they are the same as the tiles described above and the seed configuration, which can be shown in Figure 7. When all the tiles and the seed configuration are prepared, we put them

together into the reaction buffer. According to the DNA tiles prepared and the mechanism of algorithmic DNA tile self-assembly through Watson-Crick base pairing, the self-assemble process starts at the same time with the connector tiles, so the final stage can be seen in Figure 8.

We also can see that the process of the three small systems performing. The nondeterministic assigning system can give a color set "RBRYBRY" to all the edges of the graph "$e_1, e_2, e_3, e_4, e_5, e_6, e_7$" which are adjacent to the vertices "X_1, X_2, X_3, Y_1, Y_2, Y_3, Y_4" respectively. All the edges on different vertices should be arranged at the bottom of the seed configuration no matter whether the vertices adjoin only one edge or not. At the same time, the edges "$X_1e_1, X_3e_5, Y_2e_3, Y_3e_2, Y_3e_4$" are on the rightmost column of the seed configuration. The copy system can pass the colors of the edges from the bottom of the seed configuration to the upper layers, and pass the edges and their adjacent vertices on the rightmost column to the left tiles, so that the detection system can check whether the colorings of the edges sharing one common vertex are feasible. The vertex X_1 which has two adjacent edges e_1 and e_2 with different colors "R" and "B" respectively is checked up the feasibility of the colorings by the detection system and the result is "Ok". For the vertex X_3, it has three adjacent edges e_5, e_6 and e_7 which are at the bottom of the seed configuration. One of the two edges e_5, e_6 with the smaller subscripts than e_7 are on the rightmost of the column. The detection system only need verify the colors of e_5 and e_6, e_5 and e_7, e_6 and e_7, and the three comparison results are all "Ok". The method of checking the colorings of other vertices is also the same.

Finally, to output the computation result, we would implement a modification of the standard sequence-reading operation that uses a combination of PCR and gel electrophoresis. On adding these tiles, and allowing them to anneal, then we get the final tile assembly. On adding ligase to seal the bonds, we will have a single strand of DNA passing through the tiles in the final output layer, which encodes the colorings of the edges. This single strand begins with the unique nucleotide sequence labeled "Success". Therefore, the feasible assignment of the edge colorings can be obtained if and only if the symbol "Success" appears in the result DNA strand. Through using the operations, we can extract the strands of different lengths representing the output tiles in the result strands. In this example, we can obtain the feasible solution of the "RBRYBRY". The color sets "R", "B" and "Y" have the corresponding relationship with the edge sets "e_1, e_3, e_6" , "e_2, e_5" and "e_4, e_7" which are also "X_1Y_1, X_2Y_2, X_3Y_3", "X_1Y_3, X_3Y_2" and "X_2Y_3, X_3Y_4". Thus the feasible solution of the timetable problem which is also the arrangement of the courses can be described as: "X_1Y_1, X_2Y_2, X_3Y_3" are arranged in the first period, "X_1Y_3, X_3Y_2" and "X_2Y_3, X_3Y_4" in the second and third period respectively which are satisfied with the constraints in the problem.

For the nondeterministic algorithm, we give the same example to show the failure in attaching tiles in Figure 9 and don't get the right results. If the nondeterministic assigning system gives a color set "RBRYBBY" to the edges "$e_1, e_2, e_3, e_4, e_5, e_6, e_7$" respectively, there will be some conflicts in the process of the growth for the assembly complexes. When the detection system checks up the coloring of the vertex X_3, the colorings of the edges e_5 and e_6 are the same which are both "B", so the conflict generates and the result is "No tile can match", thus the self-assembly complexes can't grow any more, therefore, the coloring of the edges assigned is the infeasible solution of the problem. It means that "X_1Y_1, X_2Y_2", "X_1Y_3, X_3Y_2,X_3Y_3" and "X_2Y_3, X_3Y_4" are not the feasible arrangements of the courses, here there is a conflict that the teacher X_3 can't give a course in two different classes Y_2 and Y_3 at the same period.

Complexity analysis

The complexity of the design is considered in terms of computation time, computation space and the number of distinct tiles required. Generally, suppose there are m teachers, and n classes for the timetable problem.

It is obvious from the given examples that the upper bound of the computation time T is $T = m(n-1)+n(m-1)+mn+4+mn+mn+2 = \Theta(mn)$.

The upper bound of the computation space S taken for each assembly is the area of the assemble

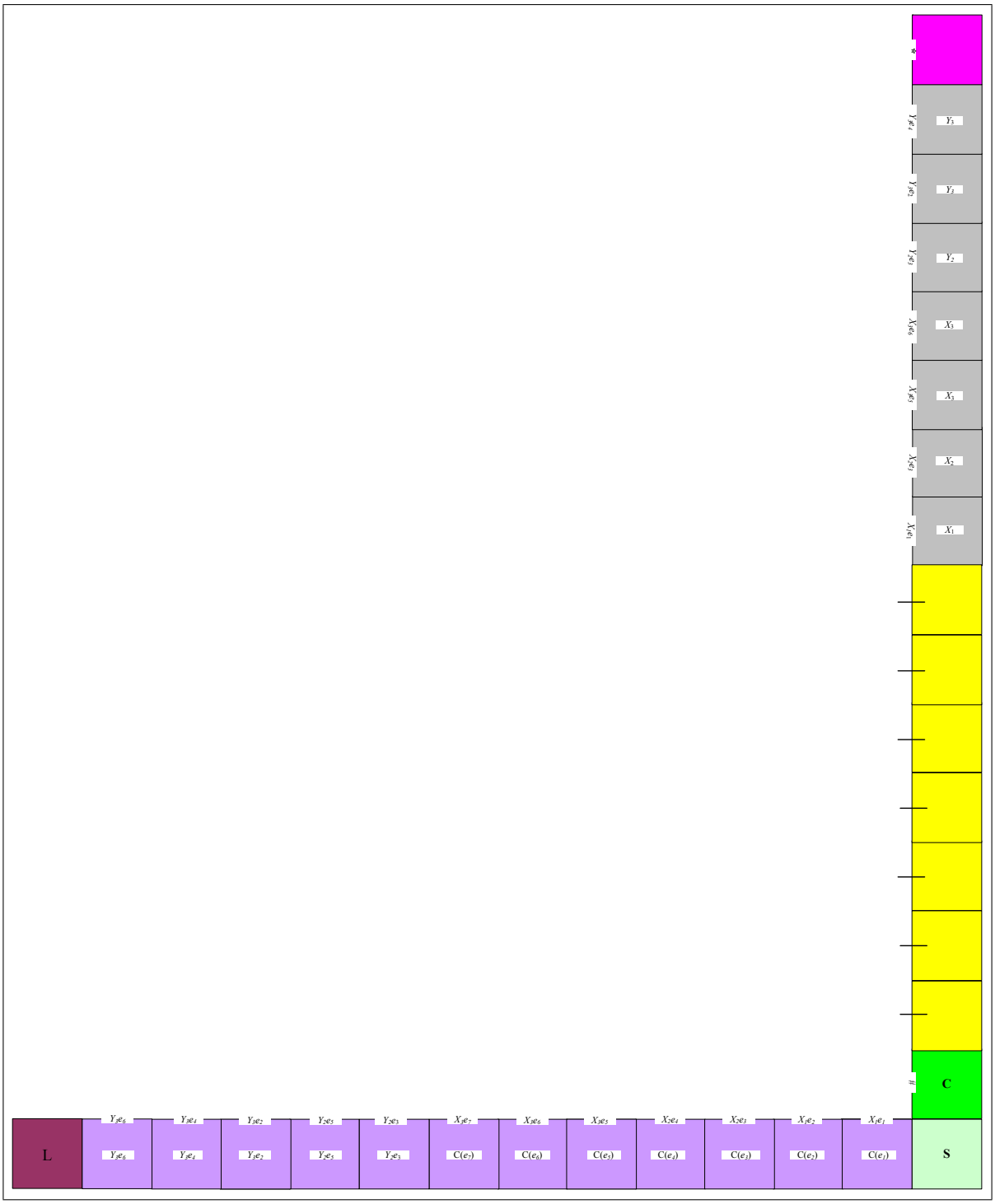


Figure 7: The seed configuration of the timetable problem in the example

Figure 8: The final stage of the successful example for the problem

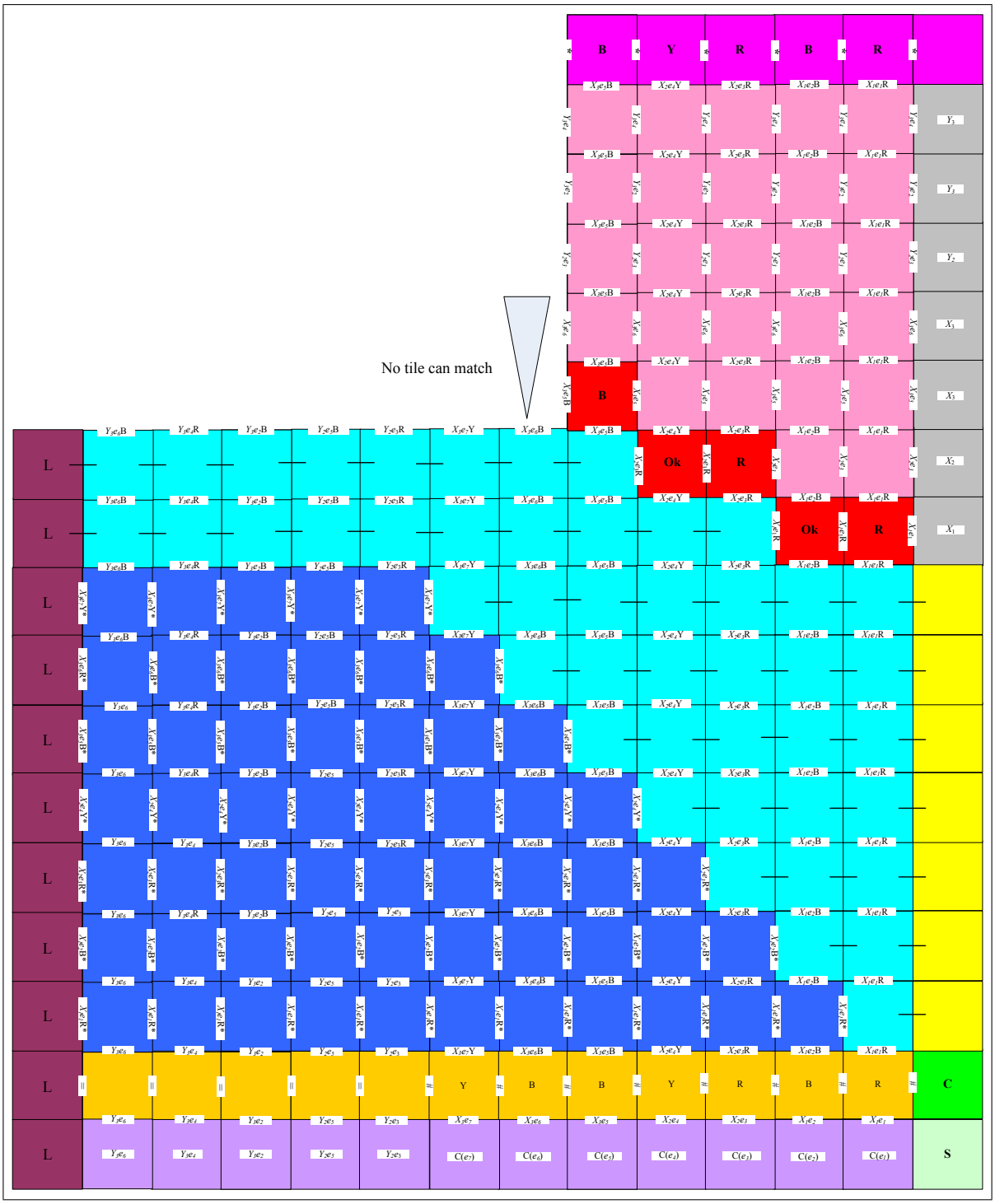


Figure 9: The failure of the example for the problem because of infeasible coloring sets

complexes represented by $S = [m(n-1)+n(m-1)+mn+4]*[mn+mn+4] = \Theta(m^2n^2)$ which is upper-bounded polynomially to the number of variables.

Finally, suppose the graph G which is converted from the given timetable is k-edge-colorable. The upper bound of tiles needed contain the following tiles:

Boundary tiles. These tile types include the input boundary tiles and computation boundary tiles. According to the number of the edges in the graph, there are $2mn$ tiles for the input tiles at the bottom of the seed configuration, and $[m(n-1)+n(m-1)+3]$ on the rightmost column. The computation boundary tiles contain $(kmn+km+3)$ types. So the upper bound of the boundary tiles is $[2mn+m(n-1)+n(m-1)+kmn+km+6]$.

Computation tiles. For the assigning operations, there must be $(kmn+mn)$ tiles with the upper bound, and they are shown in Figure 2. For the copy system which can be seen in Figure 3, the upper bound of the tile types is $[km(n-1)+kn(m-1)+kmn+kmn+kmn]$. For the detection system in Figure 4, there must be $[km(n-1)+kn(m-1)+kmn]$ tiles with the upper bound. Thus the upper bound of the computation tiles is $[9kmn+mn-k(m+n)]$.

Output tiles. Finally, there must be some tiles to output the results. The upper bound is $(2kmn+2)$ and the tiles are shown in Figure 5.

Summing up all the tile types, because the value of k can be determined for the given timetable problem, so we can have the upper bound of the total number of tiles: $[2mn+m(n-1)+n(m-1)+kmn+km+6]+[9kmn+mn-k(m+n)]+(2kmn+2)=\Theta(mn)$.

4 Summary and Conclusions

DNA tile self-assembly is looked forward to many applications in different fields. In this paper, we show how the DNA self-assembly process can be used for solving the timetable problem. The advantage of our method is that once the initial strands are constructed, each operation can compute fast parallelly through the process of DNA self-assembly without any participation of manpower, thus the algorithm is proposed which can be successfully implemented the timetable problem with the operation time complexity of $\Theta(mn)$, parallely and at very low cost. A limitation of the algorithm, which is common for most DNA computations, comes from the fact that the exponential dimension of the problem has been pushed into the physical space (volume) occupied by the DNA molecules. This will eventually become a restrictive factor. The input size and thus the DNA volume can't grow forever. This implies an upper bound to the size of instances that can be solved in practice.

While DNA tile self-assembly suffers from high error rates, the possible sources of errors are, either an error in constructing the tiles, or an erroneous binding of tiles, methods of error control and error correction may be used to decrease the error rates in the computation of DNA tile self-assembly model. Many experimental results in DNA tile self-assembly have not appealed to the advantages of crystal growth; however, these early works on the fundamentals of self-assembly and the physical experimental evidence of actual DNA tile crystals suggest a bright future for DNA tile self-assembly. The field of nanotechnology holds tremendous promise, but many technical hurdles will have to be overcome before algorithmic DNA tile self-assembly can be developed into a practical commercial technology. If the molecules and supramolecules can be controlled at will, then it may be possible to achieve vastly better performance for computers and memories. So we can see that the DNA tile self-assembly model has various applications in many fields and it also might open up a host of other applications in materials science, medicine, biology and other ways.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (Grant Nos. 60674106, 30870826, 60703047, 60533010 and 60803113), 863 Program of China (2006AA01Z104), and program for New Century Excellent Talents in University (NCET 05-0612), Ph.D. Programs Foundation of Ministry of Education of China (20070001020), Chenguang Program of Wuhan (200750731262), and the open fund of Key Lab. for Image Processing and Intelligent Control (No.200703).

Bibliography

[1] L.M. Adleman, Molecular computation of solutions to combinatorial problems, *Science*, vol. 266, pp. 1021-1024, 1994.

[2] L.Q. Pan, J. Xu, Y.C. Liu, A Surface-Based DNA Algorithm for the Minimal Vertex Problem, *Progress in Natural Science*, vol. 13, pp. 81-84, 2003.

[3] L.Q. Pan, G.W. Liu, J. Xu, Solid phase based DNA solution of the coloring problem, *Progress in Natural Science*, vol. 14, pp. 104-107, 2004.

[4] A. Carbone, N.C. Seeman, Molecular Tiling and DNA Self-assembly, *Springer-Verlag Berlin Heidelberg*, LNCS 2950, pp. 61-83, 2004.

[5] N.C. Seeman, DNA nanotechnology: novel DNA constructions, *Annu. Rev. Biophy. Biomol. Struct.*, vol. 27, pp. 225-248, 1998.

[6] C. Mao, W. Sun, N.C. Seeman, Designed two dimensional DNA Holliday junction arrays visualized by atomic force microscopy, *J. Am. Chem. Soc.*, vol. 121, pp. 5437-5443, 1999.

[7] E. Winfree, On the computational power of DNA annealing and ligation, *DNA Based Computers*, pp. 199-221, 1996.

[8] T. Eng. Linear self-assembly with hairpins generates the equivalent of linear context-free grammars. *3rd DIMACS Meeting on DNA Based Computers*, *Univ. of Penn.*,1997.

[9] R. Barish, P. Rothemund, E. Winfree, Two computational primitives for algorithmic self-assembly: Copying and counting, *Nano Letters*, vol. 12, pp. 2586-2592, 2005.

[10] Pablo Moisset de Espane′s, Ashish Goel, Toward minimum size self-assembled counters, *Springer Science Business Media B.V.*, 2008.

[11] P. Rothemund, N. Papadakis, E. Winfree, Algorithmic self-assembly of DNA Sierpinski triangles, *PLoS Biology*, vol. 12, pp. 2041-2053, 2004.

[12] M. Cook, P. Rothemund, E. Winfree, Self assembled circuit patterns, *DNA*, pp. 91-107, 2004.

[13] C. Mao, T.H. LaBean, J.H. Reif, Logical computation using algorithmic self-assembly of DNA triple-crossover molecules, *Nature*, vol. 407, pp. 493-496, 2000.

[14] Y. Brun, Arithmetic computation in the tile assembly model: Addition and multiplication, *Theoretical Computer Science*, vol. 378, pp. 17-31, 2006.

[15] G.L. Michail, T.H. LaBean, 2D DNA self-assembly for satisfiability, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. vol. 44, pp. 139-152, 1999.

[16] Y. Brun, Nondeterministic polynomial time factoring in the tile assembly model, *Theoretical Computer Science*, vol. 395, pp. 3-23, 2008.

[17] Y. Brun, Solving NP-complete problems in the tile assembly model, *Theoretical Computer Science*, vol. 395, pp. 31-36, 2008.

[18] A. Gehani, T.H. LaBean, J.H. Reif, In DNA Based Computers: Proceedings of a DIMACS Workshop, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 1999.

[19] D. Werra, An introduction to timetabling. *European Journal of Operations Research*, vol. 19, pp.151-162, 1985.

[20] D. Abramson, Constructing school timetables using simulated annealing: Sequential and parallel algorithms. *Management Science*, vol. 37, pp. 98-113, 1991.

[21] A. Colorni, M. Dorigo, V. Maniezzo, Genetic algorithms and highly constrained problems: the time-table case. *In: H.P.Schwefel and R.Manner(eds). Parallel Problem Solving from Nature. Proceedings of 1st Workshop, PPSN 1, LNCS 496, Dortmund, Germany, 1-3 October. Springer-Verlag,* pp. 55-59, 1991.

[22] L. Gaspero, A. Schaerf, Tabu search techniques for examination timetabling. *In Proceedings of the 3rd International Conference on Practice and Theory of Automated Timetabling (PATAT 2000)*, *Springer-Verlag*, LNCS 2079, pp. 104-117, 2001.

[23] H. Wang, Proving theorems by pattern recognition, *I. Bell System Technical Journal*, vol. 40, pp. 1-42, 1961.

[24] E. Winfree, Algorithmic self-assembly of DNA, Ph.D.Thesis, Caltech, Pasadena, CA, June 1998.

[25] P. Rothemund, E. Winfree, The program-size complexity of self-assembled squares, *ACM Symposium on Theory of Computing (STOC)*, pp. 459-468, 2001.

[26] Mihaela Oprea. MAS_UPUCT: A Multi-Agent System for University Course Timetable Scheduling. *International Journal of Computers*, *Communications* & *Control*, Vol. II, pp. 94-102, 2007.

[27] Zhipeng L., Jin-Kao Hao. Adaptive Tabu search for course timetabling. *European Journal of Operational Research*, doi:10.1016/j.ejor.2008.12.007, 2009.

[28] A.R. Mushi, Mathematical programming for mulations for the examinations timetable problem: the case of the university of DAR ES SALAAM. A*frican Journal of Science and Technology (AJST) Science and Engineering Series*, Vol. 5, pp. 34-40, 2004.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 4, pp. 506-516

Cereal Grain Classification by Optimal Features and Intelligent Classifiers

A. Douik, M. Abdellaoui

Ali Douik, Mehrez Abdellaoui
Ecole Nationale d'Ingénieurs de Monastir (ENIM)
Département de Génie Electrique
Laboratoire ATSI
Rue Ibn El Jazzar, 5019 Monastir Tunisie
E-mail: {Ali.douik,mehrez.abdellaoui}@enim.rnu.tn

Abstract: The present paper focused on the classification of cereal grains using different classifiers combined to morphological, colour and wavelet features. The grain types used in this study were Hard Wheat, Tender Wheat and Barley. Different types of features (morphological, colour and wavelet) were extracted from colour images using different approaches. They were applied to different classification methods.
Keywords: morphological, colour, wavelet transform, neural networks, statistical classifier, fuzzy logic.

1 Introduction

The past few years was marked by the development of researches that contribute to reach an automatic classification of cereal grains which is perceived as a possible solution to prevent human errors in the quality evaluation process. Computer vision system which is a promising technology in the quality control can replace the human operator. After hours of working the operator may loose concentration which in turn will affect the evaluation process. So a computer vision system proved to be more efficient at the level of precision and rapidity. But, the natural diversity in appearance of various cereal grains varieties makes classification by computer vision a complex work to achieve. Many researches were carried out to classify cereal grains. Characterization models were based on morphological features ([1–9]), colour features ([10–13]) or textural features ([14]). Other researchers ([15–18]) have tried to combine these features for the sake of improving the efficiency of classification. Recently, wavelet technique was integrated in cereal grains characterization ([19, 20]). This technique, developed by Mallat [21], is used in textural image analysis to make object classification more precise. The present paper is divided into four main parts. The first one will deal with the cereal image acquisition system, the second part will be devoted to present the classification features with its morphological, colour and wavelet components, the third section will focus on the different methods used in the classification process and the last one will compare the different methods accompanied with their performance evaluation.

2 Cereal image acquisition system

2.1 Image acquisition device

A high resolution colour camera (VIVITAR) with a USB 2.0 cable was used to acquire grain images. The acquired images were of 3.1 mega pixel resolution. Light sources were placed symmetrically over and under a glass plate over which the grains are spread out. All the samples were taken at constant camera settings, i.e., exposure time, saturation and gamma. The images obtained were pre-processed to eliminate background pixels using image subtraction. Indeed, the active image containing grain sample is compared to image containing background. The image we got contains the grains and a uniform background (black). This step of pre-processing makes the gains segmentation easier and more efficient.

Table 1: Freeman code features and their abbreviations

Region	Direct1	Direct2	Direct3	Direct4	Direct5	Direct6	Direct7	Direct8
Region1	Vz_{11}	Vz_{12}	Vz_{13}	Vz_{14}	Vz_{15}	Vz_{16}	Vz_{17}	Vz_{18}
Region2	Vz_{21}	Vz_{22}	Vz_{23}	Vz_{24}	Vz_{25}	Vz_{26}	Vz_{27}	Vz_{28}
Region3	Vz_{31}	Vz_{32}	Vz_{33}	Vz_{34}	Vz_{35}	Vz_{36}	Vz_{37}	Vz_{38}
Region4	Vz_{41}	Vz_{42}	Vz_{43}	Vz_{44}	Vz_{45}	Vz_{46}	Vz_{47}	Vz_{48}

2.2 Image database samples

A database of images was created from various samples of several cereal varieties obtained from different sources and for different crop years from laboratories of the Tunisian Cereal Office. Tunisian Hard Wheat (HW), Tunisian Tender Wheat (TW) and Tunisian Barley (B) are the main classes of the samples considered.

3 Classification features

For each grain type, 152 parameters are extracted from the colour images of the database (122 morphological, 18 colour, and 12 wavelet features).

3.1 Morphological features

After isolating the grain, the region of interest was selected around the boundary of the edge. The morphological features were obtained from the binary images containing only pixels of the grain edge. We can classify these features as follows:

- **Grain size measurements:** Length (L), width (l), width by lengh ratio (R_1), area (S), perimeter (P), area by perimeter ratio (R_2), angles (GrA, PtA) and radius of curvature (Rr, Rl) of the two extremities, likelihood between the grain and the nearest ellipsis for the grain (E), mean (Sx, Sy) and standard deviation (σ_x, σ_y) of horizontal and vertical symmetry.

- **Freeman code features:** After dividing the grain image in four regions as shown in the figure (1.a). We perform for every region the freeman code ([22]); it's the oldest contour descriptor and the most used today; it's mainly based on the position of the pixels set that are the nearest neighbours (NN-set) of the actual pixel. In fact, every region is coded starting from a given origin and according to the directions of the nearest neighbour that are represented in 8-connexity (coded on 3 bits) as demonstrated in figure (1.b). The features extracted from the Freeman code are 32; eight for every region. These features are summed up in table 1.

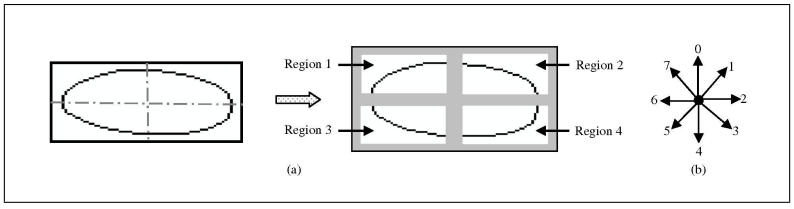


Figure 1: Freeman code extraction, (a) Dividing image in four regions to compute the Freeman code, (b) Direction codes

- **Fourier transform features:** The Fourier Transform is an important image processing tool which is used to decompose an image into its sine and cosine components ([23]). The application of

Table 2: List of colour parameters and their abbreviations

Colour Components	Mean value	Mean square value	Variance	Standard deviation	Kurtosis	Skewness
Red	RM_1	RM_2	RV	RSD	RM_3	RM_4
Green	GM_1	GM_2	GV	GSD	GM_3	GM_4
Blue	BM_1	BM_2	BV	BSD	BM_3	BM_4

this transform on the contour pixels creates a set of complex coefficients that represents the shape of the contour. From these coefficients we extract the morphological descriptors using different signatures (1).

$$a(u) = \frac{1}{N}\sum_{k=0}^{N-1} s(k) exp\left[\frac{-j2puk}{N}\right] \tag{1}$$

Where:

$u \in [0, N-1]$ $(N : number of points in contour)$

$s(k)$: the signature chosen.

$a(u)$: harmonic descriptors.

The signatures used are complex, radial distance and polar. From each signature, we selected the first 25 harmonic coefficients that can be added to the set of the morphological features.

The three signatures used are invariant by translation and consequently their Fourier descriptors (FD), but it was proved that they are sensitive to rotation. Invariance by rotation is then realized by ignoring the FD phase and by considering only modules of these Fourier descriptors.

For the complex signature all descriptors except the first (DC component) are needed to index the form. The DC component describes only the contour position, and it is useless with the form description. The descriptors standardization consists in dividing their modules by the one of second descriptor. The vector which indexes the form is given by the (2).

$$F = \left[\frac{|FD_2|}{|FD_1|}, \frac{|FD_3|}{|FD_1|}, \ldots, \frac{|FD_{N-1}|}{|FD_1|}\right] \tag{2}$$

The radial distance function and the polar coordinates are real. They have N/2 different frequencies for that half of the FD is necessary to index the form. The invariant vector (3) is obtained by dividing the module of the N/2 first descriptors by the module of the first descriptor.

$$F = \left[\frac{|FD_1|}{|FD_0|}, \frac{|FD_2|}{|FD_0|}, \ldots, \frac{|FD_{N/2}|}{|FD_0|}\right] \tag{3}$$

3.2 Colour features

For each colour image that contains an isolated grain, we perform statistical parameters on values of pixels belonging to the grain. Color parameters included: Mean value, Mean square value, Variance, Standard Deviation, Kurtosis and Skewness of the Red, Green and Blue primaries. In table 2 we present these parameters and their abbreviations.

Table 3: List of wavelet parameters and their abbreviations

Matrix type	Average value	Variance	Standard deviation
Matrix of approximation image	$MVAP$	VAP	$SDAP$
Matrix of horizontal details	$MVHD$	VHD	$SDHD$
Matrix of vertical details	$MVVD$	VVD	$SDVD$
Matrix of diagonal details	$MVDD$	VDD	$SDDD$

3.3 Wavelet features

The wavelet analysis of an image is a multi resolution analysis which is defined by linear operators allowing analyzing a signal on various frequencies. Indeed, the signal is projected on a scale function that gives a representation of the original signal at higher scale. This projection causes a back zoom of the original signal, where the approximation is performed. In order to rebuild the signal, starting from approximation coefficients, we must also project the original signal on a wavelet to recover information lost during the first projection. The second projection contains the details of the original signal.

The details of wavelet features have been reported earlier in [20]. Table 3 resumes the chosen features. They were statistically tested to extract the best parameters leading to an optimal classification.

The tests done on the 12 parameters proved that only two parameters are judged like not-significant (ADH and ADD). Thus; the number of parameters which is going to be retained for the characterization phase is 10: SDAP, AAP, VAP, SDDH, VDH, SDDV, ADV, VDV, SDDD and VDD.

4 Classification methods

Starting from the classification features extracted, we developed many methods using different approaches. The first approach is a statistical classification method that uses only morphological and colour features. The second approach is a classification using a fuzzy logic based method. The third is a combination between the first and the second. The last approach is an artificial neural network classification method that exploits all features leading to the best classification result.

In what follows, we present these different approaches and their contribution to the classification of cereal grains.

4.1 Statistical classification method

From the set of samples, we achieved statistics related to morphological and color features extracted from color images of grains. From these statistics we obtained a distribution curve of every feature. This method operates directly on the distributions intervals of the morphological and color parameters. The classification is made by successive tests on parameters according to their ranks. Conceived algorithm has been tested on images containing a mixture of grains collected from treated samples.

To classify the grain types using a statistical method we considered the morphological and color features in this approach. Classification results for the grain types using this method are illustrated in figure 2.

We notice that the recognition rate for TW is weak while working with morphological or colour features (morph. 56%; color 51%). For HW, colour features gave an optimal recognition rate exceeding 99,4%, but does not exceed 67% when working with morphological features. For Barley grains, due to their form that is different from other types of grains, the morphological features gave us a good classification result reaching 98,7%. The global recognition rate for the statistical classification method

Table 4: Test of best parameters

Parameters	Barley	Hard wheat	Tender wheat	Total
Lsb	82,43%	77,02%	90,28%	80,94%
E	77,82%	20,75%	90,91%	46,28%
GrA	72,80%	70,57%	67,13%	68,62%
RM2	64,02%	50,52%	39,18%	50,25%

is limited to 76%. This is explained by the overlapping that exists between the distribution curves of grain classes when working with all the morphological and colour features.

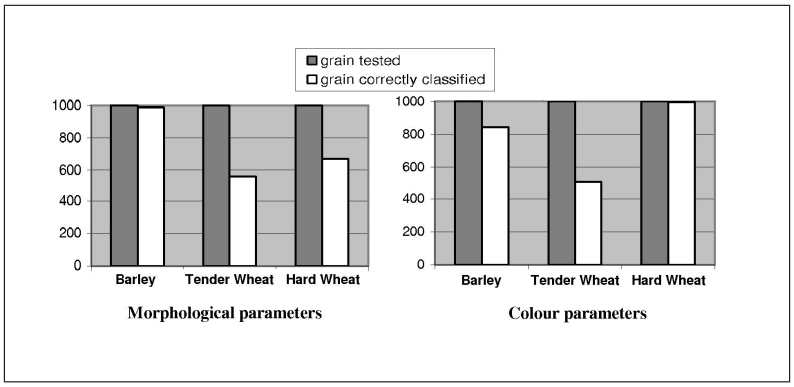


Figure 2: Results of the statistical classification method when applied to morphological and colour features

4.2 Fuzzy logic based classification method

Due to the overlaps of distribution curves of grains types we implement a classification method based on fuzzy logic techniques to improve the recognition rate issued from the statistical classification method.

Classification using the fuzzy logic is made according to the following steps:

- Classes's definition.
- Generation of the membership functions for every parameter.
- Development of inference rules.
- Decision making.

It results three classes corresponding to the different grain types considered. Membership functions are deduced from the distribution curves of the different parameters of every grain type. The membership functions were conceived by normalization of the curves and then by a Gaussian approach for every curve. The number of rules depends on the number of parameters considered. The chosen norm is the max-prod. Then, the rules form is: "IF (condition1) AND (condition2) THEN (decision)".

The choice of entries is based on a test of identification parameters. Table 4 illustrates the test of the four best parameters for the classification from the set of morphological and colour features associated to the fuzzy logic method.

From the possible combinations of the four parameters we select the best ones according to its recognition rates. The combinations selected are illustrated (Lsb and GrA : 83,42%; Lsb, GrA and RM2 : 72,71%; Lsb, GrA, E and RM2 : 68,23%). From this test we chose the parameters Lsb and GrA since

when combining them it gives the best recognition rate. The result of this method using the combination Lsb and GrA is shown in figure 3.

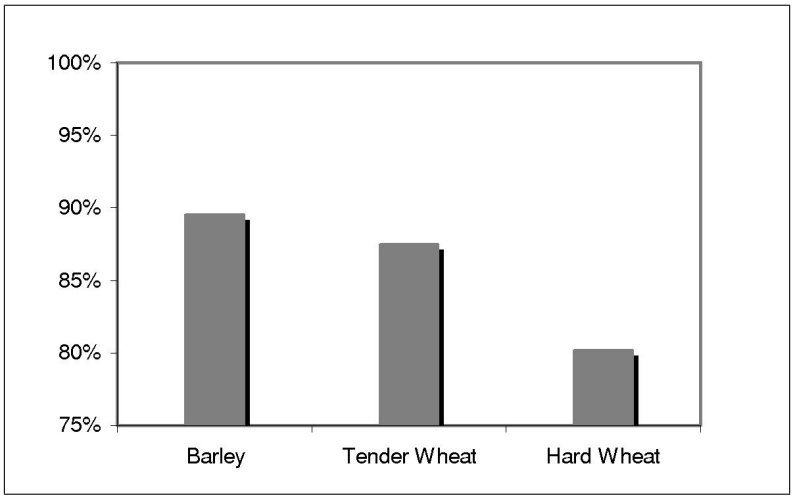


Figure 3: Results for Fuzzy Logic based classification method

Concerning the hard wheat and tender wheat grains, this method gives us a best recognition rate than the statistical one. On the other hand for the barley grains, the first method is more reliable. So that, we opt to use a method which combines the two previous methods and gives t best recognition rate.

4.3 Statistical and Fuzzy Logic combined classification method

It consists in making a decision about the grain type by the fuzzy logic method in the cases where the statistical method cannot make a decision. The fuzzy logic is used in the combined method in the cases of overlaps of all morphological and colour parameters. The improvement concerns the hard wheat and tender wheat grains only; the barley grains possess an optimal recognition rate. The results of this method are illustrated in figure 4.

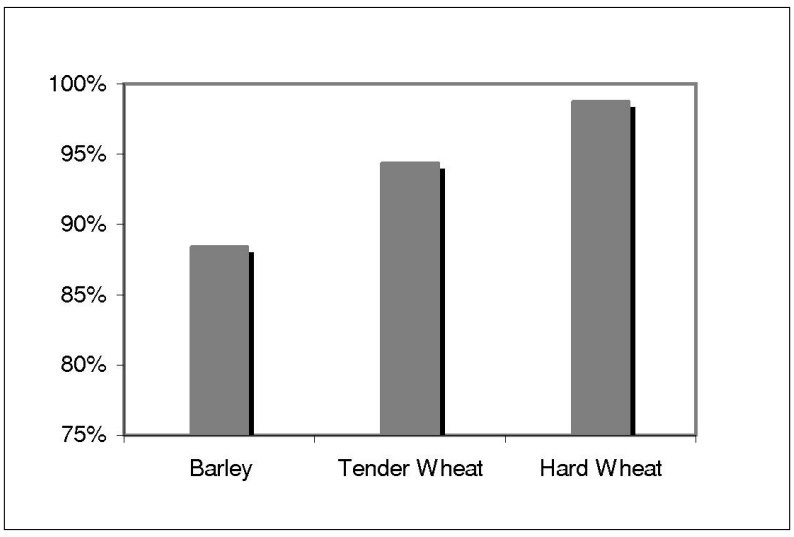


Figure 4: Results for Statistical and Fuzzy Logic combined classification method

Table 5: Number of neurons in the hidden layer

	Morphological features				Colour features	Wavelet features
	SM	FC	FT	AMF		
Number of features	15	32	25	122	18	10
Number of neurons	3	7	5	5	10	3

Table 6: Classification results for ANN classifier

	Rates (%)								
	Morphological features			Colour features			Wavelet features		
Grain type	CR	RJR	RCR	CR	RJR	RCR	CR	RJR	RCR
B	1,1	0	98,9	2,8	0	97,2	2,5	0,7	96,8
HW	1,6	0,5	97,9	7,6	0,7	91,7	1,7	1	97,3
TW	7,7	2,8	89,5	3,5	1,3	95,2	0	0	100
Mean rates	3,5	1,1	95,4	4,6	0,7	94,7	1,4	0,6	98

4.4 Artificial Neural network classification method (ANN)

Training phase and network architecture

The network architecture is a multi-layer neural network MLP. The training is done using the function "TRAINLM" from the Matlab neural network toolbox. Activation functions are hyperbolic tangent and the linear Matlab functions "tansig" and "purelin". During the training phase, we varied the neurons number in the hidden layer and we determine the training error. We chose 40000 as training iterations number since this value leads to a minimum training error. The number of neurons in the hidden layer depends of the type of features considered as entries of the network in the table 5 we illustrate the variation of the number of neurons in the hidden layer when using different types of features (for the morphological features SM means Size Measurments, FC : Freeman Code, FT : Fourier Transform and AMF : All Morphological Features) .

Classification results

For this test we used 3000 grains (1000 grains of each class), 600 grains for characterization and 400 grains for validation. The training of each class is done using 1800 grains, 600 grains will be used to learn the true membership and the others 1200 will be used to learn the system the false membership to the class. This technique seems to be very original and will make it possible to enlarge the classification space, to refine space collates and to reduce the conflict rate between various classes.
Thus this test will determine the conflict rate (CR), the rejection rate (RJR) and recognition rate (RCR). Table 6 represents the results obtained during the first test.

5 Evaluation and discussion

Figure 5 shows the classification recognition rates of the four developed methods. The ANN classifier lead to the best recognition rates for Barley (98,9%) using morphological features and Tender wheat (100%) using wavelet features whereas the statistical and fuzzy logic combined classifier was the best for Hard wheat classification (98,7%). These two methods gave better results than the first and second one.

The tables 7, 8, 9 and 10 present the confusion matrixes of the four developed classifiers. When we

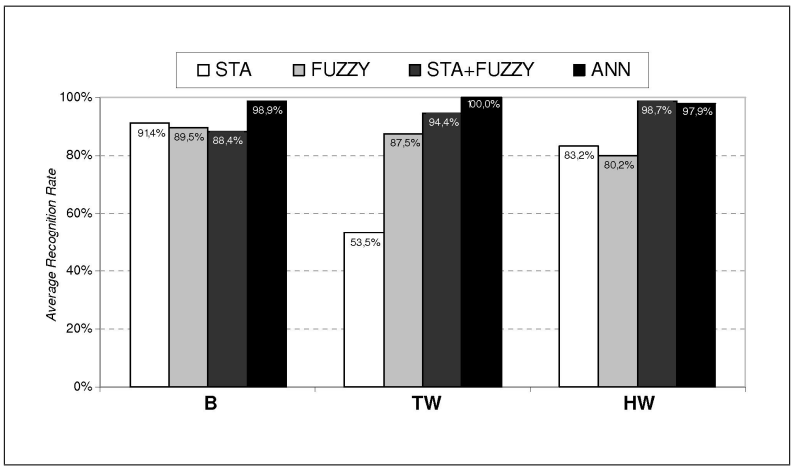


Figure 5: Comparison of the classification methods based on average recognition rates

Table 7: Confusion Matrix (%) for the statistical method

	B	TW	HW
B	91,4	1,5	7,1
TW	5,2	53,5	41,3
HW	2,9	13,9	83,2

observe these matrixes, we note that the major confusions are between Tender Wheat and Hard Wheat in the Statistical classification method (41,3% for HW and 13,9 TW), Fuzzy Logic classification based method (12% for HW and 15,7% for TW) and Statistical and Fuzzy Logic Combined classification method (5% for HW and 0,8% for TW) this is due to the similarities that exists in the morphology and the texture of these two cereal grain classes. This problem is resolved using the ANN classification method (0% for HW and 1,6% for TW).

Barley grains are more confused with Hard Wheat (STA: 7,1% ; FUZZY: 9,6% ; STA+FUZZY: 10,5% and ANN : 0,6%) than with Tender Wheat (STA: 1,5% ; FUZZY: 0,9% ; STA+FUZZY: 1,1% and ANN : 0,5%) this is due to the size that is larger than Tender Wheat and colour features.

To evaluate the time performance for each classification method we count the time in seconds that takes every algorithm to classify grains in a sample of 300 grains containing 100 grain of each type. The algorithms are developed on a Toshiba Satellite (Intel Core 2 / 1,6 Ghz) Laptop under Windows Vista environment. Table 11 presents the cost time for each method.

Table 8: Confusion Matrix (%) for the fuzzy logic based method

	B	TW	HW
B	89,5	0,9	9,6
TW	0,5	87,5	12,0
HW	4,1	15,7	80,2

Table 9: Confusion Matrix (%) for the statistical and fuzzy logic cobined method

	B	TW	HW
B	88,4	1,1	10,5
TW	0,6	94,4	5,0
HW	0,5	0,8	98,7

Table 10: Confusion Matrix (%) for the ANN method

	B	TW	HW
B	98,9	0,5	0,6
TW	0	100	0
HW	0,5	1,6	97,9

Table 11: Time performance of the different methods

Method	Time(s)
STA	72
FUZZY	43
STA+FUZZY	84
ANN	177

We have noticed that the Fuzzy Logic based classification method appears to run about 60% faster than the second fastest (Statistical classification method). The method leading to best recognition results is 4 times slower than the fastest methods. The Statistical and Fuzzy Logic combined classification method can be considered as the most performing method as it have a good recognition rate (94%) and take 50% less time than the method leading to the optimal recognition rate. While the reported execution times depend on the implementation language, we note that we have used Matlab 2007.

6 Conclusion

As dealt above, the classification of different grain types was successfully achieved using different parameters based on different types of features (morphological, colour and wavelet). These parameters were tested on different classification methods; the statistical classification method gave an average recognition rate of 76%. The second method based on fuzzy logic techniques gave an average recognition rate of 85,73%. The hybrid method, which is a combination of the two fore mentioned methods gave an average recognition rate of 93,83%. Finally, the ANN classification method was tested on all features and gave the best recognition rate reaching 98%.

Bibliography

[1] M. Abdellaoui, A. Douik, M. Annabi, Détérmination des critéres de forme et de couleur pour la classification des grains de céréales, *Proc. Nouvelles Tendances Technologiques en Génie Electrique et Informatique, GEI'2006*, Hammamet,Tunisia, 2006, pp. 393-402.

[2] D. A. Barker, T. A. Vouri, M. R. Hegedus, D. G. Myers, The use of ray parameters for the discrimination of Australian wheat varieties. Plant Varieties and Seeds 5(1) (1992) 35-45.

[3] D. A. Barker, T. A. Vouri, M. R. Hegedus, D. G. Myers, The use of slice and aspect ratio parameters for the discrimination of Australian wheat varieties, Plant Varieties and Seeds 5(1) (1992) 47-52.

[4] D. A. Barker, T. A. Vouri, M. R. Hegedus, D. G. Myers, The use of Fourier descriptors for the discrimination of Australian wheat varieties, Plant Varieties and Seeds 5(1) (1992) 93-102.

[5] D. A. Barker, T. A. Vouri, M. R. Hegedus, D. G. Myers, The use of Chebychev coefficients for the discrimination of Australian wheat varieties, Plant Varieties and Seeds 5(1) (1992) 103-111.

[6] P. D. Keefe. A dedicated wheat grain image analyzer, Plant Varieties and Seeds 5(1) (1992) 27-33.

[7] H. D. Sapirstein, J. M. Kohler, Physical uniformity of graded railcar and vessel shipments of Canada Western Red Spring wheat determined by digital image analysis, Canadian Journal of Plant Science 75(2) (1995) 363-369.

[8] J. Paliwal, N. S. Shashidhar, D. S. Jayas, Grain kernel identification using kernel signature, Transactions of the ASAE 42(6) (1999) 1921-1924.

[9] S. Majumdar, D. S. Jayas, Classification of cereal grains using machine vision. I. Morphology models, Transactions of the ASAE 43(6) (2000) 1669-1675.

[10] M. Neuman, H. D. Sapirstein, E. Shwedyk, W. Bushuk, Wheat grain colour analysis by digital image processing: I. Methodology, Journal of Cereal Science 10(3) (1989) 175-182.

[11] M. Neuman, H. D. Sapirstein, E. Shwedyk, W. Bushuk, Wheat grain colour analysis by digital image processing: II. Wheat class determination, Journal of Cereal Science 10(3) (1989) 182-183.

[12] X. Y. Luo, D. S. Jayas, S. J. Symons, Identification of damaged kernels in wheat using a colour machine vision system. Journal of Cereal Science 30(1) (1999) 49-59.

[13] S. Majumdar, D. S. Jayas, Classification of cereal grains using machine vision. II. Color models, Transactions of the ASAE 43(6) (2000) 1677-1680.

[14] S. Majumdar, D. S. Jayas, Classification of cereal grains using machine vision. III. Texture models, Transactions of the ASAE 43(6) (2000) 1681-1687.

[15] S. Majumdar, D. S. Jayas, Classification of cereal grains using machine vision. IV. Combined morphology, color, and texture models, Transactions of the ASAE 43(6) (2000) 1689-1694.

[16] J. Paliwal, N. S. Visen, D. S. Jayas, N. D. G. White, Comparison of a neural network and a non-parametric classifier for grain kernel identification, Biosystems Engineering, 85(4) (2003) 405-413.

[17] N. S. Visen, D. S. Jayas, J. Paliwal, N. D. G. White, Comparison of two neural network architectures for classification of singulated cereal grains, Canadian Biosystems Engineering 46 (2004) 3.7-3.14.

[18] M. Abdellaoui, A. Douik, M. Annabi, Hybrid method for cereal grain identification using morphological and color features, Proc. 13th IEEE International Conference on Electronics, Circuits, and Systems, (Nice, France, 2006), pp. 870-873.

[19] R. Choudhary, J. Paliwal, D. S. Jayas, Classification of cereal grains using wavelet, morphological, colour, and textural features of non-touching kernel images, Biosystems engineering 99 (2008) 330 - 337.

[20] A. Douik, M. Abdellaoui, Cereal varieties classification using wavelet techniques combined to multi-layer neural networks, Proc. 16th Mediterranean Conference on Control and Automation, (Ajaccio, France, 2008) pp1822-1827.

[21] S. G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(7) (1989) 674-693.

[22] H. Freeman, On the encoding of arbitrary geometric configurations, IEEE Trans on Electr. Comput. 10 (1961) 260-268.

[23] D. Zhang, G. Lu, A Comparative Study on Shape Retrieval Using Fourier Descriptors with Different Shape Signatures, Proc. IEEE International Conference on Multimedia and Expo, (2001), pp. 1139-1142.

Ali Douik was born in Tunis, Tunisia. He received the Master degree from the "Ecole Normale Supérieure de l'Enseignement Technique de Tunis", in 1990 and the Ph.D. degree in Automatic from the "Ecole Supérieure des Sciences et Techniques de Tunis, Tunisia", in 1996. In 2010, he received the ability degree from the "University of Monastir, Tunisia". He is presently "Maitre assistant" in the "Ecole Nationale d'Ingénieurs de Monastir". His research is related to Automatic Control and Image Processing.

Mehrez Abdellaoui was born in Tunis in 1979. He received his Electrical Engineering Diploma from Electrical Engineering Department in ENIM-Monastir in 2003 and the Master degree in Automatics from the ENIM-Monastir in 2005. He is currently a PhD student in the Electrical Engineering Department at the ENIM-Monastir. His research interests include Image Processing and Video Analysis.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 4, pp. 517-524

E-Learning & Environmental Policy: The case of a politico-administrative GIS

N.D. Hasanagas, A.D. Styliadis, E.I. Papadopoulou, L.A. Sechidis

Nikolaos D. Hasanagas, Athanasios D. Styliadis, Lazaros A. Sechidis
Kavala Institute of Technology, Department of Landscape Architecture
City of Drama, Greece, GR 66100
Email: styliadis@ath.forthnet.gr, nikolaos.hasanagas@gmail.com,
lazikas@photo.topo.auth.gr

Eleni I. Papadopoulou
Aristotle University of Thessaloniki, Faculty of Agricultural Science
Department of Agricultural Economics, City of Thessaloniki, GR 54124
E-mail epapa@agro.auth.gr

Abstract: Is an effective knowledge exchange and cooperation between academic community and practitioners possible? Implementation of e-learning in specialized policy fields pertains to the most challenging priorities of ICTs and software engineering. In multidisciplinary academic areas which combine environmental policy studies with positivist subjects (like environmental issues, forest policy, rural development, Landscape Architecture etc), the using of e-learning system in analyzing policy issues steadily gains in importance and is a method which connects the academic community and the researchers with the practitioners and field experts. Such initiatives incorporate a number of politometrics- relevant algorithms embedded in a context of political geography (i.e. visualized hierarchies in different region-related policy issues). This is the case addressed in this paper. The GIS learning management system introduced in this paper is based on certain criteria concerning organizational models and region-specific politico-administrative hierarchies. Scenarios of politico-administrative metadata achieving optimal power synergy are extracted through a sequencing technique, combining vector-algebra software and statistics and can be used for both teaching and research purposes.
Keywords: e-Learning, GIS, politometrics, forest policy, environmental issues, rural development policy, socio-informatics.

1 Introduction

Although the political structures of the "western" and "civilized" world are considered to be standardized in the framework of a single "cosmopolite" value system, the power structures in environmental policy issues are quite different between regions. Not in every region of Europe the state actors necessarily concentrate the same degree of power. Sometimes private enterprises, environmental or economic groups are the leading actors. An actor (e.g. environmental or landowner interest group) should also be adjusted to the particular condition of regional policy networks in order to succeed. Socio-informatics software like VISONE (network analysis) is based on elaborated vector-algebra algorithms [3] which are aiming at quantifying and visualizing the intangible political relations [1] and informal dynamics of environmental policy and in general of rural development policy. The application of such a computer-aided "political geometry" methodology with region-specific cases is the "cornerstone" of structuring a GIS functional for politometrics and thus of implementing a GIS Learning Managament System (GLMS) [4] in the context of a post-modern political geography depicting regional-specific (in)formal hierarchies (Archimedes findings).

The implementation of e-learning in specialized policy fields pertains to the most challenging priorities of ICTs and software engineering. In multidisciplinary academic areas which combine environmental policy studies with positivist subjects (like Landscape Architecture, Rural Sociology and Economics, Forest Science etc), the using of e-learning system in analyzing policy issues steadily gains in importance (Archimedes findings). Moreover, in cross-sectoral policy networks such as those which are developed on forest policy issues and are discussed in this paper, it is impossible to separate policy sectors; Environmental issues, rural development policy and forest policy

are an inseparable part of what is called "integrated rural development". Such networks involve not only forest owner and industrial interests but also other groups (e.g. agricultural museums, environmental NGOs, agrarian associations etc). Not only the classical rural territory but also urban-related interests are involved. Thereby, a new urban-countryside relation is developed. Thus, a policy-relevant filtering of learning objects is necessary in order to enable the exchange of knowledge between academic community (or researchers) and practitioners/ field experts.

The e-learning system suggested in this paper is expected to be appropriate for achieving not only an effective over-bridging between academic community and practitioners in forest policy and in the wider environmental and rural development issues but also an effective organisation and coordination of means of forest and rural resource management, an acceptable evaluation of forest resources and acceptable procedures of estimating or accounting the economic, material and non-material values of forest, frictionless goal-setting and decision-making involving private and public actors, and a method of examining issues of ambiguity and law-making concerning forest and wider natural resources. As long as this e-learning is implemented among target groups from different policy sectors related to forestry (e.g. spatial planning, agriculture, tourism, water management etc), a minimization of conflicts is feasible [6]. This holistic approach of policy-making is enabled through the complete analysis of policy networks, which is an operational form of system theory [8, 10]. In other words, a new systemic analysis of classical forest policy is the basis of this e-learning system.

An adaptive process that selects learning objects (region-specific policy structures and actors) from a digital repository and sequences them in a way which is appropriate for the targeted GIS learning community or individuals [2, 6, 11, 13, 16, 17], is also necessary for reducing computational time and gaining in objectivity and acceptance of politico-administrative conclusions. Such a method is also required as the rural development issues are characterized by complex and unpredictable informal procedures and there are no clear and common indicators for evaluating rural development policy in Europe (RUDI findings). Nevertheless, a qualitative and participative evaluation is necessary on the part of learners (students specialized in Forest Policy, Rural Sociology and Economics, Development and spatial engineering, lobbyists of interest groups etc). Many types of intelligent learning systems are available but without GIS functionality. In the GLMS proposed in this paper, five key components could be identified, which are common in most GIS systems: region-specific data acquisition, algebraic and statistical analysis, processing data of actors and regional networks, construction of political geography database, and calculating/ visualizing formal and informal hierarchies. Figure 1 depicts the interactions between these five GIS components [4, 18].

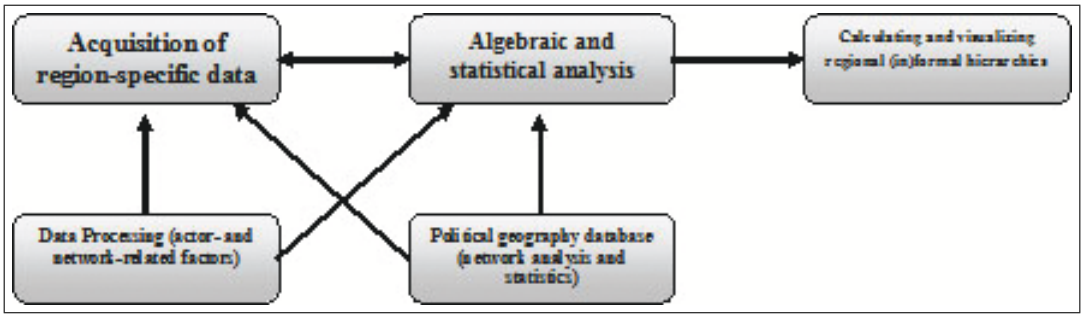


Figure 1: The main components of an intelligent GLMS.

The selection of learning content (in this case, the region-specific formal and informal hierarchies) is based on policy-research criteria depending on complex and heterogeneous cognitive styles [3, 8, 15] which may be characterized as "region-" and "administrative-based" (Archimedes findings). In this way, a wide range of learner expectations is satisfied [9, 10, 12]. Particularly, the administrative-based elements (organisational theories) are much more subjective than the regional ones. Therefore, cooperation between multidisciplinary academic and field experts (e.g. environmental, forest and agricultural scientists, public administrators, lobbyists, sociologists, informatics experts etc) is necessary in order to achieve acceptance and integrated analysis of real cases. Such a criteria set constitutes a decision support system (DSS) for learners [4] and teaching staff which enables both of them to reduce hypothetical options and produce original and accurate research results and constitutes a basic component of an intelligent GLMS [16, 19]. Although many digital DSS types have been proposed, these are applicable only to examination of human-building interaction and perceptual relations [10, 17, 18, 22, 23, 25] and they are not combined with GLMS. Other DSS types which are combined with GLMS [16,24] are strictly related to

spatial elements and not to region-specific politico-administrative issues (Archimedes findings). In this paper, the development of a politico-administrative GIS is addressed. A region- and administrative-based filtering process of learning objects is discussed. Politico-administrative metadata are proposed which can be used for learning object filtering on the basis of the Open GIS Consortium guidelines (standards) concerning GIS functionality [4, 20].

2 Digitalizing the "political geometry" in a region-specific GIS

The formal and informal hierarchy shaped in every region surveyed is composed of three power dimensions, as shown in formula (1) [21]:

$$Politico-administrative\ Power = Trust + Incentives + Uniqueness \tag{1}$$

Trust is used for leading even when surveillance is infeasible, provision of incentives is for assuring commitment, and uniqueness is useful for exerting institutional pressure. According to the RUDI findings, informal hierarchies are more decisive for the policy output than the formal ones because of the lack of detailed criteria of decision-making and evaluation in rural development policy. Trust is a relational value based on expertise, experience and personality and is accumulated through successive transfer of reputation. If e.g. the Forestry Commission trusts the Royal Scottish Forestry Society, which trusts the National Trust of Scotland and the last two actors trust the Friends of the Loch Lomond, then the last one proves to be the most trustworthy as it is able to gain the trust of all previous actors (also of the Forestry Commission indirectly). For the reputation of the actor A, it is not merely important how many actors trust A, but also how much reputation these actors gain from other actors etc. These actors can be ordered on the vertical and horizontal axis of a matrix. Thereby the network can be algebraically processed. Formula (2) which is known as Katz-status formula is applied for calculating the power status of an actor in a network:

$$T = (I - aC)^{-1} - I \tag{2}$$

where T is a matrix including the status values of all actors as elements, C is the matrix presenting the real network of trust, and a is a dumping factor. The same formula is applied in the case of the provision of incentives and uniqueness dependence relations. This algorithm is used by VISONE software.

VISONE layers vertically the actors (learning objects) according to their power status measured in % (Figure 2). The horizontal order has no politico-administrative meaning. It is obvious that in the simple polygon form, the policy networks are not disclosing any hierarchies developed in their regions. When they are layered, they acquire a pyramid-shape form. The sharper the pyramid (vertical length in relation to horizontal length), the higher the oligarchy, as defined in formula (3) [14, 21]:

$$Oligarchy = \frac{Status\ max - Status\ min}{Status\ Average} \tag{3}$$

The sharpest pyramid is this of UK1 issue network (oligarchy=2,40), while the "pyramid" of Greek network does not seem to be a pyramid at all, as the oligarchy is quite low (1,48). In Figure 2, the power status of each actor can be examined by the learners in relation to its orientation (use or conservation of natural resources) and its legal character (private or state actor). In this way, learning effects and original conclusions with academic and practical value can be made by the learners through the interpretation of this digital visualization of region-specific hierarchies with abstract but applicable politico-administrative metadata [1, 13, 20].

An output of such a GIS produced by the comparative analysis of these digital pyramids of (in)formal regional hierarchies is concisely presented in Figure 3: The power status can be examined by the learners again in relation to the legal character and to the orientation among various regions. Thereby, policy-relevant conclusions can be made regarding the winning possibility, considering these determinants (legal character and orientation).

E.g. in the Greek and Spanish regions, the private actors are much more powerful than the state ones, with noticeable difference in comparison with the European average. The inverse hierarchy can be recognized in the case of the UK1 and UK2 networks (Scotland). The conservation-oriented actors (e.g. environmental NGOs and agencies) are more powerful than the use-oriented ones in Denmark. This is a case subversive to the average power relation. The science-oriented actors (universities, research institutes) are more powerful than the other actors in the network of Finland and in one region-specific network in Spain. These results can be further interpreted by using qualitative information about the content of the policy issues (RUDI and Archimedes findings).

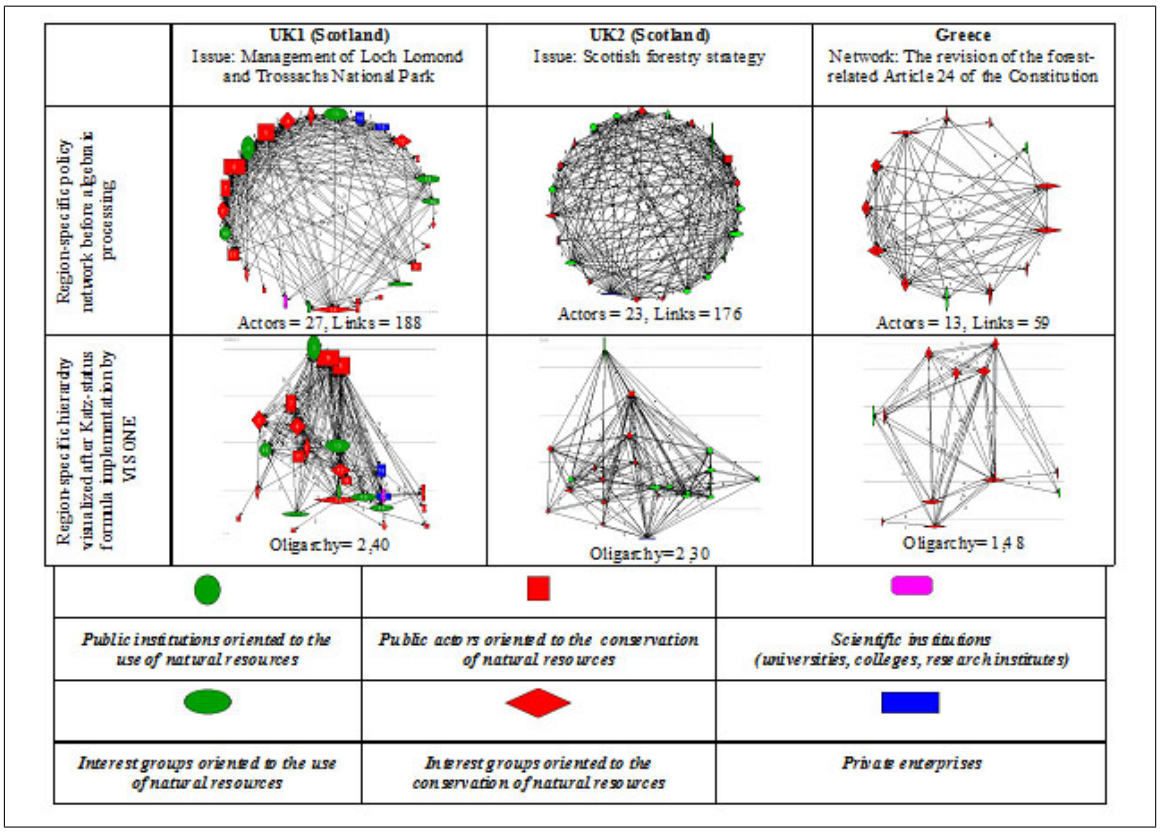


Figure 2: Region-specific networks of politico-administrative power.

A further elaboration of DSS [2,5,16] for the diagnosis of favorable (or unfavorable) participation in a region-specific network [9, 19] is possible by applying stepwise regression to this political-geographic database: Not every actor can participate in every regional network with equal chance of developing power. According to the New-Institutionalist approach the power (P) achieved by an actor does not depends only on the organizational (O) features of this actor but also on the regional network (N) conditions in which this actor is involved, as shown in the Figure 4. Which combinations of actor and regional network features lead to the optimal power? One can deduce hypotheses on the O-factors and induce the N-factors through stepwise regression [21]. In Figure 5, the procedure of the stepwise regression works as a filtering process, reduces the power-ineffective combinations and produces ideal types of politico-administrative metadata (actor- and regional network-related power determinants). An ideal type, for instance, is the following one:

An actor (e.g. an environmental NGO) with multidisciplinary team (0,284*MULTIDIS), which is not radical (-0,261*RADICALI), and has no state representatives at its board (-0,203*STATECH), can develop optimal power in a network which is composed of only a few actors (-0,427*ACTORS), provides many opportunities of lobbying (0,394*POTLOBB), is characterized by low relative importance of state (-0,296*RELIMPST), and involves only a few policy sectors (-0,243*INTERSEC).

Comparing these actor- and network specific factors with GIS outputs such as these which are described in Figure 3, it is concluded, for instance, that this type of actor described in Figure 5, has optimal chance to develop power in the regional networks of Bavaria.

O-factors were selected by using specific organisational theories (i.e. contingency model which is expressed by the absence of state representative at the board and the using of alternative expertise for surviving in heterogeneous regional- political environments). The regional N-factors were inductively selected by the stepwise regression. The combinations of O- and N-factors can also be characterized as regional-specific critical scenario analysis [2,5,7,21].

According to RUDI results, the deviation between formal and informal hierarchies and the differences between regional networks can be attributed to the inflexible bureaucracy, the complexity, the centralisation and to the lack of formal and clear criteria of decision-making and evaluation in Greece and in other European countries. Furthermore, the challenges posed by the requirement of harmonizing social, economic and environmental standards

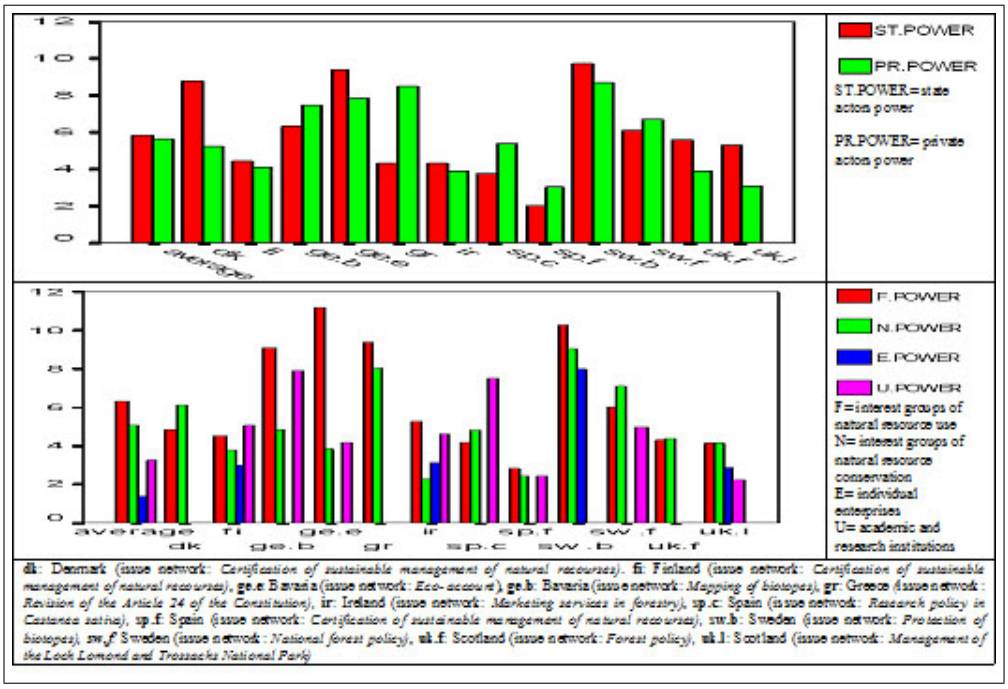


Figure 3: Politometrics-embedded GIS.

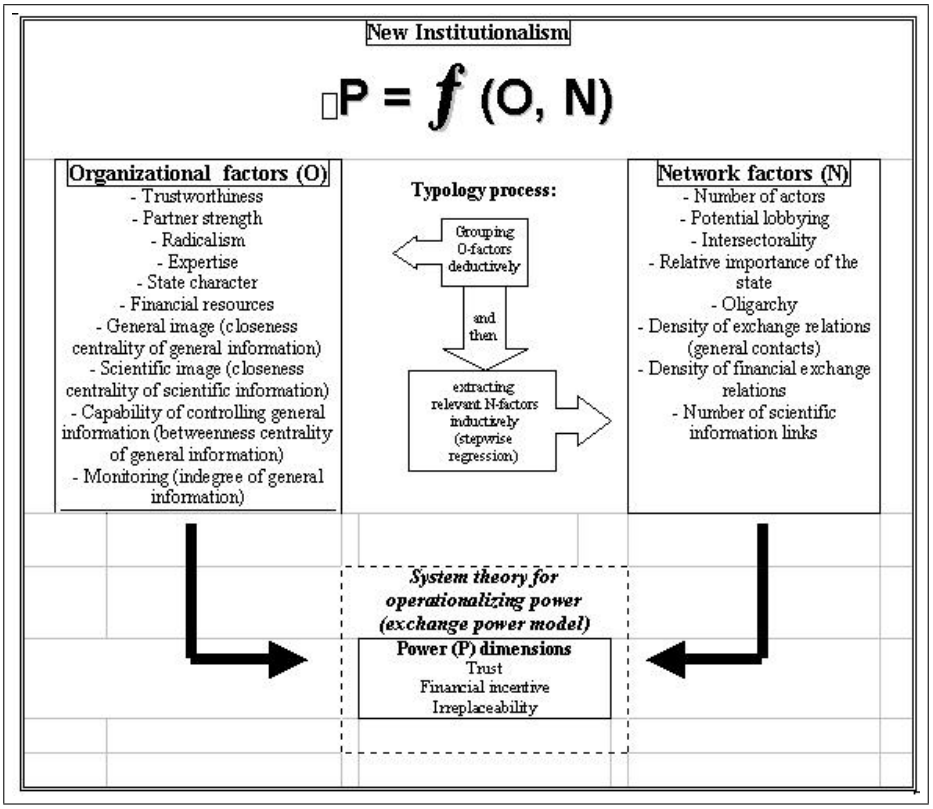


Figure 4: Optimal power synergy through politico-administrative meta-data: actor-related and region-specific power determinants.

Dependent Variable: POWER

Modell		Unstandardised coefficients		Standardised coefficients	Significance
		B	Standard error	Beta	
1	(Constant)	10,762	,965		,000
	ACTORS	-,230	,040	-,527	,000
2	(Constant)	7,969	1,064		,000
	ACTORS	-,200	,036	-,457	,000
	POTLOBB	,109	,024	,381	,000
3	(Constant)	6,647	1,151		,000
	ACTORS	-,198	,035	-,452	,000
	POTLOBB	,113	,023	,395	,000
	MULTIDIS	,354	,137	,206	,011
4	(Constant)	8,311	1,279		,000
	ACTORS	-,219	,035	-,502	,000
	POTLOBB	,111	,022	,387	,000
	MULTIDIS	,357	,132	,208	,008
	RELIMPST	-,704	,267	-,208	,010
5	(Constant)	11,702	1,727		,000
	ACTORS	-,237	,034	-,542	,000
	POTLOBB	,098	,022	,342	,000
	MULTIDIS	,401	,128	,233	,002
	RELIMPST	-,799	,259	-,236	,003
	RADICALI	-1,979	,707	-,216	,006
6	(Constant)	13,390	1,757		,000
	ACTORS	-,241	,033	-,552	,000
	POTLOBB	,103	,021	,362	,000
	MULTIDIS	,410	,123	,238	,001
	RELIMPST	-,764	,249	-,226	,003
	RADICALI	-2,574	,709	-,281	,000
	STATECH	-,627	,218	-,216	,005

Figure 5: Stepwise regression on politometrical GIS data.

complicate the policy-making.

3 Summary and Conclusions

The main goal of vector-algebra software (i.e. VISONE) is the digitalization and visualization of the formal and informal politico-administrative hierarchies of region-specific issue networks. The goal of the stepwise regression as a filtering process is the reduction of the searching space. GIS learning object repositories often contain numerous possible combinations of regional hierarchy features and actor-related characteristics. Without visual complete network analysis and statistical techniques like the stepwise regression, the examination of and familiarization with all possible learning objects would be characterized by conceptual complexity and time-consumption. This would be discouraging for the learners, especially if they were practitioners (e.g. lobbyists of environmental and industrial groups, or employees of forest services and agricultural directorates) and not only normal students. The method presented in this paper is a system for filtering learning objects which is based on knowledge domains [7, 11, 19] (i.e. organisational theories and practical experience) and seems to be appropriate for student and adult education as well.

Acknowledgements

The research initiative proposed by this paper has been supported by the EU-funded "Archimedes" Research Project (Department of Landscape Architecture, Kavala Institute of Technology, Drama, Greece), by the EU-funded research project

"RUDI: Rural Development Impacts- Assessing the impact of Rural Development policies, incl. LEADER" Consortium no: 213034, 7th Framework Programme for Research and Technological Development, (Department of Agricultural Economics, Faculty of Agricultural Science, Aristotle University of Thessaloniki, Greece), and by the Institute of Forest Policy and Nature Conservation of Goettingen University (Germany).

Bibliography

[1] M. A. Rajan, M. Girish Chandra, L.C. Reddy, P. Hiremath, *Concepts of Graph Theory Relevant to Ad-hoc Networks*. Int. J. of Computers, Communications & Control, 3, Suppl. issue: Proceedings of ICCCC, 465-469, 2008.

[2] M. Medjoudj, P. Yim, *Extraction of Critical Scenarios in a Railway Level Crossing Control System*. Int. J. of Computers, Communications and Control, 2(3), 252-268, 2007.

[3] A. Anohina, M. Vilkelis, R. Lukasenko, *Incremental Improvement of the Evaluation Algorithm in the Concept Map Based Knowledge Assessment System*. Int. J. of Computers, Communications & Control, 4(1), 6-16, 2009.

[4] A.D. Styliadis, I.D. Karamitsos, D.I. Zachariou, *Personalized e-Learning Implementation - The GIS Case*. International Journal of Computers, Communications & Control, 1(1), 59-67, 2006

[5] P. Dolog, N. Henze, W. Nejdl & M. Sintek, *Personalization in Distributed eLearning Environments*. Proc. of the 13th International World Wide Web Conference, New York, USA, 2004.

[6] M. Krott, N.D. Hasanagas, *Measuring bridges between sectors: Causative evaluation of cross-sectorality*. Forest Policy and Economics, 8(5), 555-563, 2006.

[7] M. Stanojevic, M. Vujosevic, B. Stanojevic, *Number of Efficient Points in some Multiobjective Combinatorial Optimization Problems*. Int. J. of Computers, Communications & Control, 3, Suppl. issue: Proceedings of ICCCC, 497-502, 2008.

[8] Y. Birot, G. Buttoud, R. Flies, K. Hogl, M. Pregernig, R. Päivinen, I. Tikkanen, M. Krott Voicing interests and concerns: institutional framework and agencies for forest policy research in Europe. Forest Policy and Economics, 4(4), 333-350, 2002.

[9] N. Magnani, L. Struffi, *Translation sociology and social capital in rural development initiatives. A case study from the Italian Alps*. Journal of Rural Studies, 25(2), 231-238, 2009.

[10] J. Murdoch, *Networks - a new paradigm of rural development?*. Journal of Rural Studies, 16(4), 407-419, 2000.

[11] G. McCalla,*The Fragmentation of Culture, Learning, Teaching and Technology: Implications for the Artificial Intelligence in Education Research Agenda in 2010*. International Journal of Artificial Intelligence in Education, 11, 177-196, 2000.

[12] LSAL, SCORM Best Practices Guide for Content Developers, 2003. Cernegie Mellon Learning Systems Architecture Lab. Retrieved in September 2005 from the World Wide Web: http://www.lsal.cmu.edu/lsal/expertise/projects/developersguide

[13] Kinshuk, R. Oppermann, A. Patel & A. Kashihara, *Multiple Representation Approach in Multimedia based Intelligent Educational Systems*. Artificial Intelligence in Education Journal, Amsterdam: IOS Press. 259-266, 1999.

[14] A. Real T., N.D. Hasanagas, *Complete Network Analysis in Research of Organized Interests and Policy Analysis: Indicators, Methodological Aspects and Challenges*. Connections, 26(2), 89-106, 2005.

[15] P.R. Polsani, *Use and Abuse of Reusable Learning Objects*. Journal of Digital information, (2003). Retrieved in September 2005 from the World Wide Web: http://jodi.ecs.soton.ac.uk/Articles/v03/ i04/Polsani.

[16] A.D. Styliadis, *E-Learning Documentation of Historical Living Systems with 3-D Modeling Functionality*. INFORMATICA, 18(3), 419-446, 2007.

[17] A.D. Styliadis, P.G. Patias, N.C. Zestas, *3-D Computer Modeling with Intra-Component, Geometric, Quality and Topological Constraints*. INFORMATICA, 14(3), 375-392, 2003.

[18] A.D. Styliadis, M.Gr. Vassilakopoulos, *A spatio-temporal geometry-based model for digital documentation of historical living systems*. Information & Management, 42, 349-359, 2005.

[19] I.J. Terluin, *Differences in economic development in rural regions of advanced countries: an overview and critical analysis of theories*. Journal of Rural Studies, 19(3), 327-344, 2003.

[20] M. S. Urban & E. G. Barriocanal, *On the Integration of IEEE-LOM Metadata Instances and Ontologies*. Learning Technology Newsletter, 5(1), 2003.

[21] N.D. Hasanagas, *Power factor typology through organisational and network analysis. Usinhg environmental policy networks as an illustration*. Ibidem. Stuttgart. 2004.

[22] A.D. Styliadis, *Digital documentation of historical buildings with 3-d modeling functionality*. Automation in Construction, 16, 498-510, 2007.

[23] A.D. Styliadis, *Historical photography-based computer-aided architectural design: Demolished buildings information modeling with reverse engineering functionality*. Automation in Construction, 18, 51-69, 2008.

[24] A.D. Styliadis, I.I. Akbaylar, D.A. Papadopoulou, N.D. Hasanagas, S.A. Roussa, L.A. Sexidis, *Metadata-based heritage sites modeling with e-learning functionality*. Journal of Cultural Heritage, 10, 296-312, 2009.

[25] A.D. Styliadis, D.G. Konstantinidou, K.A. Tyxola, *ECAD System Design - Applications in Architecture*. Int. J. of Computers, Communications & Control, 3(2), 204-214, 2008.

Nikolaos D. Hasanagas Born in 1974. Assistant Professor in environment-related subjects at the Kavala Institute of Technology, Drama, Greece. BSc and MSc eq. in Environmental Sciences (Aristotle Univ. of Thessaloniki, Greece), BA and MA eq. in Social Sciences, PhD in Environmental Policy Analysis (Goettingen Univ., Germany).

Athanasios D. Styliadis Born in 1956. Professor of digital architecture and design computing at the Department of Landscape Architecture at the Kavala Institute of Technology, Drama, Greece. Diploma in Surveying Engineering, MSc in Computer Science (Dundee Univ., Scotland), PhD in CAAD and GIS (Aristotle Univ. of Thessaloniki, Greece).

Eleni I. Papadopoulou Born in 1957. Assistant Professor in Rural Policy at the Faculty of Agricultural Science at the Aristotle University of Thessaloniki, Greece. BSc in Agriculture Engineering, MSc in Agricultural Economics (Univ. of Reading, UK), PhD (Aristotle University of Thessaloniki).

Lazaros A. Sechidis Born in 1968. Assistant Professor in Geodesy at the Department of Landscape Architecture at the Kavala Institute of Technology, Drama, Greece. Dipl. in Surveying Engineering, PhD in Photogrammetry (Aristotle Univ. of Thessaloniki).

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 4, pp. 525-531

Fingerprints Identification using a Fuzzy Logic System

I. Iancu, N. Constantinescu, M. Colhon

Ion Iancu, Nicolae Constantinescu, Mihaela Colhon
Department of Informatics
University of Craiova,
Al.I. Cuza Street, No. 13, Craiova RO-200585, Romania
E-mail: i_iancu@yahoo.com, nikyc@central.ucv.ro, mghindeanu@yahoo.com.

Abstract: This paper presents an optimized method to reduce the points number to be used in order to identify a person using fuzzy fingerprints. Two fingerprints are similar if n out of N points from the skin are identical. We discuss the criteria used for choosing these points. We also describe the properties of fuzzy logic and the classical methods applied on fingerprints. Our method compares two matching sets and selects the optimal set from these, using a fuzzy reasoning system. The advantage of our method with respect to the classical existing methods consists in a smaller number of calculations.
Keywords: fuzzy models, fingerprint authentication, cryptographic signature model.

1 Introduction

Fingerprint identification is the most mature biometric method being implemented at an early level since 1960. The recognition of a fingerprint can be done with two methods: "one-to-one" (verification) and "one-to-many" ($1:N$ identification). The first method is applied when we have two fingerprints and we want to verify if they belong to the same person. The second one is used when we have one fingerprint and we search it in a data base. The verification is much easier and faster because we have the two fingerprints and we just need to compare them. On the other hand, the identification implies more time for extracting the fingerprint because there are needed much more details.

The fingerprints are not compared with images, they use a method based on characteristic points named "minutiae". These points are characterized by *ridge ending* (the abrupt end of a ridge), *ridge bifurcation* (a single ridge that divides in two ridges), *delta* (a Y-shaped ridge meeting), *core* (a U-turn in ridge pattern), etc. All these features are grouped in three types of lines: *line ending*, *line bifurcation* and *short line*. After the minutiae points are localized, a map with all their locations on the finger is created. Every minutiae point has associated two coordinates (x,y), an angle for orientation and a measure for the fingerprint quality. The matching of two fingerprints depends on the position and on the rotation. For this reason, every fingerprint is represented, not only, as a group of points with two coordinates, but also, as a group of points with coordinates relative to other points. This allows obtaining an unique positioning of a point regarding to other three points. The three selected points must not be collinear. When two fingerprints are compared, first are compared the relative coordinates. If this stage ends successfully, these coordinates are transformed in 2D coordinates and verified.

After verifying the fingerprints, the result will tell us if they are from the same person with a high probability. Still, the cases when the belonging probability of a fingerprint is 0 (false) or 1(true) are rarely. In most of the cases, the probability will be a number $p \in [0,1]$. This fact leads to a fuzzy logic. The values in fuzzy logic can range between 0 and 1 (1 is for absolute truth, 0 for absolute falsity). A fuzzy value for an element x will express the degree of membership of x in a set X. It is essential to realize that fuzzy logic uses truth degrees as a mathematical model of the vagueness phenomenon while probability is a mathematical model of randomness.

2 State of the Art

Two fingerprints are similar if n out of N points match. To verify this, Freedman et al. introduced the fuzzy matching protocols [3]. Using these protocols, the information about the fingerprint we want to identify (or verify) will not be revealed if no match is found. To describe the fuzzy private matching problem we will take a set of words $X = x^1 \dots x^N$ where $\{x^i\}$ are the letters. Two words $X = x^1 \dots x^N$ and $Y = y^1 \dots y^N$ match only if: $n \leq |\{k : x^k = y^k \mid 1 \leq k \leq N\}|$ and this relation is denoted with $X \approx_n Y$. In the subsequent we will name the set X as the *total set for selection*. The input of the protocol will be two sets of words ($X = X_1 \dots X_m$ for the client and $Y = Y_1 \dots Y_s$ for the server) and the parameters m, s, N and n. While the output of the server is empty, the output of the client will be a set $\{Y_i \in Y | \exists X_i \in X : X_i \approx_n Y_i\}$, where $A \approx_n B$ means that the points A and B are very close. This set is, in fact, the intersection of the two input sets [1]. It was demonstrated that this protocol leads information about the input even if no match is found [1]. Another protocol, based on Freedman's protocol, was presented in [1]. It uses σ as a combination of n different indices $\gamma_1, \gamma_2 \dots \gamma_n$ and $\sigma(X) = x^{\gamma_1} || \dots || x^{\gamma_t}$ for a word X. After the parameters and the public key are sent, the client constructs a polynomial representation of the points set:

$$P_\sigma = (x - \sigma(X_1)) * (x - \sigma(X_2)) * \dots * (x - \sigma(X_m))$$

This is a feedback polinomial value for a set of fingerprints. Then he sends $\{P_\sigma\}_{k_2}$ to the server. The server analyzes every received polynomial $\{P_\sigma\}$ at the point $\sigma(Y_i)$ and computes $\{w_i^\sigma\}_{k_2} = \{r * P_\sigma(\sigma(Y_i)) + Y_i\}_{k_2}$, where r is a random value. After all the calculations, the server sends $\{w_i^\sigma\}_{k_2}$ to the client. The client will decrypt all the messages and if w_i^σ matches with any word from X then it is added to the output set. $\{w_i^\sigma\}_{k_2}$ is a combination between fingerprint points value and the parameter which characterizes the common information between a base set and the current set of collected values.

A particular scheme of fingerprint authentication describes a method which is not based on the minutiae points [12], but by the texture of the finger, called FingerCode. Such a FingerCode is a vector composed from 640 values between 0 and 7. The vector is ordered and stable in size. The method uses Euclidean distance to find the matching. After estimating the block orientation, a curvature estimator is designed for each pixel. Its maximal value is, in fact, the morphological searched center. Using a properly tuned Gabor filter ([11, 13]) we can catch ridges and valleys from the fingerprint. The FingerCode is computed as the average absolute deviation from the mean of every sector of each image. Error-correction for the FingerCode would never be efficient enough to recognize a user. In [12], the method proposed uses a secret $d + 1 - letter\ word$, which correspond to the $d + 1$ coefficients of a polynomial p of degree d. The public key will be extract from (F, p), where F is the FingerCode. Then, we choose n random point of p. These points will be hidden like in a fuzzy commitment scheme. To find the polynomial p, each point is decoded. If at least $d + 1$ points are decoded then p can, also, be retrieved.

A method based on the minutiae points and, also, on the pattern of the finger was presented in [4]. All the ridges that cross a line (x, y) where x and y are minutiae points are counted. Then, are presented all the possible combinations of three minutiae points and the ridges crossing that line. Such a combinations' list needs C_n^3 entries, where n is the number of minutiae points. This method is more complex because before all the calculations are done we need to identify the minutiae points and then combine them.

3 Our Method

3.1 System Description

A commercial fingerprint-based authentication system requires a very low False Reject Rate (FRR) for a given False Accept Rate (FAR) where FAR is *the probability that the system will incorrectly identify* and FRR is *the probability of failure in identification*.

Our method is, also, based on the minutiae points of the fingerprints. We can identify at least 40 minutiae points on a fingerprint, depending on its quality. In general, the number of the minutiae points varies from 0 to 100. All the methods mentioned above can be applied to a fingerprint verification. But, for an identification we need an algorithm with a low level of complexity because the data bases used in practice have millions of fingerprints. To reduce the search time and complexity, we first propose to classify the fingerprints, and then, to identify the input fingerprints only in one subset of the data base. To choose the right subset the fingerprint is matched at a coarse level to one of the existing types. After that, it is matched at a finer level to all the fingerprints of the subset. The FBI in the United States recognize eight different types of patterns [5]. For example, we have an input fingerprint and we want to identify it in a data base with 15000 entries. We will take the minim number of minutiae points, 40. If no classification is made we have to do at least $40 \times 15000 = 600000$ operations. But, if we use a classification with eight types (each subset has the same number of fingerprints $15000/8 = 1875$) we will have at least $(8+1875) \times 40 = 75320$ calculations. This is because we will first compare the input fingerprint with each group and after that it will be compared with each element of the chosen group. As we can see, the calculations are reduced to only $12,5\%$. The classification of the fingerprints is preferred to have more than three types of subsets. This is because a higher accuracy is achieved. Such a classification, also, helps to reduce the number of calculations with a higher percentage.

3.2 Fuzzy Mathematical Background

A fuzzy set A in X is characterized by its membership function:

$$\mu_A : X \to [0,1]$$

where $\mu_A(x) \in [0,1]$ represents the membership degree of the element x in the fuzzy set A. We will work with membership functions represented by trapezoidal fuzzy numbers. Such a number $N = (\underline{m}, \overline{m}, \alpha, \beta)$ is defined as

$$\mu_N(x) = \begin{cases} 0 \ \ for \ \ x < \underline{m} - \alpha \\ \frac{x - \underline{m} + \alpha}{\alpha} \ \ for \ \ x \in [\underline{m} - \alpha, \underline{m}] \\ 1 \ \ for \ \ x \in [\underline{m}, \overline{m}] \\ \frac{\overline{m} + \beta - x}{\beta} \ \ for \ \ x \in [\overline{m}, \overline{m} + \beta] \\ 0 \ \ for \ \ x > \overline{m} + \beta \end{cases}$$

The rules are represented by fuzzy implications. Let X and Y be two variables whose domains are U and V, respectively. The rule

$$if \ X \ is \ A \ then \ Y \ is \ B$$

is represented by its conditional possibility distribution ([14], [15]) $\pi_{Y/X}$:

$$\pi_{Y/X}(v,u) = \mu_A(u) \to \mu_B(v), \ \forall u \in U, \ \forall v \in V$$

where $\to$ is an implication operator ([2]) and μ_A and μ_B are the membership functions of the fuzzy sets A and B, respectively. One of the most important implication is Lukasiewicz implication [2], $I_L(x,y) = \min(1 - x + y, 1)$.

3.3 Proposed Fuzzy Logic System

Fuzzy control provides a formal methodology for representing, manipulating and implementing human's heuristic knowledge about how to control a system. In a fuzzy logic controller, the expert knowledge is of the form

IF (*a set of conditions are satisfied*) *THEN* (*a set of consequences are inferred*)
where the antecedents and the consequences of the rules are associated with fuzzy concepts (linguistic terms). The most known systems are: Mamdani, Tsukamoto, Sugeno and Larsen which work with crisp data as inputs. A Mamdani type model which works with interval inputs is presented in [10].

In this paper we use a version of Fuzzy Logic Control (FLC) system from [9] in fingerprints identification. This version is characterized by:

- the *linguistic terms* (or values), that are represented by trapezoidal fuzzy numbers
- *Lukasiewicz implication*, which is used to represent the rules
- the *crisp control action of a rule*, computed by Middle-of-Maxima method
- the *overall crisp control actions*, computed by discrete Center-of-Gravity.

We assume that the facts can be given by crisp data, intervals and/or linguistic terms and a rule is characterized by:

- a set of linguistic variable A, having as domain an interval $I_A = [a_A, b_A]$
- n_A linguistic values $A_1, A_2, ..., A_{n_A}$ for each linguistic variable A
- membership function $\mu^0_{A_i}(x)$ for each value A_i, where $i \in \{1, 2, ..., n_A\}$ and $x \in I_A$.

According to the structure of a FLC, the following steps are necessary in order to work with our system.

Firing levels

We consider an interval input $[a,b]$ with $a_A \leq a < b \leq b_A$. The membership function of A_i is modified ([10]) by membership function of $[a,b]$ as follows

$$\forall x \in I_A, \mu_{A_i}(x) = min(\mu^0_{A_i}(x), \mu_{[a,b]}(x))$$

where

$$\mu_{[a,b]}(x) = \begin{cases} 1 & if \quad x \in [a,b] \\ 0 & otherwise \end{cases}$$

It is obvious that, any t-norm T can be used instead of *min* (see, for instance, [6–8]).

The firing level, generated by the input interval $[a,b]$, corresponding to the linguistic value A_i is given by:

$$\mu_{A_i} = \max\{\mu_{A_i}(x) | x \in [a,b]\}.$$

The firing level μ_{A_i}, generated by a linguistic input value A'_i is

$$\mu_{A_i} = \max\{\min\{\mu^0_{A_i}(x), \mu_{A'_i}(x)\} | x \in I_A\}.$$

The firing level μ_{A_i}, generated by a crisp value x_0 is $\mu^0_{A_i}(x_0)$.

Fuzzy inference

We consider a set of fuzzy control rules

$$R_i : if\ X_1\ is\ A^1_i\ and\ ...\ and\ X_r\ is\ A^r_i\ then\ Y\ is\ C_i$$

where the variables $X_j, j \in \{1, 2, ..., r\}$, and Y have the domains U_j and V, respectively. The firing levels of the rules, denoted by $\{\alpha_i\}$, are computed by

$$\alpha_i = T(\alpha^1_i, ..., \alpha^r_i)$$

where T is a t-norm and α^j_i is the firing level for $A^j_i, j \in \{1, 2, ..., r\}$. The conclusion inferred from the rule R_i, using the Lukasiewicz implication is

$$C'_i(v) = I(\alpha_i, C_i(v)), \forall v \in V.$$

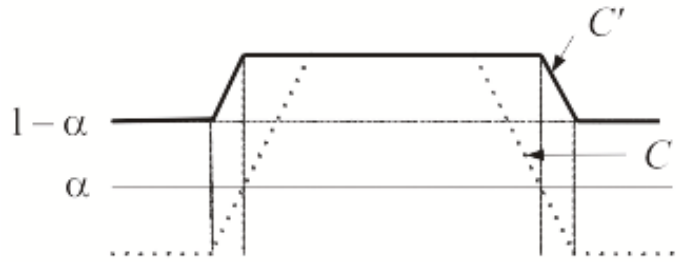


Figure 1: Conclusion obtained with Lukasiewicz implication

Defuzzification

The fuzzy output C_i' of the rule R_i is transformed in a crisp output z_i using the Middle-of-Maxima operator. The crisp value z_0 associated to a conclusion C' inferred from a rule having the firing level α and the conclusion C represented by the fuzzy number $(\underline{m}_C, \overline{m}_C, \alpha_C, \beta_C)$ is:

$$z_0 = \frac{\underline{m}_C + \overline{m}_C + (1-\alpha)(\beta_C - \alpha_C)}{2}$$

The overall crisp control action is computed by the discrete Center-of-Gravity method: if the number of fired rules is N then the final control action is

$$z_0 = \left(\sum_{i=1}^{N} \alpha_i z_i \right) / \sum_{i=1}^{N} \alpha_i$$

where α_i is the firing level and z_i is the crisp output of the i-th rule.

4 An application in fingerprint identification

For the proposed FLC we consider rules with two inputs and one output. The input variables are $\sigma_1 = \{x_i | x_i \in X\}$ and $\sigma_2 = \{x_j | x_j \in X, x_i \neq x_j, \forall i, j\}$ where X is the set defined in Section 2. By σ_1 we represent the values set of basic input data and σ_2 is a user data to be evaluated and authenticated. These values sets will be denoted by S_1 and S_2 respectively. The σ set values denote the optimal points set to be used for the fuzzy matching authentication, according with S output variable. The fuzzy rule-base consists of

R1: If S_1 is *Low* and S_2 is *Very Low* then S is *Low*
R2: If S_1 is *Very Low* and S_2 is *Low* then S is *Low*
R3: If S_1 is *Very Low* and S_2 is *Very Low* then S is *Very Low*
R4: If S_1 is *Very Low* and S_2 is *Very Low* then S is *Low*
R5: If S_1 is *Very Low* and S_2 is *Middle* then S is *Middle*
R6: If S_1 is *Very Low* and S_2 is *Middle* then S is *Low*
R7: If S_1 is *High* and S_2 is *Very High* then S is *High*
R8: If S_1 is *Very High* and S_2 is *High* then S is *High*
R9: If S_1 is *Very High* and S_2 is *Very High* then S is *High*
R10: If S_1 is *Low* and S_2 is *Very High* then S is *Middle*

The value *Very Low* for a variable S_1, S_2 or S represents a minimum degree of trust while the value *Very High* represents the maximum degree. There are five linguistic values for every variable

$$\{Very\ Low,\ Low,\ Middle,\ High,\ Very\ High\}.$$

We consider the universes of discourse $[0, 10]$. The membership functions corresponding to the linguistic values are represented by the following trapezoidal fuzzy numbers:

- for the variable $S1$: $\{(0,1,0,1.5),(3,4,1,0.5),(5,6,1,0.5),(7.5,8.5,3,1),(9.5,10,0.5,0)\}$
- for the variable $S2$: $\{(0,1.5,0,1),(2.5,4,0.5,0.5),(5,6,2,0),(6.5,8,1.5,0),(8.5,10,0.5,0)\}$,
- for the variable S: $\{(0,1,0,1.5),(2.5,4,0.5,0.5),(5,7,2.5,0.5),(8,8.5,2,0.5),(9.5,10,1,0)\}$

We consider the following interval input values: $[1.5,2.2]$ for $S1$ and $[3.2,4.2]$ for $S2$. The positive firing levels corresponding to the linguistic values of the input variable $S1$ are

$$\mu_{VeryLow} = 0.666, \mu_{Low} = 0.2$$

and the positive firing levels corresponding to the linguistic values of the input variable $S2$ are:

$$\mu_{Low} = 1, \mu_{Middle} = 0.6$$

The fired rules and their firing levels, computed with t-norm Product $T(x,y) = xy$, are:

$$R_2 \text{ with firing level } \alpha_2 = 0.666,$$
$$R_5 \text{ and } R_6 \text{ with } \alpha_5 = \alpha_6 = 0.3996.$$

The fired rules give the following crisp values as output:

$$z_2 = 3.25,\ z_5 = 5.3996,\ z_6 = 3.25;$$

then the overall crisp control action is

$$z_0 = 3.836.$$

These values represent the matching approach for every subset points which are candidate to be in the final set, and are computed using a fuzzy merging comparison between selection sets σ_1 and σ_2. The optimal selection set (which has less points) is represented by the output variable σ.

5 Conclusions

Among all the biometric techniques, the identification based on fingerprints is used in the most applications. The uniqueness of the fingerprint can be determinate by the pattern of ridges and the minutiae points. For identifying an input fingerprint, the proposed method uses a fuzzy classification of the data. The proposed system is much more efficient than the FLC presented in [8]. This is because, in order to reduce the necessary points number, we find the minutiae points by using a fuzzy logic reasoning system which compare two points sets matching values. In a practical application, it is recommended to use the proposed system with a set of implications and aggregate the results given by every implication, in order to obtain the overall output; in this way can be obtained a stronger base for more accurate results of our system. We intend to use these results in a future work, by maping their relative placement on the finger, and comparing all its points with the ones of the fingerprints for the right subset.

Bibliography

[1] L. Chmielewski and J. H. Hoepman, Fuzzy Private Matching (Extended Abstract), *ARES '08: Proceedings of the 2008 Third International Conference on Availability, Reliability and Security*, IEEE Computer Society, pp. 327–334, 2008.

[2] E. Czogola and J. Leski, On equivalence of approximate reasoning results using different interpretations of fuzzy if-then rules, *Fuzzy Sets and Systems*, vol. 117, no. 2, pp. 279–296, 2001.

[3] M. Freedman, K. Nissim and B. Pinkas, Efficient private matching and set intersection, *Advances in Cryptology, EUROCRYPT 2004*, Springer-Verlag, pp. 1–19, 2004

[4] R. S. Germain, A. Califano and S. Colville, Fingerprint Matching Using Transformation Parameter Clustering, *IEEE Comput. Sci. Eng.*, vol. 4, no. 4, pp. 42–49, 1997.

[5] M. R. Hawthorne, *Fingerprints. Analysis and Understanding*, CRC Press, 2009

[6] I. Iancu, T-norms with threshold, *Fuzzy Sets and Systems. International Journal of Soft Computing and Intelligence*, vol. 85, no. 1, pp. 83–92, 1997.

[7] I. Iancu, Operators with n-thresholds for uncertainty management, *Journal of Applied Mathematics & Computing*, Springer Berlin, vol. 19, no. 1-2, pp. 1–17, 2005.

[8] I. Iancu, Generalized Modus Ponens Using Fodor's Implication and T-norm Product with Threshold, *International Journal of Computers, Communications & Control (IJCCC)*, vol. 4, no. 4, pp. 330–343, 2009.

[9] I. Iancu, Extended Mamdani Fuzzy Logic Controller, *The Fourth IASTED International Conference on Computational Intelligence CI 2009*, ACTA Press, vol. 5, pp. 143–149, 2009.

[10] F. Liu, H. Geng and Y. Q. Zhang, Interactive Fuzzy Interval Reasoning for Smart Web Shopping, *Applied Soft Computing, Elsevier*, vol. 5, no. 4, pp. 433–439, 2005.

[11] D. G. Radojevic, Fuzzy Set Theory in Boolean Frame, *International Journal of Computers, Communications & Control (IJCCC)*, vol. 3, no. 5, pp. 121–131, 2008.

[12] V. V. T. Tong, H. Sibert, J. Lecoeur and M. Girault, Biometric fuzzy extractors made practical: a proposal based on Finger Codes, *Advances in Biometrics*, Springer Berlin / Heidelberg, pp. 604–613, 2009.

[13] T. Vesselenyi, S. Dzitac, I. Dzitac and M.J. Manolescu, Fuzzy and Neural Controllers for a Pneumatic Actuator, *International Journal of Computers, Communications & Control (IJCCC)*, vol. 4, no. 2, pp. 375–387, 2007.

[14] L. A. Zadeh, A theory of approximate reasoning, *Machine Intelligence 9, Elsevier*, pp. 149–194, 1979.

[15] L. A. Zadeh, Fuzzy sets as a basis for a theory of a possibility, *Fuzzy Sets and Systems*, vol. 100, pp. 9–34, 1999.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 4, pp. 532-539

A Modeling Method of JPEG Quantization Table for QVGA Images

G.-M. Jeong, J.-D. Lee, S.-I. Choi, D.-W. Kang

Gu-Min Jeong
School of Electrical Engineering
Kookmin University, Seoul, Korea
E-mail: gm1004@kookmin.ac.kr

Jong-Duck Lee
Avionics Center
LIG Nex1, Daejeong, Korea
E-mail: jdlee81@lignex1.com

Sang-Il Choi
School of Electrical Engineering and Computer Science
Seoul National University, Seoul, Korea
E-mail: kara@csl.snu.ac.kr

Dong-Wook Kang
Corresponding author
School of Electrical Engineering
Kookmin University, Seoul, Korea
E-mail: dwkang@kookmin.ac.kr

Abstract: This paper presents a new JPEG quantization table design method for mobile phone images. Although the screen size of mobile phones is very small, the full information of the image should nevertheless be represented. Moreover, the high frequency components of the mobile phone images may contain important information. In order to enhance the performance of mobile JPEG images, these high frequency components should be compensated using an appropriate quantization table. Considering these characteristics, we propose a modeling method of the quantization table for compensating the high frequency components of the mobile images while sacrificing their low frequency components. We select the optimized pre-emphasis factor and bias factor using various sets of 240× 320 images and show that the proposed method improves the performance in terms of size and PSNR.
Keywords: JPEG, Quantization Table, Mobile QVGA Image, Frequency Compensation

1 Introduction

In still image coding, JPEG [1] [2] has become a *de facto* standard and shows good performance for digital pictures. In the case of mobile phone images, JPEG is also widely used. Nowadays, although VGA (480×640) or SVGA (600×800) screens are already used in smartphones, QVGA (240×320) LCDs, which are very small compared to those used in PCs or digital cameras, are generally adopted in handsets. For these reasons, the handset images may have different characteristics from those of PCs or digital cameras [3].

In mobile phone images, the whole information must be described within a small image size. When comparing small and large size images of the same scene, the frequency characteristics in the 8×8 blocks can be different from each other. That is, the effect of the high frequency components can be increased compared to that of the low frequency components in small size images. Therefore, in handset images,

especially QVGA images, the high frequency components can be relatively more important than in PC or digital camera images, due to the differences in the image size. Considering these characteristics, in order to improve the image quality and compression ratio for mobile images, it is possible to design a specific quantization table by decreasing the quantization values for the high frequency components and increasing those for the low frequency components.

In this paper, we propose a new JPEG quantization table design for 240×320 mobile images extending the results in [4]. We present a new modeling scheme of the quantization table by considering the characteristics of mobile images and optimize the pre-emphasis factor and the bias factor using the 240×320 images which are serviced by a telecommunication company [5]. Especially, in contrast to R-D optimization [6], we model the quantization table by making full use of the standard quantization table. There is no need to send the quantization table, as in the case of R-D optimization. Since only the pre-emphasis factor and the bias factor are needed in the proposed method, it can be simply applied to the JPEG encoder/decoder. The simulation results show that the proposed method works well.

The remainder of this paper is organized as follows. In Section 2, the characteristics of mobile images are conceptually discussed. In Sections 3 and 4, we present the proposed quantization table modeling scheme and pre-emphasis factor optimization using tests, respectively. The conclusions are presented in Section 5.

2 The characteristics of handset images

Recently, with the support of 3G wireless communication which provides a high speed data rate, there is a growing demand to increase the size of the screen in order for the user to enjoy various multimedia contents. However, to guarantee the portability and mobility of mobile phones, the extent to increase the screen size of mobile phones is limited.

(a) (b)

Figure 1: Sample image and $1/4 \times 1/4$ image

(a) (b)

Figure 2: The right-lower blocks of the images in Fig. 1(a) and Fig. 1(b)

The small size of the mobile phone causes the frequency characteristics of mobile images to differ from those of PC or digital camera images. For example, let us consider the images shown in Fig. 1. Fig. 1(a) and Fig. 1(b) are 512 × 512 and 128 × 128 images, respectively. Fig. 2(a) and Fig. 2(b) are the

right-lower 8×8 blocks in those images in Fig. 1. As seen in Fig. 2, the right-lower 8×8 block in Fig. 2(b) has more high frequency components than that in Fig. 2(a).

As can be seen in Fig. 1 and Fig. 2, we can think conceptually that the importance of the high frequency components increases in the 8×8 blocks for the small size images.

Likewise, to design the JPEG quantization table for mobile QVGA images, the high frequency components should be compensated. However, compensating the high frequency components may increase the compressed file size. For these reasons, we propose a quantization table modeling scheme for compensating the high frequency components while sacrificing the low frequency components in order not to increase the compressed file size. In this way, we hope to achieve a better PSNR and bpp than those obtained using the standard table for QVGA images.

3 Quantization table modeling for mobile images

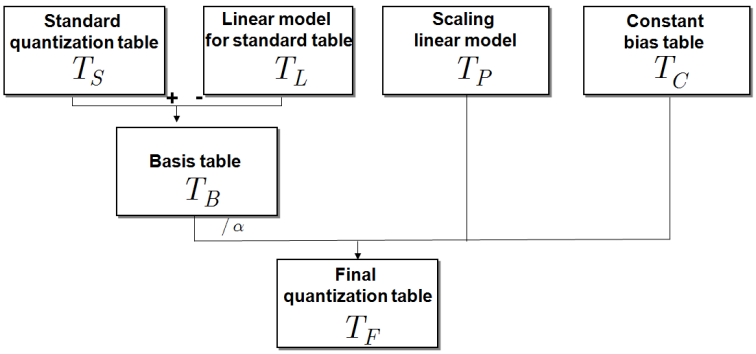


Figure 3: The proposed modeling scheme of quantization table

We propose a new modeling method for the JPEG quantization table used for mobile images which compensates the high frequency components. Fig. 3 shows the overall design of the proposed method. To make full use of the standard quantization table, we obtain the final quantization table T_F from the standard quantization table T_S, pre-emphasis factor α and bias factor β.

First, the linear model of the standard table T_L is derived from T_S. We obtain the basis table T_B by subtracting T_L from T_S. Next, the scaling linear model T_P and the constant bias table T_C are calculated from the pre-emphasis factor α and the bias factor β respectively. As a result, the final quantization table T_F is derived using $T_F = T_P + T_B/\alpha + T_C$.

Let us describe the proposed modeling method in more detail. We set $T_L(1,1) = T_S(1,1)$ and $T_L(8,8) = T_S(8,8)$. Based on $T_L(1,1)$ and $T_L(8,8)$, we can obtain T_L using (1). Using T_S and T_L, T_B is calculated as $T_B = T_S - T_L$. Note that (1) is also used for the calculation of T_P.

$$\begin{aligned} T(x,x) &\equiv T(x) = \frac{T(8)-T(1)}{7}(x-1)+T(1), \text{if } x=y \\ T(x,y) &= T(\frac{x+y}{2}), \qquad \text{if } x+y \text{ is even} \\ T(x,y) &= \frac{T(\frac{x+y-1}{2})+T(\frac{x+y+1}{2})}{2}, \qquad \text{if } x+y \text{ is odd} \end{aligned} \tag{1}$$

where $T(x,y)$ means the (x,y) component of the 8×8 matrix table T.

In the proposed modeling, we compensate for the high frequency components by increasing the low frequency components of the table and decreasing the high frequency components of the table. Therefore,

we set $T_P(1,1)$ and $T_P(8,8)$ according to the pre-emphasis factor α as follows:

$$T_P(1,1) = \alpha T_L(1,1), \quad T_P(8,8) = \frac{1}{\alpha} T_L(8,8).$$

T_P is also a linear model and can be calculated from $T_P(1,1)$ and $T_P(8,8)$ using (1). Also T_C is a constant bias table using bias factor β, which is obtained by $T_C = \beta I$.

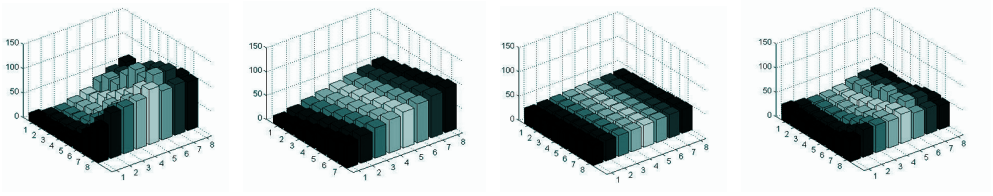
We can obtain the final quantization table T_F as follows:

$$T_F = T_P + T_B/\alpha + T_C = T_P + (T_S - T_L)/\alpha + T_C. \tag{2}$$

As α increases, the effect of the high frequency components increases, while that of the low frequency components decreases. Also, the detailed bias is adjusted using β.

Table 1: T_S, T_L, T_P and T_F when $\alpha = 2, \beta = 0$

16	11	10	16	24	40	51	61	16	21	27	33	39	45	51	57	32	33	34	35	37	38	39	40	32	28	25	26	29	35	39	42
12	12	14	19	26	58	60	55	21	27	33	39	45	51	57	63	33	34	35	37	38	39	40	42	28	26	25	27	28	42	41	38
14	13	16	24	40	57	69	56	27	33	39	45	51	57	63	69	34	35	37	38	39	40	42	43	27	25	25	27	33	40	45	36
14	17	22	29	51	87	80	62	33	39	45	51	57	63	69	75	35	37	38	39	40	42	43	44	25	26	26	28	37	54	48	37
18	22	37	56	68	109	103	77	39	45	51	57	63	69	75	81	37	38	39	40	42	43	44	45	26	26	32	39	44	63	58	43
24	35	55	64	81	104	113	92	45	51	57	63	69	75	81	87	38	39	40	42	43	44	45	47	27	31	39	42	49	58	61	49
49	64	78	87	103	121	120	101	51	57	63	69	75	81	87	93	39	40	42	43	44	45	47	48	38	43	49	52	58	65	63	52
72	92	95	98	112	100	103	99	57	63	69	75	81	87	93	99	40	42	43	44	45	47	48	49	47	56	56	55	60	53	53	49



(a) T_S (b) T_L (c) T_P (d) T_F

Figure 4: T_S, T_L, T_P and T_F when $\alpha = 2, \beta = 0$

Table 1 and Fig. 4 show T_S, T_L, T_P and T_F when $\alpha = 2$ and $\beta = 0$, respectively. As shown in T_F, the low frequency values are increased and the high frequency values are decreased in the quantization table.

4 Experiments for 240×320 handset images

To obtain the quantization table for QVGA images,we selected 300 images from the SK Telecom photo service site [5], which consist of 100 facial images, 100 whole body images and 100 background images. Fig. 5 shows sample pictures among the test images.

Based on the proposed method, quantization tables are selected in three ways, which are optimizing a performance index for α, optimizing that performance index for both α and β, and selecting a quantization table in order to reduce the size of compressed images preserving image quality.

Next, to validate the selected pre-emphasis factors and bias factors, we choose another 300 images, as shown in Fig. 6 and adopt these selected α's and β's to these examples.

4.1 Pre-emphasis factor optimization for 240×320 handset images [4]

We choose the optimum pre-emphasis factor α to minimize the Lagrangian cost given by $J(\alpha) = D(\alpha) + \lambda R(\alpha)$, where $D(\alpha)$ and $R(\alpha)$ are the distortion of the reconstructed images and the rate required

Figure 5: Samples of test images to select quantization tables

Figure 6: Samples of test images to validate the selected quantization tables

to encode the test images with the corresponding quantization table, respectively, and λ is the Lagrangian multiplier. An exhaustive search for the optimum value of λ concludes that the Lagrangian cost tends to be minimized at $\lambda = 1.125$. Here, we set $\beta = 0$.

The cost values are optimal for 1.6, 1.7 and 1.9 for the facial, body and background images, respectively, and we select these values. (Also, as shown in Table 2, if we do not divide the images into these three categories, without any loss of generality, we can use $\alpha = 1.9$.)

4.2 Pre-emphasis factor and bias factor optimization for 240×320 handset images

We choose the optimum pre-emphasis factor α and bias factor β to minimize the Lagrangian cost given by $J(\alpha,\beta) = D(\alpha,\beta) + \lambda R(\alpha,\beta)$ with respect to the same $\lambda = 1.125$ as in Section 4.1. Table 3 shows the selected pre-emphasis factors and bias factors considering the cost function. Comparing to the result in Table 2, there is a little improvement in the cost values.

4.3 Selection of quantization table considering image size preserving image quality

Also, we select α and β in order to minimize the image size preserving image quality. Table 4 shows the selected pre-emphasis factors and bias factors.

Table 2: Cost values $J(\alpha)$ for variable α ($\beta = 0$)

α	Face	Body	Background	total
1.4	20.31	44.50	38.35	103.16
1.5	20.32	44.47	38.07	102.86
1.6	**20.16**	44.19	37.52	101.87
1.7	20.21	**44.16**	37.25	101.62
1.8	20.29	44.24	37.19	101.72
1.9	20.28	44.24	**36.93**	**101.45**
2.0	20.56	44.40	37.19	102.15
2.1	20.72	44.59	37.22	102.53

Table 3: Cost values $J(\alpha, \beta)$ for selected α's and β's

	α	β	Cost value
Face	2.3	-4	20.12
Body	2.1	-1	44.06
Background	2.3	-1	36.77

4.4 Experimental results

Let us denote the selected quantization tables in Section 4.1, Section 4.2 and Section 4.3 as T_{F_1}, T_{F_2} and T_{F_3}, respectively.

Table 5 shows the performance improvement using the proposed method for the images in Fig. 5. For T_{F_2}, there are improvements of 7.58%, 7.93% and 2.85% in the size and 0.18dB, 0.16dB and 0.53 dB in the PSNR, for the face, body and background images, respectively. Also, for T_{F_3}, there are improvements of 10.23%, 11.47% and 9.45% in the size and 0.03dB, 0.01dB and 0.18 dB in the PSNR, respectively.

Next, to validate the selected pre-emphasis factors and bias factors, we apply the selected factors to other 300 images as shown in Fig. 6. Table 6 shows the performance improvement using the proposed method for the images in Fig. 6. For T_{F_2}, there are improvements of 8.51%, 6.16% and 7.14% in the size and 0.22dB, 0.05dB and 0.19 dB in the PSNR, for the face, body and background images, respectively. Also, for T_{F_3}, there are improvements of 11.03%, 9.77% and 13.13% in the size and 0.08dB, 0.2dB and 0.02 dB in the PSNR, respectively.

Table 4: Selected α's and β's

	α	β
Face	2.1	-1
Body	2.3	1
Background	2.1	4

Table 5: Performance improvement using the proposed method compared to the standard quantization table

	Face		Body		Background	
	bpp	dB	bpp	dB	bpp	dB
T_S	1.192	35.32	1.520	31.71	1.576	32.18
T_{F_1}	1.120	35.40	1.424	31.86	1.543	32.69
Improvements	6.04%	0.08dB	6.31%	0.15dB	2.09%	0.51dB
T_{F_2}	1.107	35.50	1.407	31.87	1.532	32.71
Improvements	7.58%	0.18dB	7.93%	0.16dB	2.85%	0.53dB
T_{F_3}	1.07	35.35	1.346	31.72	1.427	32.36
Improvements	10.23%	0.03dB	11.47%	0.01dB	9.45%	0.18dB

Table 6: Performance improvement using the proposed method compared to the standard quantization table

	Face		Body		Background	
	bpp	dB	bpp	dB	bpp	dB
T_S	1.151	35.42	1.402	33.47	1.226	34.61
T_{F_1}	1.069	35.56	1.331	33.83	1.161	34.85
Improvements	7.12%	0.14dB	5.06%	0.36dB	5.30%	0.24dB
T_{F_2}	1.060	35.64	1.320	33.52	1.143	34.8
Improvements	8.51%	0.22dB	6.16%	0.05dB	7.14%	0.19dB
T_{F_3}	1.024	35.5	1.265	33.67	1.065	34.63
Improvements	11.03%	0.08dB	9.77%	0.2dB	13.13%	0.02dB

5 Conclusion

In this paper, we presented a new JPEG quantization table design method for mobile images and the experimental results for the selected images. Based on the characteristics of mobile images, we proposed a new quantization table model. Also, considering the R-D cost function, the pre-emphasis factors and the bias factors were selected for different image groups. The experimental results showed the validity of the proposed method. Since the model is obtained from the standard quantization table in the proposed scheme, only the pre-emphasis factor and the bias factor need to be transmitted. The proposed scheme can be easily applied to the JPEG codec and can be utilized for the display of 240×320 images or other size images in mobile phones.

Acknowledgments

This work was supported in part by the research program 2010 of Kookmin University, Korea and also supported in part by the Ministry of Knowledge Economy (MKE), Korea, under the Information Technology Research Center (ITRC) support program supervised by the Institute for Information Technology Advancement (IITA) under Grant IITA-2009-C1090-0904-0002.

Bibliography

[1] G. K. Wallace, The JPEG Still-Picture Compression Standard, *Communications of the ACM*, Vol. 34, No. 4, pp. 30-44, 1991.

[2] Independent JPEG Group, http://www.ijg.org.

[3] G.-M. Jeong, J.-H. Kang, Y.-S. Mun and D.-H. Jung, JPEG Quantization Table Design for Photos with Face in Wireless Handset, *Lecture Notes in Computer Science*, Vol. 3333, pp.681-688, 2004

[4] G.-M. Jeong, J.-D. Lee and D.-W. Kang, A JPEG Quantization Table for Mobile QVGA Images, *The Journal of The Institute of Webcasting, Internet Television and Telecommunication (in Korean)*, Vol. 8, No. 1, pp.19-24, 2008

[5] SK Telecom, http://www.sktelecom.com

[6] M. Crouse and K. Ramchandran, Joint Thresholding and Quantizer Selection for Transform Image Coding: Entropy-Constrained Analysis and Applications to Baseline JPEG, *IEEE Transactions on Image Processing*, Vol. 6, No. 2, pp.285-297, 1997

Gu-Min Jeong received the B.S. and M.S. degrees from the Dept. of Control and Instrumentation Eng., Seoul National University, Seoul, Korea, in 1995 and 1997, respectively, and Ph.D. degree from School of Electrical Eng. and Computer Science, Seoul National University, Seoul, Korea in 2001. He was a Senior Engineer at NeoMtel, Korea from 2001-2004 and a Manager at SK Telecom, Korea from 2004-2005. Currently, he is an Associate Professor of School of Electrical Engineering, Kookmin University, Seoul, Korea. His research area includes wireless communication service, mobile multimedia, and embedded systems.

Jong-Duck LEE received the B.S. and M.S. degree from the School of Electrical Eng. Kookmin University, Seoul, Korea, in 2007 and 2009, respectively. Currently, he is working for LIG Nex1, Korea. His research area includes embedded systems and mobile multimedia.

Sang-Il Choi received his B.S. degree in the Division of Electronic Engineering from Sogang University in 2005, and received M.S. and Ph. D. degrees from School of Electrical Eng. and Computer Science, Seoul National University, Seoul, Korea, in 2007 and 2010, respectively. Currently, He is a post doctoral researcher in BK21 information technology in Seoul National University, Korea. His research interests include image processing, face recognition, feature extraction and their applications.

Dong-Wook Kang received the B.S., M.S., and Ph.D. degrees from the Dept. of Electronics Eng., Seoul National University, Seoul, Korea, in 1986, 1988, and 1995, respectively. He joined as a faculty member for the Dept. of Electrical Eng., Kookmin University, Seoul, Korea, in 1995 and now is a Professor of School of Electrical Eng., Kookmin University. He is a trustee of the Korean Society of Broadcasting Engineers. His research area includes video coding, multimedia signal processing and digital culture technology.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 4, pp. 540-550

Solving `Vertex Cover` Problem by Means of Tissue P Systems with Cell Separation

C. Lu, X. Zhang

Chun Lu
Key Laboratory of Image Processing and Intelligent Control
Department of Control Science and Engineering
Huazhong University of Science and Technology
Wuhan 430074, Hubei, People's Republic of China
E-mail: luchun.et@gmail.com(corresponding author)

Xingyi Zhang
School of Computer Science and Technology, Anhui University
Hefei 230601, Anhui, People's Republic of China
E-mail: xyzhanghust@gmail.com

Abstract: Tissue P systems is a computing model in the framework of membrane computing inspired from intercellular communication and cooperation between neurons. Many different variants of this model have been proposed. One of the most important models is known as tissue P systems with cell separation. This model has the ability of generating an exponential amount of workspace in linear time, thus it allows us to design cellular solutions to NP-complete problems in polynomial time. In this paper, we present a solution to the `Vertex Cover` problem via a family of such devices. This is the first solution to this problem in the framework of tissue P systems with cell separation.

Keywords: Membrane Computing, Tissue P System, Cell Separation, `Vertex Cover`

1 Introduction

Membrane computing is an emergent branch of natural computing, which is inspired by the structure and the function of living cells, as well as the organization of cells in tissues, organs and other higher order structures. The devices in membrane computing, called *P systems*, provide distributed parallel and non-deterministic computing models. Since Gh. Păun introduced the P system in [10], this area has received important attention from the scientific community, such as computer scientists, biologists, formal linguists and complexity theoreticians.

In the last years, many different models of P systems have been proposed (a comprehensive bibliography can be found in [14]). The most studied variants are the *cell-like* models of P systems, where membranes are hierarchically arranged in a tree-like structure. Various models of cell-like P systems have been successfully used to design solutions to NP-complete problems in polynomial time (see [4]). These solutions are obtained by generating an exponential amount of workspace in polynomial time and using parallelism to check simultaneously all the candidate solutions. In general, cell division, cell creation and cell separation are the three efficient ways to obtain exponential workspace in polynomial time, thus obtaining three corresponding variants of P systems: *cell division*, where the new workspace is generated by membrane division, *cell creation*, where the new membranes are created from objects, and *cell separation*, where the new workspace is generated by membrane separation. It has been proved that all of the three models can efficiently solve NP-complete problems, but technically they are pretty different in the way of designing solutions.

Another interesting class of P systems is known as *tissue P systems*, where membranes are placed in the nodes of a graph. This variant has two biological inspirations (see [6]): intercellular communication

and cooperation between neurons. The common mathematical model of these two mechanisms is a net of processors dealing with symbols and communicating these symbols along channels specified in advance, based on symport/antiport rules [9]. Tissue P systems can also efficiently solve NP-complete problems provided that some ingredients are added into such systems, as in the case of cell-like P systems. The first attempt in this respect is to consider cell division in tissue P systems, yielding *tissue P systems with cell division* [12]. In this model, the two new cells generated by a division rule have exactly the same objects except for at most a pair of different objects. This model was shown to efficiently solve NP-complete: SAT [12], 3-coloring [1], Subset Sum [2], Vertex Cover [3], etc.

Recently, another class of tissue P systems is proposed based on cell separation, that is, *tissue P systems with cell separation*, and a polynomial-time solution to the NP-complete problem SAT is given in [8]. In this model, the contents of the two new cells evolved from a cell by separation rules can be different, thus leading to a significant difference in specific techniques for designing solutions to concrete NP-complete problems. In this paper, we shall explore the possibility of using such a model to solve another NP-complete problem–Vertex Cover. Specifically, a family of tissue P systems with cell separation is constructed, in which each system can solve all instances of Vertex Cover of a fixed size in a polynomial time. Although the Vertex Cover problem has been considered in the framework of other models in membrane computing (for instance, cell-like P systems with active membrane, tissue P systems with cell division, and so on), here the first solution for this problem is presented in the framework of tissue P systems with cell separation.

The paper is organized as follows: in Sections 2 and 3 preliminaries and the definition of tissue-like P systems with cell separation are recalled, respectively. In Section 4, recognizer tissue P systems are briefly described. A polynomial-time solution to Vertex Cover problem is presented in Section 5, including a short overview of the computation and of the necessary resources. Finally, some conclusions and new open research lines are presented.

2 Preliminaries

An *alphabet*, Σ, is a finite and non-empty set of abstract symbols. An ordered sequence of symbols is a *string*. Let Σ be a (finite) alphabet; then Σ^* is the set of all strings over Σ. The number of symbols in a string u is the *length* of the string, and it is denoted by $|u|$. As usual, empty string (with length 0) is denoted by λ. The set of strings of length n built with symbols from the alphabet Σ is denoted by Σ^n and $\Sigma^* = \cup_{n\geq 0}\Sigma^n$.

Let A be a (finite) set, $A = \{a_1, \cdots, a_n\}$. Then a finite multiset m over A is a function $f : A \to \mathbb{N}$. If $m = (A, f)$ is a multiset then its *support* is defined as $supp(m) = \{x \in A \,|\, f(x) > 0\}$. The size of the multiest m is $|m| = \Sigma_{x\in A} f(x)$. A multiset is empty (resp. finite) if its support is the empty set (resp. finite).

A multiset m over A can also be represented by any string x that contains exactly $f_m(a_i)$ symbols a_i for all $1 \leq i \leq n$, e.g., by $a_1^{f(a_1)} a_2^{f(a_2)} \ldots a_k^{f(a_k)}$. Thus, superscripts indicate the multiplicity of each element, and if $f(x) = 0$ for any $x \in A$, then this element is omitted.

We suppose that the reader is already familiar with the basic notions and the terminology of P systems. For details, see [11].

3 Tissue P Systems with Cell Separation

According to the first works on tissue P systems [5, 6] the membrane structure did not change along the computation. A new model based on the cell-like model of *tissue P systems with cell separation* is presented in [7]. The biological inspiration of them is clear: alive tissues are not *static* network of cells, since membrane fission generates new cells in a natural way.

Formally, a *tissue P system with cell separation* of initial degree $q \geq 1$ is a construct

$$\Pi = (\Gamma, O_1, O_2, w_1, \ldots, w_q, \mathcal{E}, \mathcal{R}, i_o),$$

where:

1. Γ is the *alphabet* of *objects*, $\Gamma = O_1 \cup O_2$, $O_1, O_2 \neq \emptyset$, $O_1 \cap O_2 = \emptyset$;
2. $w_1, \ldots, w_q$ are strings over Γ, describing the multisets of objects placed in the cells of the system at the beginning of the computation;
3. $\mathcal{E} \subseteq \Gamma$ is the set of objects present in the environment in arbitrarily copies each;
4. $\mathcal{R}$ is a finite set of rules of the following forms:
   - **(a)** $(i, u/v, j)$, for $i, j \in \{0, 1, 2, \ldots, q\}, i \neq j$, $u, v \in \Gamma^*$;
     Communication rules; $1, 2, \cdots, q$ identify the cells of the system, 0 is used as the label of the environment. This rule $(i, u/v, j)$ can be applied over two cells i and j such that u is contained in cell i and v is contained in cell j. The application of this rule means that the objects of the multisets represented by u and v are interchanged between the two cells;
   - **(b)** $[a]_i \rightarrow [O_1]_i [O_2]_i$, where $i \in \{1, 2, \ldots, q\}$ and $a \in \Gamma$;
     Separation rules; under the influence of object a, the cell with label i is separated into two cells with the same label; at the same time, the object a is consumed; the objects from O_1 are placed in the first cell, those from O_2 are placed in the second cell;
5. $i_o \in \{0, 1, 2, \ldots, q\}$ is the output region.

Rules are used in the non-deterministic maximally parallel manner as customary in membrane computing. In each step, all cells which can evolve must evolve in a maximally parallel way (in each step a multiset of rules which is maximal is applied, no further rule can be added). This way of applying rules has only one restriction: when a cell is separated, the separation rule is the only one which is applied for that cell in that step; the objects inside that cell do not evolve by means of communication rules. The daughter cells will participate to the interaction with other cells or with the environment by means of communication rules in the next step, if they are not separated once again. Their labels precisely identify the rules which can be applied to them.

A sequence of transitions which starts from the initial configuration is called a computation with respect Π. A computation is completed only if it halts and the computations give a result, and result is the multiset of objects present in region i_o in the halting configuration.

4 Recognizer Tissue P Systems with Cell Separation

NP-completeness has been usually studied in the framework of *decision problems*. Let us recall that a decision problem is a pair (I_X, θ_X) where I_X is a language over a finite alphabet (whose elements are called *instances*) and θ_X is a total Boolean function over I_X.

The notions from classical *computational complexity theory* are adapted for membrane computing to study the computing efficiency for solving decision problems. *Recognizer tissue P systems* are introduced in [12] for tissue P systems with the same idea of *recognizer P systems* introduced into cell-like P systems [13].

A recognizer tissue P system with cell separation of degree $q \geq 1$ is a construct

$$\Pi = (\Gamma, O_1, O_2, \Sigma, w_1, \ldots, w_q, \mathcal{E}, \mathcal{R}, i_{in}, i_o)$$

where:

- $(\Gamma, O_1, O_2, w_1, \ldots, w_q, \mathcal{E}, \mathcal{R}, i_o)$ is a tissue P system with cell separation of degree $q \geq 1$ (as defined in the previous section).
- The working alphabet Γ has two distinguished objects yes and no, at least one copy of them present in some initial multisets $w_1, \ldots, w_q$, but not present in $\mathcal{E}$.
- Σ is an (input) alphabet strictly contained in Γ.
- $i_{in} \in \{1, \ldots, q\}$ is the input cell.
- The output region i_o is the environment.
- All computations halt.
- If $\mathcal{C}$ is a computation of Π, in the last step of the computation either the object yes or the object no (but not both) have to be send out to the environment.

The computations of the system Π with input $w \in \Sigma^*$ start from a configuration of the form $(w_1, w_2, \ldots, w_{i_{in}} w, \ldots, w_q; \mathcal{E})$, that is, after adding the multiset w to the contents of the input cell i_{in}. We say that the multiset w is *recognized* by Π if and only if the object yes is sent to the environment, in the last step of the corresponding computation. We say that $\mathcal{C}$ is an accepting computation (respectively, rejecting computation) if the object yes (respectively, no) appears in the environment associated to the corresponding halting configuration of $\mathcal{C}$.

Definition 1. A decision problem $X = (I_X, \theta_X)$ is solvable in polynomial time by a family of recognizer tissue P systems $\Pi = \{\Pi(n) \mid n \in \mathbb{N}\}$ with cell separation, if the following holds:

- The family Π is *polynomially uniform* by Turing machines, that is, there exists a deterministic Turing machine constructing $\Pi(n)$ from $n \in \mathbb{N}$ in polynomial time.
- There exists a polynomial-time coding (cod, s) form I_X to Π such that:
  - for each instance $u \in I_X$, $s(u)$ is a natural number and $cod(u)$ is an input multiset of the system $\Pi(s(u))$;
  - the family Π is *polynomially bounded* with regard to (X, cod, s), that is, there exists a polynomial function p, such that for each $u \in I_X$ every computation of $\Pi(s(u))$ with input $cod(u)$ is halting and, moreover, it performs at most $p(|u|)$ steps;
  - the family Π is *sound* with regard to (X, cod, s), that is, for each $u \in I_X$, if there exists an accepting computation of $\Pi(s(u))$ with input $cod(u)$, then $\theta_X(u) = 1$;
  - the family Π is *complete* with regard to (X, cod, s), that is, for each $u \in I_X$, if $\theta_X(u) = 1$, then every computation of $\Pi(s(u))$ with input $cod(u)$ is an accepting one.

We denote by $\mathbf{PMC}_{TS}$ the set of all decision problems which can be solved by means of recognizer tissue P systems with cell separation in polynomial time.

5 A Solution to the Vertex Cover Problem

The vertex cover of a non-directed graph is a subset of its vertices such that for each edge of the graph at least one of its endpoints belongs to that subset. The size of the vertex cover is the number of vertices in the subset. The Vertex Cover problem considered in this paper is formulated as follows: given a non-directed graph, $G = (V, E)$, and a natural number $k \leq |V|$, determine whether or not G has a vertex cover of size at most k.

We shall prove that `Vertex Cover` can be solved in linear time (in the number of nodes and edges of the graph) by a family of recognizer tissue-like P systems with cell separation. We construct a family $\Pi = \{\Pi(\langle n,m,k\rangle) \mid n,m,k \in \mathbb{N}\}$ where each system of the family will process every instance u of the problem given by a graph with n vertices and m edges, and by a size k of the vertex cover (that is, $s(u) = \langle n,m,k\rangle$, where $\langle a,b\rangle = \frac{(a+b)(a+b+1)}{2} + a$ and $\langle a,b,c\rangle = \langle\langle a,b\rangle,c\rangle$. In order to provide a suitable encoding of these instances, we will use the objects A_{ij}, with $1 \leq i < j \leq n$, to represent the edges of the graph, and we will provide $cod(u) = \{A_{ij} \mid 1 \leq i < j \leq n \wedge (v_i,v_j) \in E\}$ as the initial multiset for the system.

With an instance u of the `VC` problem, the system $\Pi(s(u))$ with input $cod(u)$ decides that instance by a brute force algorithm, implemented in the following four stages:

- *Generation Stage*: The initial cell labeled by 2 is separated into two new cells; the separations are iterated until a cell has been produced for each possible candidate solution.
- *Pre-checking Stage*: After obtaining all possible subsets of vertices encoded in cells labeled by 2, this stage only select the subsets of size k.
- *Checking Stage*: For each of these subsets, it is checked if there exists an edge of the graph for which none of its endpoints is in the subset.
- *Output Stage*: The system sends to the environment the right answer according to the results of the previous stage.

$\Pi(\langle n,m,k\rangle) = (\Gamma(\langle n,m,k\rangle), \Sigma(\langle n,m,k\rangle), w_1, w_2, \mathcal{R}(\langle n,m,k\rangle), \mathcal{E}(\langle n,m,k\rangle), i_{in}, i_0)$, for each $n,m,k \in \mathbb{N}$. The family Π contains the following systems:

- $\Gamma(\langle n,m,k\rangle) = O_1 \cup O_2$,

$$
\begin{aligned}
O_1 \;=\; & \{c_{i,j}, A_{i,j}, z_{i,j}, P_{i,j} \mid 1 \leq i < j \leq n\} \cup \{j_i \mid 1 \leq i \leq 2n+1\} \\
& \cup \{A_i, B_i, B_i', C_i', T_i, F_i' \mid 1 \leq i \leq n\} \cup \{d_i \mid 1 \leq i \leq n+1\} \\
& \cup \{D_{i,j} \mid 1 \leq i,j \leq n\} \cup \{a_{1,i}, b_{1,i}, d_{1,i}, g_i, h_i, l_i, e_i \mid 1 \leq i \leq n-1\} \\
& \cup \{a_i \mid 1 \leq i \leq 5n+m+\lceil \lg n \rceil + 9\} \cup \{a_{2,i}, b_{2,i}, d_{2,i} \mid 2 \leq i \leq n-1\} \\
& \cup \{a_{i,j,k}, b_{i,j,k}, d_{i,j,k} \mid 1 \leq i < j \leq n, 1 \leq k \leq n-1\} \\
& \cup \{C_{i,j}, B_{i,j} \mid 1 \leq i \leq n, 1 \leq j \leq m\} \cup \{L_i \mid 1 \leq i \leq m+\lceil \lg n \rceil + 7\} \\
& \cup \{P_i \mid 1 \leq i \leq m+\lceil \lg n \rceil + 6\} \cup \{H_i \mid 1 \leq i \leq \lceil \lg m \rceil + 1\} \\
& \cup \{G_i \mid 1 \leq j \leq \lceil \lg n \rceil + 1\} \cup \{b, z, f_1, y, s, E_0, E_1, E_2, T, N, yes, no\}, \\
O_2 \;=\; & \{c_{i,j}', A_{i,j}', z_{i,j}'\} \mid 1 \leq i < j \leq n\} \cup \{T_i', F_i \mid 1 \leq i \leq n\} \cup \{y', z', f'\}.
\end{aligned}
$$

- $\Sigma(\langle n,m,k\rangle) = \{c_{i,j}, A_{i,j}, A_{i,j}' \mid 1 \leq i < j \leq n\}$.
- $w_1 = a_1 a_{1,1} g_1 a_{i,j,1}$`yes no`.
- $w_2 = c_{i,j} A_{i,j} A_1$.
- $\mathcal{R}(\langle n,m,k\rangle)$ is the set of rules:

1. **Separation rule:**
   $r_1 \equiv [s]_2 \to [O_1]_2 [O_2]_2$.
2. **Communication rules:**
   $r_{2,i} \equiv (1, a_i / a_{i+1}, 0)$ for $1 \leq i \leq 5n+m+\lceil \lg n \rceil + 8$;
   $r_{3,i,j,k} \equiv (1, a_{i,j,k} / b_{i,j,k}, 0)$ for $1 \leq i < j \leq n, 1 \leq k \leq n-1$;
   $r_{4,i,j,k} \equiv (1, b_{i,j,k} / c_{i,j}^2 d_{i,j,k}^2, 0)$ for $1 \leq i < j \leq n, 1 \leq k \leq n-1$;
   $r_{5,i,j,k} \equiv (1, d_{i,j,k} / a_{i,j,k+1}, 0)$ for $1 \leq i < j \leq n, 1 \leq k \leq n-2$;

$r_{6,i} \equiv (1, g_i/h_i, 0)$ for $1 \leq i \leq n-1$;
$r_{7,i} \equiv (1, h_i/l_i^2 A_{i+1}^2, 0)$ for $1 \leq i \leq n-1$;
$r_{8,i} \equiv (1, l_i/g_{i+1}, 0)$ for $1 \leq i \leq n-2$;
$r_{9,i} \equiv (1, a_{1,i}/b_{1,i}, 0)$ for $1 \leq i \leq n-1$;
$r_{10,i} \equiv (1, b_{1,i}/c^2 d_{1,i}^2 e_i^2, 0)$ for $1 \leq i \leq n-1$;
$r_{11,i} \equiv (1, d_{1,i}/a_{1,i+1}, 0)$ for $1 \leq i \leq n-2$;
$r_{12,i} \equiv (1, e_i/a_{2,i+1}, 0)$ for $1 \leq i \leq n-2$;
$r_{13,i} \equiv (1, a_{2,i}/b_{2,i}, 0)$ for $2 \leq i \leq n-1$;
$r_{14,i} \equiv (1, b_{2,i}/c^2 d_{2,i}^2, 0)$ for $2 \leq i \leq n-1$;
$r_{15,i} \equiv (1, d_{2,i}/a_{2,i+1}, 0)$ for $2 \leq i \leq n-2$;
$r_{16,i,j} \equiv (2, c_{i,j} A_{i,j}/z_{i,j} z'_{i,j} A_{i,j} A'_{i,j}, 0)$ for $1 \leq i < j \leq n$;
$r_{17,i,j} \equiv (2, c_{i,j} A'_{i,j}/z_{i,j} z'_{i,j} A_{i,j} A'_{i,j}, 0)$ for $1 \leq i < j \leq n$;
$r_{18,i} \equiv (2, cT_i/zz'T_iT'_i, 0)$ for $1 \leq i \leq n-1$;
$r_{19,i} \equiv (2, cT'_i/zz'T_iT'_i, 0)$ for $1 \leq i \leq n-1$;
$r_{20,i} \equiv (2, cF_i/zz'F_iF'_i, 0)$ for $1 \leq i \leq n-1$;
$r_{21,i} \equiv (2, cF'_i/zz'F_iF'_i, 0)$ for $1 \leq i \leq n-1$;
$r_{22} \equiv (2, A_n/T_nF_nf_1f'_1s, 0)$;
$r_{23,i} \equiv (2, A_i/T_iF_iyy'zz's, 0)$ for $1 \leq i \leq n-1$;
$r_{24,i} \equiv (2, y/A_i, 1)$ for $2 \leq i \leq n$;
$r_{25,i} \equiv (2, y'/A_i, 1)$ for $2 \leq i \leq n$;
$r_{26} \equiv (2, z/c, 1)$;
$r_{27} \equiv (2, z'/c, 1)$;
$r_{28,i,j} \equiv (2, z_{i,j}/c_{i,j}, 1)$ for $1 \leq i < j \leq n$;
$r_{29,i,j} \equiv (2, z'_{i,j}/c_{i,j}, 1)$ for $1 \leq i < j \leq n$;
$r_{30} \equiv (1, z/\lambda, 0)$;
$r_{31} \equiv (1, z'/\lambda, 0)$;
$r_{32,i,j} \equiv (1, z_{i,j}/\lambda, 0)$ for $1 \leq i < j \leq n$;
$r_{33,i,j} \equiv (1, z'_{i,j}/\lambda, 0)$ for $1 \leq i < j \leq n$;
$r_{34} \equiv (2, f/j_1 d_1, 0)$;
$r_{35} \equiv (2, f'/j_1 d_1, 0)$;
$r_{36,i,j} \equiv (2, d_j T_i/D_{i,j}, 0)$ for $1 \leq i, j \leq n$;
$r_{37,i,j} \equiv (2, d_j T'_i/D_{i,j}, 0)$ for $1 \leq i, j \leq n$;
$r_{38,i,j} \equiv (2, D_{i,j}/B_i d_{j+1}, 0)$ for $1 \leq i, j \leq n$;
$r_{39,i} \equiv (2, j_i/j_{i+1}, 0)$ for $1 \leq i \leq 2n$;
$r_{40} \equiv (2, j_{2n+1} d_{k+1}/E_0, 0)$;
$r_{41} \equiv (2, E_0/L_1E_1, 0)$;
$r_{42,i} \equiv (2, L_i/L_{i+1}, 0)$ for $i = 1, \ldots, m + \lceil \lg n \rceil + 6$;
$r_{43} \equiv (2, E_1/P_1E_2, 0)$;
$r_{44} \equiv (2, E_2/G_1H_1, 0)$;
$r_{45,i} \equiv (2, P_i/P_{i+1}, 0)$ for $i = 1, \ldots, m + \lceil \lg n \rceil + 5$;
$r_{46,i} \equiv (2, G_i/G_{i+1}^2, 0)$ for $i = 1, \ldots, \lceil \lg n \rceil$;
$r_{47,i} \equiv (2, H_i/H_{i+1}^2, 0)$ for $i = 1, \ldots, \lceil \lg m \rceil$;
$r_{48,i,j} \equiv (2, A_{i,j} H_{\lceil \lg m \rceil + 1}/P_{i,j}, 0)$ for $1 \leq i < j \leq n$;
$r_{49,i,j} \equiv (2, A'_{i,j} H_{\lceil \lg m \rceil + 1}/P_{i,j}, 0)$ for $1 \leq i < j \leq n$;
$r_{50,i} \equiv (2, G_{\lceil \lg n \rceil + 1} B_i/C_i, 0)$ for $i = 1, \ldots, n$;
$r_{51,i} \equiv (2, C_i/C_{i,1} B_{i,1}, 0)$ for $i = 1, \ldots, n$;
$r_{52,i,j} \equiv (2, B_{i,j}/B_{i,j+1} B'_i, 0)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$;
$r_{53,i,j} \equiv (2, C_{i,j}/C_{i,j+1} C'_i, 0)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$;
$r_{54,i,j} \equiv (2, B'_i P_{i,j}/\lambda, 0)$ for $1 \leq i < j \leq n$;

$r_{55,i,j} \equiv (2, C'_j P_{i,j}/\lambda, 0)$ for $1 \leq i < j \leq n$;
$r_{56,i,j} \equiv (2, P_{m+\lceil \lg n \rceil + 5} P_{i,j}/N, 0)$ for $1 \leq i < j \leq n$;
$r_{57} \equiv (2, L_{m+\lceil \lg n \rceil + 7} P_{m+\lceil \lg n \rceil + 6}/T, 0)$;
$r_{58} \equiv (1, b/T, 2)$;
$r_{59} \equiv (1, a_{5n+m+\lceil \lg n \rceil + 9} b/N, 2)$;
$r_{60} \equiv (1, T \ \texttt{yes}/\lambda, 0)$;
$r_{61} \equiv (1, N \ \texttt{no}/\lambda, 0)$;

- $\mathcal{E}(\langle n,m,k \rangle) = \Gamma(\langle n,m,k \rangle) - \{yes, no\}$.
- $i_{in} = 2$ is the *input cell*.
- $i_o = 0$ is the *output region*.

We will show that the family $\Pi = \{\Pi(\langle n,m,k \rangle) \mid n,m,k \in \mathbb{N}\}$ defined above is polynomially uniform by Turing machines. To this aim it will be proved that $\Pi(\langle n,m,k \rangle)$ is built in polynomial time with respect to the size parameter n, m and k of instances of `Vertex Cover` problem.

It is easy to check that the rules of a system $\Pi(\langle n,m,k \rangle)$ of the family are defined recursively from the values n, m and k. The necessary resources to build an element of the family are of a polynomial order, as shown below:

- Size of the alphabet: $n^2 + 5mn + 26n + 7m + 4\lceil \lg n \rceil + \lceil \lg m \rceil + 27 \in O(n^2 + mn)$.
- Initial number of cells: $2 \in O(1)$.
- Initial number of objects: $3m + 6 \in O(m)$.
- Number of rules: $5mn + 3n^2 + 26n + 10m + 4\lceil \lg n \rceil + \lceil \lg m \rceil + 6 \in O(n^2 + mn)$.
- Maximal length of a rule: $6 \in O(1)$.

Therefore, a deterministic Turing machine can build $\Pi(\langle n,m,k \rangle)$ in a polynomial time with respect to n, m and k.

5.1 An Overview of the Computation

A family of recognizer tissue P systems with cell separation is constructed in the previous section. In the following, we informally describe how the recognizer tissue P system with cell separation $\Pi(s(\gamma))$ with input $cod(\gamma)$ works. Let us start with the *generation stage*, where all the possible subsets of the vertices of the graph are generated. This stage has several parallel processes, which we describe in several items.

– In the cells with label 2, in the presence of $c_{i,j}$, by the rules $r_{16,i,j}$, $r_{17,i,j}$, the objects $c_{i,j}A_{i,j}$, $c_{i,j}A'_{i,j}$ introduce the objects $z_{i,j}z'_{i,j}A_{i,j}A'_{i,j}$, respectively. In the next step, primed objects and non-primed objects are separated into the new daughter cells with label 2. The objects $z_{i,j}$ and $z'_{i,j}$ in cells with label 2 are exchanged with the objects $c_{i,j}$ in the cell with label 1 by the rules $r_{28,i,j}$ and $r_{29,i,j}$. In this way, the cycle of duplication-separation can be iterated.

– In parallel with the above duplication-separation process, the objects c are used to duplicate the objects T_i, T_i', F_i and F_i' by the rules $r_{18,i} - r_{21,i}$ (in general $T_i(T_i')$ and $F_i(F_i')$ correspond to the values $true$ and $false$ of vertex A_i); the rules r_{26} and r_{27} take care of introducing the object c from the cell with label 1 to cells with label 2.

– In the initial configuration of the system, the cell with label 2 contains an object A_1 (A_i encodes the i-th variable in the propositional formula). The objects T_1, F_1', z, z', y, y' and s are brought in the cell with label 2, in exchange of A_1, by the rule $r_{23,i}$. In the next step they are separated into the new daughter cells with label 2 by separation rule, because $(T_1, F_1') \in O_1$ and $(F_1, T_1') \in O_2$. The object s is used to activate the separation rule r_1, and is consumed during the application of this rule. The objects y and y' are used to introduce A_2 from the cell with label 1, and the process of truth-assignment for variable v_2 can continue. In this way, in $3n-1$ steps, we get 2^n cells with label 2, and each one contains one of the 2^n possible truth-assignments for the n variables.

– In parallel with the operations in the cells with label 2, the objects $a_{i,j,k+1}$ from the cell with label 1 are traded for objects $b_{i,j,k+1}$ from the environment at the step $3k+1$ ($0 \leq k \leq n-3$) by the rule $r_{2,i,j,k}$. In the next step, each object $b_{i,j,k+1}$ is traded for two copies of objects $c_{i,j}$ and $d_{i,j,k+1}$ by the rule $r_{3,i,j,k}$. At step $3k+3$ ($0 \leq k \leq n-3$), the object $d_{i,j,k}$ is traded for object $a_{i,j,k+2}$ by the rule $r_{4,i,j,k}$. Especially, at step $3n-5$, $a_{i,j,n-1}$ is traded for $b_{i,j,n-1}$ by the $r_{2,i,j,k}$, at step $3n-4$, each copy of object $b_{i,j,n-1}$ is traded for two copies of $c_{i,j}$ by the $r_{4,i,j}$. After step $3n-4$, there is no object $a_{i,j,k}$ appears in the cell with label 1, and the group of rules $r_{3,i,j,k} - r_{5,i,j,k}$ will not be used again. Note that the subscript k of the object $a_{i,j,k}$ grows by 1 in every 3 steps until reaching the value $n-1$, and the number of copies of $a_{i,j,k}$ is doubled in every 3 steps. At step $3k+3$ ($0 \leq k \leq n-2$), the cell with label 1 contains 2^{k+1} copies of object $c_{i,j}$. At the same time, we have 2^{k+1} cells with label 2, and each cell with label 2 contains one copy of object $z_{i,j}$ (or $z'_{i,j}$). Due to the maximality of the parallelism of using the rules, each cell with label 2 gets exactly one copy of $c_{i,j}$ from the cell with label 1 by the rules $r_{28,i,j}$ and $r_{29,i,j}$. The object $c_{i,j}$ in cell with label 2 is used for duplication as described above.

– The objects $a_{1,i}$ and $a_{2,i}$ in the cell with label 1 has a similar role as object $a_{i,j,k}$ in cell 1, which introduces appropriate copies of object c for the duplication of objects T_i, T_i', F_i and F_i' by the rules $r_{9,i} - r_{15,i}$. Note that at step $3k+3$ ($0 \leq k \leq n-2$), there are $(k+1)2^{k+1}$ copies of object c which, by the maximality of the parallelism of using the rules, ensures that each cell with label 2 gets $k+1$ copies of object c .

– The object g_{i+1} in the cell with label 1 is traded for h_{i+1} from the environment at step $3i+1$ ($0 \leq i \leq n-3$) by the rule $r_{6,i}$. In the next step, the object h_{i+1} is traded for two copies of objects l_{i+1} and A_{i+2} by the rule $r_{13,i}$. At the step $3i+3$ ($0 \leq i \leq n-3$), the object l_{i+1} is traded for two copies of g_{i+2}, so that the process can be iterated, until the subscript i of g_i reaches $n-1$. At step $3n-5$, object g_{n-1} is traded for h_{n-1} by the rule $r_{6,i}$. At step $3n-4$, each object h_{n-1} is traded for two copies of A_n. After step $3n-4$, no object g_i appears in the cell with label 1, and the group of rules $r_{15,i} - r_{18,i}$ will not be used again. At the step $3i+3$ ($0 \leq i \leq n-2$), the cell with label 1 contains 2^{i+1} copies of A_{i+2}, and we have 2^{i+1} cells with label 2, each of them containing one copy of object y or one copy of object y'. Due to the maximality of the parallelism of using the rules, each cell with label 2 gets exactly one copy of A_{i+2} from cell 1 by the rules $r_{24,i}$ and $r_{25,i}$. In this way, the truth-assignment for the vertex A_{i+1} can continue.

– The objects $z_{i,j}$, $z'_{i,j}$, y, y', z and z' in the cell with label 1 are removed by the rules $r_{28,i,j}$, $r_{29,i,j}$, r_{30}, r_{31}.

Note that this non-deterministic generation stage is performed by the successive application of the separation rules, and at the end of the stage the same configuration is always reached. Thus, the system is confluent in this stage and performs $3n+1$ steps.

Now that all the subsets of vertices of the graph are generated, the *pre-checking stage* selects only those of size k. This stage is activated by rules r_{34} and r_{35}, which interchange the object f (or f') of each 2-cell (recall that there are 2^n of them) from the environment, and then each of the latter in each 2-cell

introduces an object d_1 and an object j_1 from the environment (recall that there are infinitely many of them).

The objects d_1 and j_1 start two processes of counting in each 2-cell. The first process counts the steps of this stage with counter j_i using rules $r_{39,i}$.

The second process counts the number of vertices in the subset. It is performed using rules $r_{36,i,j}$ and $r_{37,i,j}$, which interchange the objects T_i in the 2-cells by objects B_i (indicating this way that the corresponding vertex has been counted) and increase the counter d_j (the only purpose of the objects D_{ij} is to reduce the length of the rules). Note that this is a non-deterministic process, since the vertex "counted" in each step is chosen in a non-deterministic way. However, as the size of the subsets of vertices is bounded by n, after $2n$ steps of this process, the same configuration is always reached, so the system is also confluent in this stage.

For the counter d_j of a 2-cell to increase, it is necessary and sufficient that in that cell there exist objects B_i left. This means that at the end of the process explained in the previous paragraph, the only 2-cells that contain objects encoding subsets of vertices of size k are those containing the object d_{k+1}. At this moment, those cells also contain the counter j_{2n+1}, which then in two steps cause (using rules r_{40} and r_{41}, and the intermediate object E_0 for rules size reduction) the object d_{k+1} to be interchanged by objects L_1 and E_1 from the environment.

The total number of steps of the *pre-checking stage* is $2n+2$.

The *checking stage* starts now, but before checking if any of the subsets of vertices of size k selected in the previous stage is a vertex cover of the graph, we need some preparation steps. First of all, the objects L_i will be used as a counter, controlled by rules $r_{42,i}$, of the number of steps performed. On the other hand, rule r_{43} introduces another counter P_i, controlled by rules $r_{45,i}$, which runs in parallel, but with a delay of one step. Also, in each 2-cell encoding a subset of vertices of size k objects G_1 and H_1 are introduced by rules r_{43} and r_{44}, and are then multiplied by rules $r_{45,i}$ and $r_{46,i}$ until obtaining n copies of the former and m copies of the latter.

The objects $H_{\lceil \lg m \rceil+1}$ are used by rules $r_{48,i,j}$ and $r_{49,i,j}$ to change into objects P_{ij} encoding the edges of the graph. On the other hand, rules $r_{50,i}$, $r_{51,i}$, $r_{52,i,j}$ and $r_{53,i,j}$ produce, from objects $G_{\lceil \lg n \rceil+1}$ and B_i and by successive interchanges of objects between the 2-cells and the environment, m copies of objects B'_i and C'_i for each and all of the vertices in the subset encoded into the 2-cell.

As the copies of objects B'_i and C'_i are being produced, rules $r_{54,i,j}$ and $r_{55,i,j}$ eliminate from the 2-cell, in a non-deterministic way, edges of the graph (encoded by objects P_{ij}) such that at least one of its endpoints is contained in the subset encoded in the corresponding 2-cell. Once this stage has performed $m+\lceil \lg n \rceil+6$ steps, we are sure that if there is any object P_{ij} left in the 2-cell, then the subset of vertices encoded in that cell is not a vertex cover of the graph, and rule $r_{56,i,j}$ eliminates the counter P_i in an additional step.

The *answer stage* starts at step $5n+m+\lceil \lg n \rceil+9$, when the object $l_{m+\lceil \lg n \rceil+7}$ appears in every 2-cell encoding a subset of vertices of size k. If the counter P has survived in any of these 2-cells, it means that it encoded a vertex cover of the graph, and rule r_{57} interchanges the two counters with an object T from the environment, which is then sent to the 1-cell of the system by rule r_{58}. Then, rules r_{59}, r_{60} and r_{61} control if this cell has received at least one object T from any of the 2-cells of the system. If this is the case, it is detected at step $5n+m+\lceil \lg n \rceil+9$, when an object `yes` is sent to the environment and the system halts. Otherwise, it is detected at step $5n+m+\lceil \lg n \rceil+10$, when an object `no` is sent to the environment and the system halts.

5.2 Main Results

From the discussion in the previous section, the family Π is polynomially bounded, sound and complete with regard to (VC,*cod*,*s*). We have the following result:

Theorem 5.1. `Vertex Cover` $\in$ $\mathbf{PMC}_{TS}$.

Corollary 2. $\mathbf{NP} \cup \mathbf{co}-\mathbf{NP} \subseteq \mathbf{PMC}_{TS}$.

Proof: It suffices to make the following observations: the Vertex Cover problem is NP-complete, Vertex Cover$\in \mathbf{PMC}_{TS}$ and this complexity class is closed under polynomial-time reduction and under complement.

6 Discussion

The main purpose of this paper is to provide a polynomial time solution for Vertex Cover problem based on tissue P systems with cell separation. We showed that the membrane separation is an important feature that could hold the power to solving computationally hard problems in polynomial time. Following this direction, it remains as further work to describe classical complexity classes below PSPACE with this framework.

7 Acknowledgements

The authors acknowledge the support of National Natural Science Foundation of China (60674106, 30870826, 60703047, and 60533010), Program for New Century Excellent Talents in University (NCET-05-0612), Ph.D. Programs Foundation of Ministry of Education of China (20060487014), Chenguang Program of Wuhan (200750731262), HUST-SRF (2007Z015A), and Natural Science Foundation of Hubei Province (2008CDB113 and 2008CDB180).

Bibliography

[1] D. Díaz-Pernil, M. A. Gutiérrez-Naranjo, M. J. Pérez-Jiménez, A. Riscos-Núñez. A Linear–time Tissue P System Based Solution for the 3–coloring Problem. *Electronic Notes in Theoretical Computer Science*, Vol. 171, pp. 81–93, 2007.

[2] D. Díaz-Pernil, M. A. Gutiérrez-Naranjo, M. J. Pérez-Jiménez, A. Riscos-Núñez. Solving Subset Sum in Linear Time by Using Tissue P Systems with Cell Division. In: J. Mira, J. R. Alvarez, J. R. Ivarez (Eds.) *2nd International Work-Conference*, IWINAC 2007, *Interplay between natural and artificial computation Lecture Notes in Computer Science*, Vol. 4527, pp. 170–179, 2007.

[3] D. Díaz-Pernil, M. J. Pérez-Jiménez, A. Riscos-Núñez, A. Romero. Computational Efficiency of Cellular Division in Tissue-like Membrane Systems. *Romanian Journal of Information Science and Technology*, Vol. 11(3), pp. 229–241, 2008.

[4] M. A. Gutiérrez-Naranjo, M. J. Pérez-Jiménez, F. J. Romero-Campero. A Linear solution for QSAT with Membrane Creation. *Lecture Notes in Computer Science*, Vol. 3850, pp. 241–252, 2006.

[5] C. Martín Vide, J. Pazos, Gh. Păun, A. Rodríguez-Patón. A New Class of Symbolic Abstract Neural Nets: Tissue P Systems. *Lecture Notes in Computer Science*, Vol. 2387, pp. 290–299, 2002.

[6] C. Martín Vide, J. Pazos, Gh. Păun, A. Rodríguez-Patón. Tissue P Systems. *Theoretical Computer Science*, Vol. 296, pp. 295–326, 2003.

[7] L. Pan, T.-O. Ishdorj. P Systems with Active Membranes and Separation Rules. *Journal of Universal Computer Science*, Vol. 10(5), pp. 630–649, 2004.

[8] L. Pan, M. J. Pérez-Jiménez. Efficiency of Tissue P Systems with Cell Separation. In M. A. Martínez-del-Amor, E. F. Orejuela-Pinedo, Gh. Păun, I. Pérez-Hurtado, A. Riscos-Núñez, *Seventh Brainstorming Week on Membrane Computing*, Sevilla, Report RGNC 02/2009, 169–196, 2009.

[9] A. Păun, Gh. Păun. The Power of Communication: P Systems with Symport/Antiport. *New Generation Computing*, Vol. 20(3), pp. 295–395, 2002.

[10] Gh. Păun. Computing with Membranes. *Journal of Computer and System Sciences*, Vol. 61(1), 108–143, 2000.

[11] Gh. Păun. *Membrane Computing, An Introduction*, Springer–Verlag, Berlin, 2002.

[12] Gh. Păun, M. J. Pérez-Jiménez, A. Riscos-Núñez. Tissue P System with Cell Division. In Gh. Păun, A. Riscos-Núñez, A. Romero-Jiménez, F. Sancho-Caparrini (eds.), *Second Brainstorming Week on Membrane Computing*, Sevilla, Report RGNC 01/2004, 380–386, 2004.

[13] M. J. Pérez-Jiménez, A. Romero-Jiménez and F. Sancho-Caparrini, A Polynomial Complexity Class in P Systems Using Membrane Division, In E. Csuhaj-Varjú, C. Kintala, D. Wotschke and Gy. Vaszyl (eds.), *Proceedings of the 5th Workshop on Descriptional Complexity of Formal Systems, DCFS 2003*, pp. 284–294, 2003.

[14] The P System Web Page: http://ppage.psystems.eu

Chun Lu is a Ph.D candidate in Huazhong University of Science and Technology, Wuhan, China. He received his master degree in Systems Engineering from Huazhong University of Science and Technology in 2008. Currently, his main research interests cover membrane computing, neural computing, automata theory and its application.

Xingyi Zhang was born in China on June 6, 1982. He received his doctor degree at Huazhong University of Science and Technology in 2009. Currently, he works in School of Computer Science and Technology, Anhui University. His main research fields are formal language theory and its applications, unconventional models of computation, especially, membrane computing. He has published several scientific papers in international journals.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 4, pp. 551-557

A Secure and Efficient Off-line Electronic Payment System for Wireless Networks

H. Oros, C. Popescu

Horea Oros, Constantin Popescu
Department of Mathematics and Computer Science, University of Oradea
Str. Universitatii 1, Oradea, Romania
E-mail: {horos,cpopescu}@uoradea.ro

Abstract: An electronic cash system allows the exchange of digital coins with value assured by the bank's signature and with concealed user identity. In an electronic cash system, a user can withdraw coins from the bank and then spends each coin anonymously and unlinkably. In this paper we propose a secure and efficient off-line electronic payment system based on bilinear pairings and group signature schemes. The anonymity of the customer is revocable by a trustee in case of a dispute. Because the amount of communication in the payment protocol is about 480 bits, the proposed off-line electronic payment system can be used in wireless networks with limited bandwidth.
Keywords: Electronic payment system, bilinear pairings, group signatures, membership certificate.

1 Introduction

Chaum suggested the first electronic cash system [5] in 1982. In this system the technique of blind signatures was used to guarantee the privacy of users. Various extended systems have been proposed, which provide functionalities such as anonymity, double spending prevention, unforgeability, untraceability and efficiency [1], [4], [8]. Off-line electronic cash systems were first introduced in [6] and then developed further in [9], [10], [11], [12]. In off-line systems the bank's involvement in the payment transaction between a customer and a merchant was eliminated. Customers withdraw electronic coins from the bank and use them to pay a merchant (a shop). The merchant subsequently deposits the coins back to the bank.

In this paper we propose a secure off-line electronic payment system based on bilinear pairings and group signature schemes. In order to construct our electronic cash system, we use the group signature scheme of D. Yao and R. Tamassia [16] and the blind signature of Schnorr [13]. Due to the low amount of communication in the payment protocol that is about 480 bits, our off-line electronic payment system can be used in wireless networks with limited bandwidth.

The rest of this paper is organized as follows. In the next section we present our off-line electronic cash system. Furthermore, we discuss some aspects of security and efficiency in section 3. Finally, section 4 concludes the work of this paper.

2 The Proposed Off-Line Electronic Payment System

An e-cash system is a set of parties with their interactions, exchanging money and goods. A typical e-cash system has three parties:

- Customer: purchases goods or services from the merchant using the e-cash.
- Merchant: sells goods or services to the customer, and deposits the e-cash to the bank.
- Bank: issues the e-cash and maintains the bank account for customers and merchants.

And there are also three protocols: withdrawal, payment and deposit. A customer withdraws electronic coins from the bank and pays the coins to a merchant. Finally, the merchant deposits the paid coins to the bank.

Our electronic payment system consists of four types of participants: customers, merchants, banks and trusted parties. The customers honestly withdraw money from the bank and pay money to the merchant. The merchants get money from customers and deposit it in the bank. The banks manage customer accounts, issue and redeem money. The bank can legally trace a dishonest customer with the help of the trusted parties. An e-cash system is

anonymous if the bank in collaboration with the merchant cannot trace the coin to the customer. The system is off-line if during payment the merchant does not communicate with the bank.

In our off-line electronic cash system, all customers who open a bank account form a group and a trusted party is the group manager. When a customer wants to withdraw an electronic coin from his account, the bank applies a blind signature protocol [13] to this coin and decreases appropriate amount from the customer's account. Everyone including the merchant can verify the validity of the blind signature. The withdrawals are made by the bank by applying the blind signature of Schnorr [13] to a coin randomly selected by a customer and the payments are made by the customer by applying the group signature scheme of D. Yao and R. Tamassia [16] to the random coin.

2.1 System Parameters

This operation outputs the system parameters and public/private keys of users that will be used in the system.

- The group manager chooses a set of public parameters $Y = (G_1, G_2, e, P, H, H', H'')$, where G_1 and G_2 are groups of a large prime order q, G_1 is a gap group, $e : G_1 \times G_1 \to G_2$ is a bilinear map, P is a generator of G_1 and $H : \{0,1\}^* \to G_1$, $H' : \{0,1\}^* \to Z_q$ and $H'' : \{0,1\}^* \times G_1 \to Z_q$ are three collision-resistant hash functions. The group manager chooses his secret key $s_A \in Z_q^*$ and computes the public key $P_A = s_A P$.
- The customer chooses a secret $s_u \in Z_q^*$ as his private key and computes the product $P_u = s_u P$ as its public key.
- The bank selects a random secret x_b from the interval $[1, q-1]$ and calculates the point $P_b = x_b P$. The public key of the bank is P_b and the corresponding secret key is x_b.

The process for selecting the parameters and generating G_1, G_2, q, e, P is given in [2].

2.2 The Registration Protocol

We assume that communication between the customer and the group manager is secure, i.e., private and authentic.

Any customer who wants to withdraw a coin from the bank has to interact with the group manager and obtains two type of certificates from the group manager. One is long-term group membership certificate, which certifies the customer's public key information. The other is one-time signing permit, which certifies the customer's one-time signing key information. The latter is used for issuing signatures in the payment protocol.

The registration protocol involves the customer and the group manager as follows:

1. A customer obtains a long-term group membership certificate *Cert* from the group manager. The group manager computes $Cert = s_A H(info \| s_u P)$, where s_A is the private key of the group manager, $s_u P$ is the customer's public key and $info$ contains information such as group name and membership expiration date. *Cert* is given to the customer.
2. A customer also obtains one-time signing permits from the group manager. The customer randomly chooses a number of secrets $x_1, ..., x_l$ and computes one-time signing secret keys $x_1 P, ..., x_l P$ and one-time signing public keys $s_u x_1 P, ..., s_u x_l P$. The keys $s_u P$ and $s_u x_i P$ are sent to the group manager, for all $i = 1, ..., l$. The customer also sends *Cert* to the group manager.
3. The group manager first checks if the customer with public key $s_u P$ is a valid group member. This is done by verifying the following equality:

$$e(Cert, P) = e(P_A, H(info \| s_u P))$$

where P_A is the group manager's public key and $s_u P$ is the customer's public key. The protocol terminates if $e(Cert, P) \neq e(P_A, H(info \| s_u P))$. Then the group manager tests if $e(s_u x_i P, P) = e(s_u P, x_i P)$ for all $i = 1, ..., l$. If the test fails, the protocol terminates. Otherwise, the group manager computes:

$$S_i = s_A H(info \| s_u x_i P)$$

for all $i = 1, ..., l$. S_i is an one-time signing permit and is given to the customer. The group manager adds the tuple $(s_u P, x_i P, s_u x_i P)$ to its record for all $i = 1, ..., l$.

2.3 The Withdrawal Protocol

We assume that communication between the customer and the bank is secure, i.e., private and authentic. The withdrawal protocol allows a customer to withdraw e-coins from the bank. After having open a bank account, the customer withdraws an e-coin from his account by using the blind signature. Therefore, the bank cannot link the e-coin to the identity of the customer but can debit to the account correctly. The withdrawal protocol involves the customer and the bank in which the customer withdraws an electronic coin from the bank. First, the customer proves his identity to the bank using the elliptic curve version of the signature scheme of Shao [14]. Then, the bank uses the elliptic curve version of the blind Schnorr signature scheme [13] to sign the e-coin.

The customer must perform the following protocol with the bank:

1. The customer sets his electronic cash requirement:

$$m = H'(\textit{withdrawal require}||ID)$$

where ID is the identity of the customer. Then, the customer chooses a random value $k_u \in [1, q-1]$ and signs the message m using the elliptic curve version of the signature scheme of Shao [14]:

$$f = H'(m) \quad (1)$$
$$R = k_u f P \quad (2)$$
$$h = H''(m,R) \quad (3)$$
$$s = k_u - h s_u. \quad (4)$$

The customer sends m and its signature (h,s) to the bank.

2. The bank verifies the signature (h,s) of the message m:
   (a) The bank first computes $f = H'(m)$, $R' = f(hP_u + sP)$ and $h' = H''(m,R')$.
   (b) Then, the bank checks that the following equality holds:

$$h = h'.$$

   (c) If $h \neq h'$ the protocol terminates.
3. Then, the bank uses the elliptic curve version of the blind Schnorr signature [13] to sign the e-coin: selects $k' \in [1, q-1]$, computes the point $R'' = k'P$ and sends R'' to the customer.
4. The customer establishes a random coin c, randomly selects $\alpha, \beta \in [1, q-1]$, computes $R_b = R'' + \alpha P + \beta P_b, c_b = H''(c||\alpha, R_b)$ and blinds the e-coin by computing $c' = c_b - \beta \bmod q$. The customer sends the value c' to the bank.
5. The bank computes: $s' = k' - c' x_b \bmod q$ and forwards s' to the customer.
6. The customer computes $s_b = s' + \alpha \bmod q$. The pair (c_b, s_b) is a valid e-coin signature issued by the bank.
7. The customer verifies the blind signature (c_b, s_b) of the coin c, issued by the bank, by checking that the following equation holds:

$$s_b P + c_b P_b = R_b \quad (5)$$

8. The blind signature of the coin c is the pair (c_b, s_b).

The customer gets the coin c from his account.

2.4 The Payment Protocol

The payment protocol involves the customer and the merchant and should be done through a secure channel (i.e., data privacy and integrity). In the proposed system, during payment the merchant does not communicate with the bank. After withdrawing e-coins, the customer can pay for what the merchant provided. Then the merchant verifies the validity of the received e-coins.

In order to sign the coin c, the customer uses the protocol of Yao and Tamassia [16]. The merchant first sends a challenge c_m to the customer. Then, the customer produces a signature S_u of the coin c and merges the signature S_u with his one-time signing permit S_i associated with the secret $s_u x_i$. The details are as follows:

1. The merchant sends challenge $c_m = H'(ID_m||T)$ to the customer, where ID_m is the merchant's identity and T is the recorded time of the transaction.
2. The customer computes:
$$c_u = H'(c||c_m||c_b||s_b) \tag{6}$$
3. The customer computes $S_u = s_u x_i H(c_u)$.
4. The customer computes the signature $S = S_u + S_i$, where $S_i = s_A H(info||s_u x_i P)$.
5. The customer sends c, c_u and the signature $S = S_u + S_i$ of the coin c to the merchant.
6. The merchant verifies the signature S of the coin c as follows:
   (a) Computes the hash digest $H(c_u)$ and the hash digest $h' = H(info||s_u x_i P)$ of one-time signing permit.
   (b) The signature S is accepted if
$$e(S,P) = e(P_A, h')e(s_u x_i P, H(c_u)). \tag{7}$$
   If the test fails, the protocol terminates.

2.5 The Deposit Protocol

The deposit protocol permits the merchant to deposit the received e-coins to the bank. When receiving the deposited requirement from the merchant, the bank first verifies the validity of received e-coins and then credits the account of the merchant.

In on-line e-cash systems this protocol is part of the payment protocol as executed by the merchant. In our e-cash system, the deposit protocol is executed at a later moment, preferably in batch mode. The bank holds a record of spent cash to prevent double spending of e-cash. The bank cannot link deposited coins to a customer without collaboration from the group manager.

The deposit protocol involves the merchant and the bank as follows:

1. The merchant sends c, c_m, c_u, c_b, s_b to the bank.
2. The bank verifies the signature as given in the equation (6).
3. After verification succeeds, the bank checks if c obtained from the merchant exists in its database. If the coin c is in the database of the bank, then the bank finds the signature S' for the deposited coin in its database and sends it to the merchant (detection of double spending).
4. If the merchant receives S' from the bank, he/she checks whether $S' = S$. If $S' = S$, then the merchant rejects performing protocol (double spending). Otherwise, the merchant sends c_u and T to the bank.
5. The bank verifies the validity of the signature S using the equation (7).
6. If the signature S of the coin c is valid, then the bank accepts the coin c. Then, the bank will deposit the cash to the merchant's account and the merchant sends the goods to the customer. The bank stores c and $(c_u, s_u x_u P)$ in its database.
7. If the bank finds out that c and $(c_u, s_u x_u P)$ has been stored before but different T and c_m, then the coin c has been double spending. The bank performs the tracing protocol and detects the identity of the double spender with the help of the group manager.

2.6 The Tracing Protocol

The bank can legally trace the customer of a paid coin with the help of the group manager. The tracing protocol involves the bank and the group manager. Given a signature S and its associated public information P_A and $s_u x_i P$, the group manager verifies the signature S. If the signature S is valid, the group manager can identify a customer's public key $s_u P$ from $s_u x_i P$ value, by consulting the customer group record. The details are as follows:

1. The bank sends c_u and the signature S of the coin c to the group manager.
2. The group manager verifies the signature S using the equation (7).

3. The group manager can easily identify the customer from $s_u x_i P$. The group manager can provide a proof that it is indeed the customer's signature from the following equations:

$$e(s_u x_i P, P) = e(s_u P, x_i P) \tag{8}$$

4. The group manager searches through the group customer list to get the identity of the customer and sends it to the bank.

Similar to what is shown in the group signature scheme of Chen et al. [7], the group manager cannot misattribute a signature to frame the customer unless he can compute bP given q, P, aP and dP which satisfies:

$$a \equiv db \bmod q \tag{9}$$

The authors in [7] define this problem the Reversion of Computation Diffie-Hellman Problem. They prove that the Reversion of Computation Diffie-Hellman Problem is equivalent to Computational Diffie-Hellman Problem in G_1.

3 Security and Efficiency Analysis

In this section we discuss some aspects of security and efficiency of our off-line electronic payment system. We prove that our off-line electronic payment system is secure against tracing a honest customer by the bank and the proposed system is secure against forgery of the coin.

Theorem 1. *Our off-line electronic payment system is secure against existential forgery of the coin c.*

Proof: Long-term membership certificates, one-time signing permits and customer's signatures using one-time secret signing keys are generated by the sign protocol of the signature scheme of Boneh, Gentry, Lynn and Shacham [3]. The authors in [3] shown that their signature scheme is secure against existential forgery attacks. Therefore, if an adversary can forge any of these signatures, she can also forge signatures in the signature scheme of Boneh et al. [3]. Note that a signature computed with one-time secret signing key is in the form of $s_u x_i H(c_u)$, rather than $s_u H(c_u)$ as in the signature scheme [3]. It can be easily shown that if an adversary can forge a signature in a form of $s_u x_i H(c_u)$, then she can forge a signature in the form of $s_u H(c_u)$. Also, since the blind signature of Schnorr is secure against existential forgery, this allows only the legal bank to generate the signature for coin. As the hash function H' has the feature of collision free, the customer cannot find a value $c' \neq c$ with $H'(c'||c_m) = H'(c||c_m)$. Thus, our payment system satisfies unforgeability of the coin. □

Theorem 2. *The both valid signatures S and (c_b, s_b) in our payment system contain a proof of the group membership without revealing the identity of the customer.*

Proof: A valid signature S is obtained from an one-time signing permit of a customer and the customer's signature using the corresponding one-time signing key. That is $S = S_i + S_u$, where $S_i = s_A H(info||s_u x_i P)$ and $S_u = s_u x_i H(c_u)$. Because of the definition of signatures [3], a valid signature S implies that S_i is valid. This proves that the holder of key $s_u x_i P$ is a certified customer. A valid S also means that S_u is valid, therefore S_u is generated with the secret key $s_u x_i$. Thus, S contains a proof of the customer membership. Because the signing key $s_u x_i$ is one-time signing key and x_i is chosen randomly by the customer, the identity of the customer is not revealed. Also, since $c_u = H'(c||c_m||c_b||s_b)$ and the blind signature (c_b, s_b) of the coin c can not give any information for the coin c, the bank can not link the blind coin with the identity of the customer. □

Table 1: Storage space of the payment systems

	Our system	Wang	Lee	Au	Canard
Withdrawal	1120 bits	1824 bits	800 bits	8160 bits	6420 bits
Payment	480 bits	1282 bits	1304 bits	5188 bits	30740 bits
Deposit	960 bits	3232 bits	1656 bits	5164 bits	27648 bits

Table 2: Computation cost of the payment systems

	Our system	Wang	Lee	Au	Canard
Withdrawal Protocol					
multi-EXP	8	9	15	2156	5
Pairing	0	0	0	22	0
Payment Protocol					
multi-EXP	2	11	9	34	1673
Pairing	3	0	0	14	0
Deposit Protocol					
multi-EXP	0	5	7	10	14
Pairing	3	0	0	0	0

Next, we evaluate the storage space and computational time of the costly operations. Table 1 and Table 2 summarize the storage space and computation cost respectively, of different protocols of our e-cash system and the schemes in [1], [4], [9] and [15]. The overall efficiency is improved in our electronic cash system compared to Au et al.'s system [1], Canard et al.'s system [4], Lee et al.'s system [9] and Wang et al.'s e-cash system [15] in terms of the storage space and the computation cost. Our system has a point P of 160 bits and q of 160 bits. The off-line e-cash system proposed by Lee et al. has a point P of 160 bits and 160 bits prime q and the system of Wang et al. has 160 bits prime q and 321 bits prime p. Spending a coin in [15] requires 11 multi-based exponentiations and a total bandwidth of 1282 bits. The payment protocol in [9] requires 9 multi-based exponentiations and a total bandwidth of 1304 bits. For a moderate value $L = 10$ and $t = 40$, the payment protocol in [4] requires 1673 multi-based exponentiations and a total bandwidth of 30740 bits. The payment protocol in [1] requires 34 multi-based exponentiations, 14 pairings and a total bandwidth of 5188 bits. In contrast, the payment protocol in our e-cash system requires 2 multi-based exponentiation, 3 pairings and a total bandwidth of 480 bits.

4 Conclusions

In this paper we presented a secure and efficient off-line electronic payment system based on bilinear pairing and group signature schemes. We used the group signature scheme of Yao and Tamassia and the blind signature of Schnorr. Because the amount of communication between customer and merchant is about 480 bits, the proposed off-line payment system can be used in the wireless networks with the limited bandwidth.

Bibliography

[1] M. Au, W. Susilo, Y. Mu, Practical anonymous divisible e-cash from bounded accumulators, *Proceedings of Financial Cryptography and Data Security*, Lecture Notes in Computer Science 5143 Springer-Verlag, pp. 287-301, 2008.

[2] D. Boneh and M. Franklin, Identity-based encryption from the Weil pairings. *Advances in Cryptology-Crypto 2001*, Lecture Notes in Computer Science 2139, Springer-Verlag, pp.213-229, 2001.

[3] D. Boneh, C. Gentry, B. Lynn, and H. Shacham, Aggregate and verifiably encrypted signatures from bilinear maps. *In Advances in Cryptology - Eurocrypt'03*, Lecture Notes in Computer Science 2656, Springer-Verlag, pp. 416-432, 2003.

[4] S. Canard, A. Gouget, Divisible e-cash systems can be truly anonymous, *Proceedings of EUROCRYPT 2007*, Lecture Notes in Computer Science 4515, Springer-Verlag, pp. 482-497, 2007.

[5] D. Chaum, Blind signature for untraceable payments. *Proceddings of Eurocrypt'82*, Plenum Press. pp.199-203, 1983.

[6] D. Chaum, A. Fiat, M. Naor, Untraceable electronic cash, *Proceedings of the Crypto'88*, pp. 319-327, 1990.

[7] X. Chen,F. Zhang, K. Kim, A New ID-based Group Signature Scheme from Bilinear Pairings. *Journal of Electronics*, 23, pp. 892-900, 2006.

[8] C. Fun, Ownership-attached unblinding of blind signatures for untraceable electronic cash, *Information Science*, 176(3), pp. 263-284, 2006.

[9] M. Lee, G. Ahn, J. Kim, J. Park, B. Lee, K. Kim, H. Lee, Design and implementation of an efficient fair off-line e-cash system based on elliptic curve discrete logarithm problem, *Journal of Communications and Networks*, 4(2), pp. 81-89, 2002.

[10] T. Okamoto, K. Ohta, Universal electronic cash, *Proceedings of the 11th Annual International Cryptology Conference on Advances in Cryptology*, pp. 324-337, 1992.

[11] T. Okamoto, An efficient divisible electronic cash scheme, *Proceedings of Crypto'95*, Lecture Notes in Computer Science 963, Springer-Verlag, pp. 438-451, 1995.

[12] C. Popescu, An Electronic Cash System Based on Group Blind Signatures. *Informatica*, 17(4), pp. 551-564, 2006.

[13] C.P. Schnorr, Efficient signature generation for smart cards, *Journal of Cryptology*, 4(1991), pp. 239-252, 1991.

[14] Zuhua Shao, A provably secure short signature scheme based on discrete logarithms, *Information Sciences: an International Journal*, vol.177(23), pp. 5432-5440, 2007.

[15] H. Wang, J. Cao, Y. Zhang, A flexible payment scheme and its role-based access control. *IEEE Transactions Knowledge Data Engeneering*, 17, pp. 425-436, 2005.

[16] D. Yao, R. Tamassia, Cascaded Authorization with Anonymous-Signer Aggregate Signatures. *Proceedings of the Seventh Annual IEEE Systems, Man and Cybernetics Information Asssurance Workshop*, USA, pp.84-91, 2006.

Horea Oros (b. August 22, 1977) received his PhD in Computer Science (2009) from "Babeş Bolyai" University of Cluj-Napoca, Romania. Since 2001 he is working within the Department of Mathematics and Computer Science, Faculty of Sciences, University of Oradea, Romania, where currently he is a lecturer. He also is lecturer at Agora University of Oradea. He co-authored three books in the filed of computer science and published 19 articles in several journals and proceedings of prestigious international conferences. His main research interest is in the field of cryptology and computer security.

Constantin Popescu (b. October 21, 1967) received his PhD in Computer Science (2001) from "Babeş Bolyai" University of Cluj-Napoca, Romania. Since 2005 he is a professor at the Department of Mathematics and Computer Science, University of Oradea, Romania. His research interests include cryptography, network security, group signatures, security protocols and electronic payment systems. He co-authored 7 books in the field of computer science and published 49 articles in several journals and proceedings of prestigious international conferences. He is reviewer for 10 journals and several prestigious international conferences.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 4, pp. 558-566

Some Aspects about Vagueness & Imprecision in Computer Network Fault-Tree Analysis

D. E. Popescu, M. Lonea, D. Zmaranda, C. Vancea, C. Tiurbe

Daniela Elena Popescu, Doina Zmaranda, Codruta Vancea, Cristian Tiurbe
University of Oradea
Romania, 410087 Oradea, 1 Universitatii St.
E-mail: {depopescu,zdoina,cvancea,ctiurbe}@uoradea.ro

Madalina Lonea
"Politehnica" University of Timisoara
Romania, Timisoara, 2-4 V. Parvan Blvd.
E-mail: madalina_lonea@yahoo.com

Abstract: Based on the available information (eg.multiple functional faults or sensor errors give rise to similar alarm patterns or outcomes), some states in the behaviour of a network can not be distinguished from one another. So, the computer network's fault tree reliability analysis frequently relies on imprecise or vague input data. The paper will use a Dempster-Shafer Theory to accommodate this vagueness and it will show how imprecision can give rise to false-negative, and false-positive inferences; there will be assigned upper and lower bounds for the probability on elements of the state space. After illustrating the computational simplicity of incorporating the Dempster-Shafer Theory probability assignments, we will apply them for analyzing the reliability of the network of our department.
Keywords: reliability analysis, networks, Dempster-Shafer Theory, fault tree.

1 Introduction

The probabilities are no longer appropriate to represent vagueness in risk and reliability analyses; fuzzy set theory was proposed instead to quantify vagueness in this area [1]. Development in DST have shown how probability can be adapted to incomplete or vague information, especially information that is based on human judgment or human-machine interaction.

False positives and false-negatives are often the end products of vagueness or imprecision. They can arise from imprecision due to noise in monitored data, sensor device failure, or from the ambiguity about the logic rules in the fault tree.

So, when the researcher has only imprecise information, he must appeal to fuzzy-set theory techniques, either the common laws or logic his appreciation for the logic relations in fault trees. Fuzzy data can be incorporated through Dempster Shafer Theory (DST) [2] [3] [4] [5] in conventional fault-tree analysis yield meaningful results.

2 Dempster-Shafer Theory mass assignments

DST generalizes classical probability theory by assigning upper und lower bounds for probabilities, as opposed to point values, to both the elements and the subsets of the state space. For a given state space, Ω, mass (probability) is assigned over the set of all possible subsets of Ω. Because each element of Ω is also a subset of Ω (comprising 1 element) any classical probability assignment can be represented in DST. Just as the probabilities of a distribution sum to 1, so do the masses of a DST-distribution.

The standard risk analysis possibility set considered is Ω={*True, False*}. Therefore, the power set, 2^{Ω} = {Φ, T, F, (T, F)} contains an element (T,F) that represents an observation that could be either true or false, but not both.

The components of faults trees - the initiating events, consequences, and rules - are typically given fail or not fail possibilities. In some engineering applications, observation of particular initiating events is imprecise, so an analyst cannot tell whether a given event occurred. By using DST method of assigning probability to the (T,F) event, fault tree analysts can quantitatively depict the imprecision [7].

Based on the fact that [sum of all masses] = 1 and based on the fact that in DST m(Φ) = 0, after DST renormalization of the Φ, the state space become: T, F, (T, F), and:

m (T) + m (F) + m (T,F) = 1

Ordinary fault trees are systems of Boolean equations with components joined by Boolean AND & OR gates. The Boolean AND gate for the members of the effective state space is summarized in Figure 1 [7], and the Boolean OR gate is given in Figure 2.

Figure 1: Boolean Truth Table for the AND gate

$\wedge$	T	F	(T,F)
T	T	F	(T,F)
F	F	F	F
(T,F)	(T,F)	F	(T,F)

Figure 2: Boolean Truth Table for the OR gate

$\vee$	T	F	(T,F)
T	T	T	T
F	T	F	(T,F)
(T,F)	T	(T,F)	(T,F)

The mass for each entry in the table is obtained by multiply-ing the masses of the edge entries and adding all masses of like entries. For example:

Let a_1, a_2, a_3 numbers denoting *ma* over the possibilities $T, F, (T, F)$ where a1+a2+a3=1

Let b_1, b_2, b_3 numbers denoting *mb* over the possibilities $T, F, (T, F)$ where b1+b2+b3=1

For the Boolean OR gate from table 2 we have:

$$\begin{aligned} m\{A \vee B\} &= (a_1b_1 + a_1b_2 + a_1b_3 + a_2b_1 + a_3b_1; a_2b_2; a_2b_3 + a_3b_2 + a_3b_3) \\ &= (a_1 + a_2b_1 + a_3b_1; a_2b_2; a_2b_3 + a_3b_2 + a_3b_3) \end{aligned} \quad (1)$$

Thus, for $(A \vee B)$ the T element has mass assignment $(a_1 + a_2b_1 + a_3b_1)$, the F element has mass assignment a2b2, the (T,F) element has mass assignment $(a_2b_3 + a_3b_2 + a_3b_3)$.

Similarly, for the Boolean gate from Figure 1, we have:

$$\begin{aligned} mA \wedge B &= (a_1b_1; a_1b_2 + a_2b_1 + a_2b_2 + a_2b_3 + a_3b_2; a_1b_3 + a_3b_1 + a_3b_3) \\ &= (a_1b_1; a_1b_2 + a_2 + a_3b_2; a_1b_3 + a_3b_1 + a_3b_3) \end{aligned} \quad (2)$$

3 Fault Tree illustration

We will apply the above techniques for the fault tree that corresponds to the network monitoring system of our department. Our network uses 3 routers and it accesses the Internet through the university

router (Figure 3). It comprises the following elements:

R1	"HN" Router	R1's fault represent the E1 event
R2	"Cazarma" Router	R2's fault represent the E2 event
R3	"CNLAB" Router	R3's fault represent the E3 event
R4	"Univ.Oradea" Router	R4's fault represent the E4 event
S1	Files Server	S1's fault represent the E5 event
S2	FTP Server	S2's fault represent theE6 event
S3	Database Server	S3's fault represent the E7 event
S4	Domain Server	S4's fault represent the E8 event

The network's fault tree representation is shown in Figure 4.

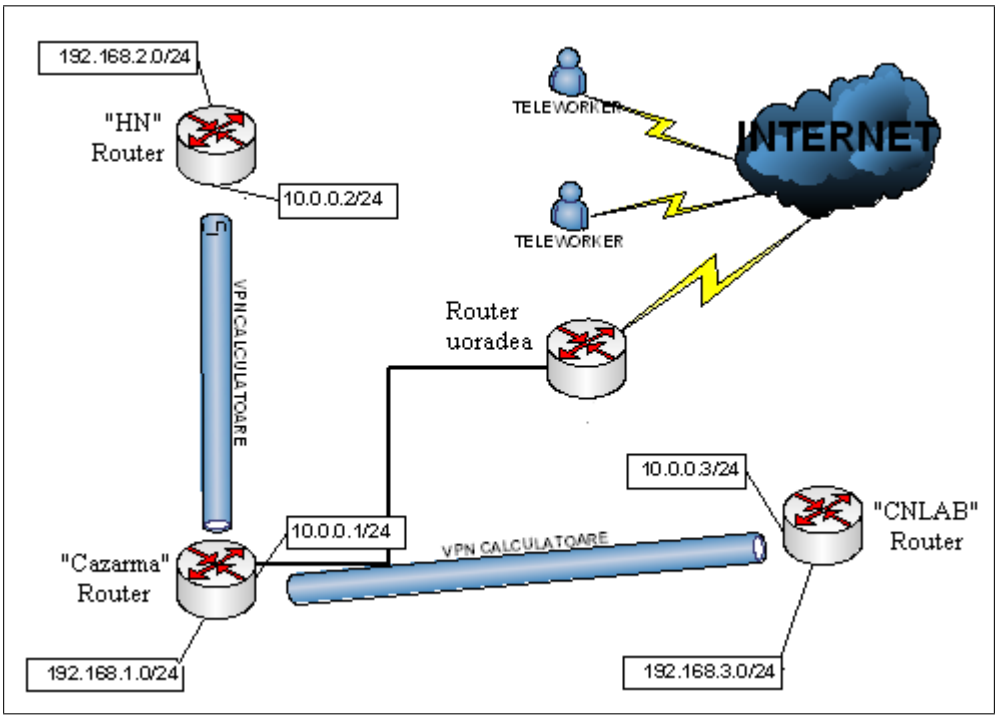


Figure 3: Computer Science Network

The top event is labelled T on the tree represented in Figure 4. The initiating events Ei (i=1, 2, 3, 4, 5, 6, 7, 8) lead to intermediate results (levels), Ij (j=1, 2, 3, 4, 5, 6). The fault tree comprises the following Boolean equations:

$$T = I_1 \vee I6 \tag{3}$$

$$I_1 = E_3 \vee E_4 \tag{4}$$

$$I6 = I_4 \vee I_5 \tag{5}$$

$$I_4 = I_2 \vee I_3 \tag{6}$$

$$I_2 = E_5 \vee E6 \tag{7}$$

$$I_3 = E_7 \vee E8 \tag{8}$$

$$I_5 = E_1 \vee E_2 \tag{9}$$

The equal signs in equations (3) $\div$ (9) mean that the logical implications works in both directions. To assess the risk of the top event, masses can be assigned to the initiating events Ei, and then propagated

up the tree based on relations (1) and (2). The main problem is the determining of the probabilities of basic initiating events (how are they assigned or determined).

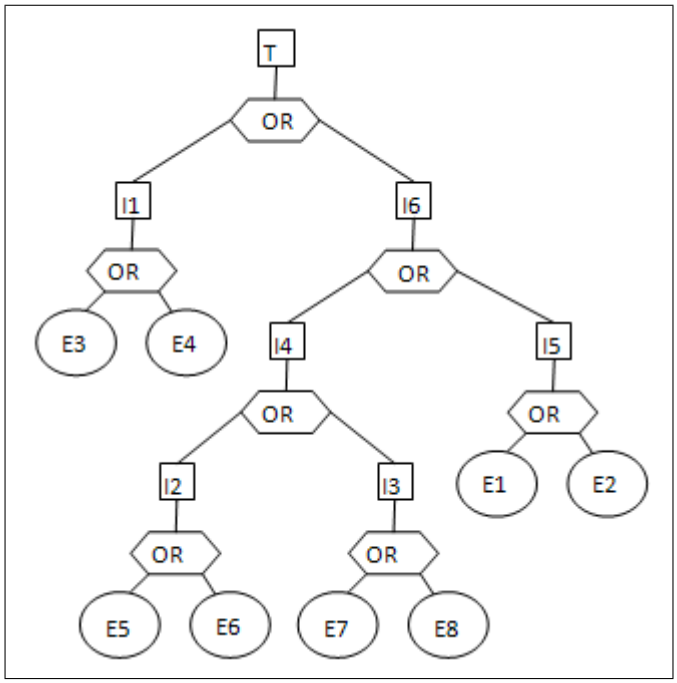


Figure 4: The Fault Tree

Each of (3) ÷ (9) can be interpreted as an "if ... then ..." rule of the fault tree and considered as much part of the fault tree as the initiating events. So, let Ri be a parameter about the failure rates of rule i = 3, ..., 9, where the i-th corresponds to equations i [6]. When joined to our existing fault tree, the Ri depict a further constraint on the fault tree, through which the initiating events must pass before an intermediate level is reached. The Ri can be interpreted in several ways. For example, to model "silent alarms" when the true state of a system is abnormal, the Ri could represent incidents in which the alarms are not working properly. If a rule were incorporated in the design of the fault tree such that the rule were true in only 9 out of 10 trials, then coupling Ri with a Boolean AND gate to the existing tree could filter the number of sure fire observations from the rule to 9/10..

Joining the Ri to the fault tree through an AND gate is useful in modelling false negatives. When the Ri are joined to the fault tree by AND gates and some sensors fail during abnormal conditions, the anticipated consequence might not occur on the fault tree because it has been falsely stopped by the rule parameter.

We will focus only on false-negatives, and thus, the Ri are joined to the existing fault tree through some Boolean AND gates. The new fault tree graphically represents the seven equations:

$$T = (I_1 \vee I6) \wedge R_3 \tag{10}$$

$$I_1 = (E_3 \vee E_4) \wedge R_4 \tag{11}$$

$$I6 = (I_4 \vee I_5) \wedge R_5 \tag{12}$$

$$I_4 = (I_2 \vee I_3) \wedge R6 \tag{13}$$

$$I_2 = (E_5 \vee E6) \wedge R_7 \tag{14}$$

$$I_3 = (E_7 \vee E8) \wedge R8 \tag{15}$$

$$I_5 = (E_1 \vee E_2) \wedge R_9 \tag{16}$$

We consider 3 adjustable assignments of mass for the initiating events and rules variables on the fault tree; these are given in Table 1.

Table 1: Mass Assignment

cwev	Case I	Case II	Case III
E1	(0.9, 0.1, 0)	(0.8, 0, 0.2)	(0.8, 0, 0.2)
E2	(0.9, 0.1, 0)	(0.8, 0, 0.2)	(0.8, 0, 0.2)
E3	(0.9, 0.1, 0)	(0.8, 0, 0.2)	(0.8, 0, 0.2)
E4	(0.9, 0.1, 0)	(0.8, 0, 0.2)	(0.8, 0, 0.2)
E5	(0.8, 0.2, 0)	(0.9, 0, 0.1)	(0.9, 0, 0.1)
E6	(0.8, 0.2, 0)	(0.7, 0, 0.3)	(0.9, 0, 0.1)
E7	(0.8, 0.2, 0)	(0.9, 0, 0.1)	(0.9, 0, 0.1)
E8	(0.8, 0.2, 0)	(0.9, 0, 0.1)	(0.9, 0, 0.1)
R9	(0.8, 0, 0.2)	(0.8, 0, 0.2)	(0.9, 0.1, 0)
R8	(0.7, 0, 0.3)	(0.9, 0, 0.1)	(0.8, 0.2, 0)
R7	(0.9, 0, 0.1)	(0.9, 0, 0.1)	(0.8, 0.2, 0)
R6	(0.9, 0, 0.1)	(0.9, 0, 0.1)	(0.8, 0.2, 0)
R5	(0.9, 0, 0.1)	(0.7, 0, 0.3)	(0.7, 0.3, 0)
R4	(0.8, 0, 0.2)	(0.8, 0, 0.2)	(0.8, 0.2, 0)

1. The initiating events have the usual probability p assigned to true and the remaining (1-p) is assigned to false. The rule variables have the reliability factor assigned to true and the remaining non-true mass assigned to $\veebar$ (T,F). The rule variables correspond to information that each of the rule has held, eg, in 8 out of 10 trials. In the 2 remaining trials, there are not sufficient informations to validate or invalidate the rule.

2. The reliability factor is assigned to true and the remaining non-true mass on both the rule parameters and the initiating events is assigned to the $\veebar$ (T,F) term. There are no sufficient information to conclude whether the event or rule failed. The remaining mass is assigned to $\veebar$ (T,F) term - it is better than assigning the remaining mass to True or False.

3. The evidence is precise on the rule validations but imprecise on the initiating events. There is the situation of a good appreciation of the logic structure of the engineering system but there are problems with the devices. The mass assignments can be obtained from relations (1) and (2) together with the values from Figure 1 For our case study we made our calculation in Microsoft Excel. Our first calculation was made based on relations (3) - (9) and the results are given in Figure 5, and then we made the calculus based on relations (10) - (16) and the results are given in Figure 6.

By analyzing the results from Figure 5 and Figure 6 we conclude that it is obvious more realistic to compute the probability of the T event with DST than without it.

On the other hand, the results of the three cases considered by us are given in Figure 7, Figure 8 and Figure 9.

In Case I (Figure 7) the mass is distributed on all 3 elements. The top event occurs with probability 0,6213 with the remaining non-true mass on the $\veebar$ (T,F) and on the false elements.

In Case II (Figure 8) the remaining non-true mass on the intermediate and final events of the fault tree is assigned to the false element of the consequences. This is due to the allocation of zero mass to the

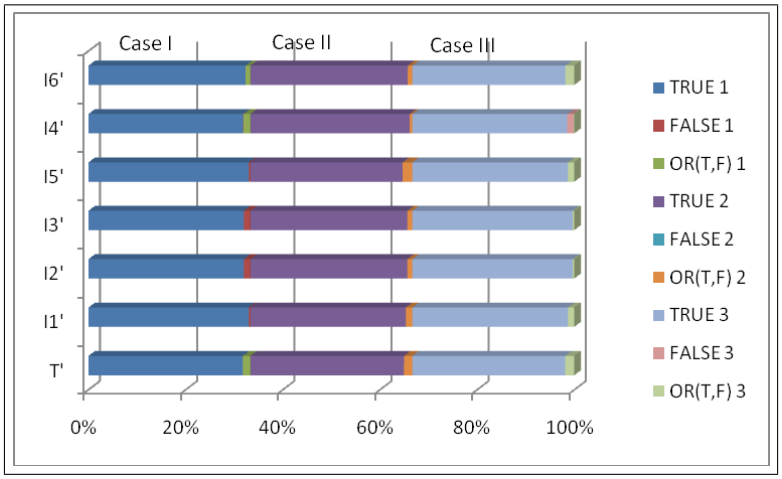


Figure 5: Calculus without imprecision on rules

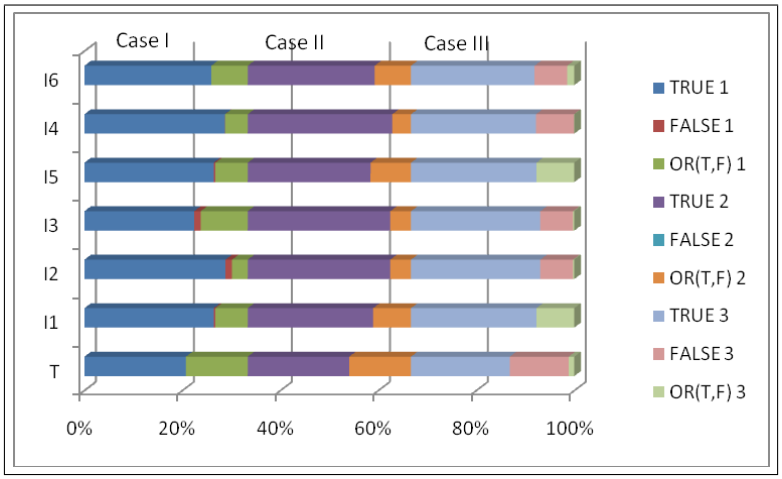
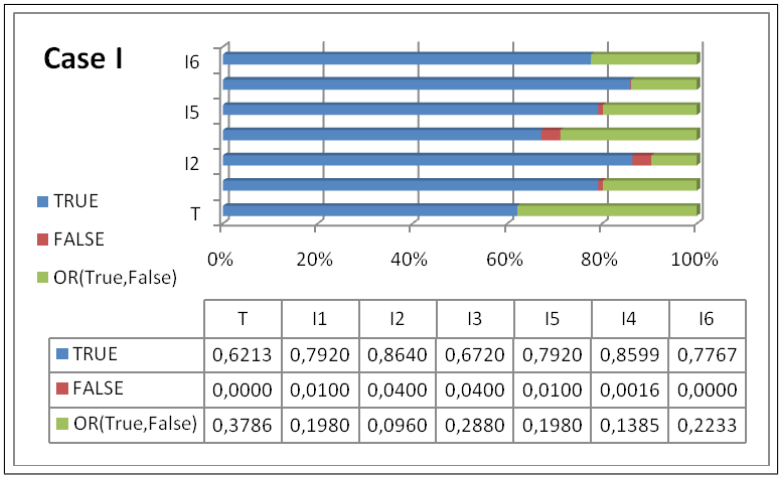


Figure 6: Calculus with imprecision on rules

false elements of the initiating events and rules. The mass of the true element of T is 0,6218, with the remaining 0,3782 mass distributed to the uncommitted term.

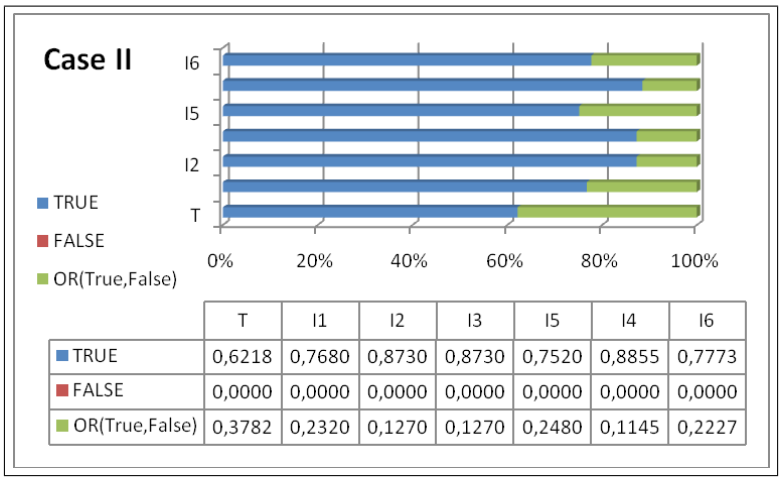
The Case III (Figure 9) reflects the most realistic case with fuzzy input data on the initiating events but clear knowledge of the logic rules. The resulting probability assignments from Figure 9 reflect the distribution of all 3 elements.

Cases I, II and III offer interesting comparisons. In case I we have only vague information on the accuracy of the rules, while in case III we have vague observations of the initiating events. The numerical estimates for the true elements are the same in both cases because both assume the same mass assignments for the true elements of the initiating events and rule parameters. Apparently, the effects of the uncertainly on the rules do not begin to offset the certainly on the initiating events until the mass is propagated up through the majority of the Boolean gates.



	T	I1	I2	I3	I5	I4	I6
TRUE	0,6213	0,7920	0,8640	0,6720	0,7920	0,8599	0,7767
FALSE	0,0000	0,0100	0,0400	0,0400	0,0100	0,0016	0,0000
OR(True,False)	0,3786	0,1980	0,0960	0,2880	0,1980	0,1385	0,2233

Figure 7: Mass Assignment for Case I

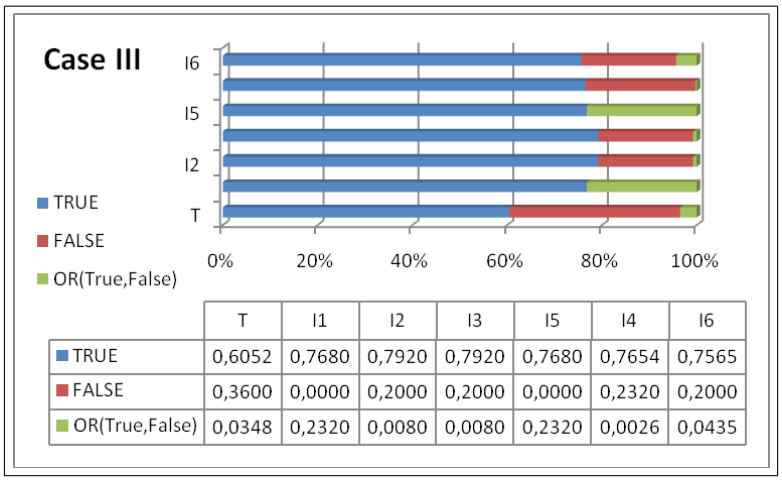


	T	I1	I2	I3	I5	I4	I6
TRUE	0,6218	0,7680	0,8730	0,8730	0,7520	0,8855	0,7773
FALSE	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
OR(True,False)	0,3782	0,2320	0,1270	0,1270	0,2480	0,1145	0,2227

Figure 8: Mass Assignment for Case II

4 Conclusions

This paper illustrate the advantages of using DST methodology (acting on binary state space with either true or false elements) for representing vagueness and imprecision in reliability analysis of net-



	T	I1	I2	I3	I5	I4	I6
TRUE	0,6052	0,7680	0,7920	0,7920	0,7680	0,7654	0,7565
FALSE	0,3600	0,0000	0,2000	0,2000	0,0000	0,2320	0,2000
OR(True,False)	0,0348	0,2320	0,0080	0,0080	0,2320	0,0026	0,0435

Figure 9: Mass Assignment for Case III

works. We used DST probability assignments for the component of fault trees and a separate parameter on each Boolean rule to show how a pattern of false-negative can be observed.

So, DST offers a more accurate representation of knowledge.

Therefore, when there is incomplete knowledge or a limited database upon which to make probability assignments, DST offers a clear advantage over binary (true, false) assignments in representing vagueness.

Bibliography

[1] Stephen D.Unwin, *A fuzzy set theory foundation for vagueness in uncertainly analysis*, Risk Analysism vol.6, num I. 1986, pp.27.34

[2] Arthur P. Dempster, *Upper and Lower probabilities induced by a multi-valued mapping* Ann Mathematical Statistics, vol.38, 1967, pp.325-339

[3] Glenn Shafer, *A Mathematical Theory of Evidence*, 1976, Princeton University Press

[4] Glen Shafer, *Bayes's two arguments for the rule conditioning*, Ann.Statistics, vol.10, 1982, pp 1075-1089

[5] Glen Shafer, *The Combination of evidence*, Int'l J. Intelligent Systems, vol.I, num.3, 1986, pp.155-176

[6] Henry Prade, *A computational approach to approximate and plausible reasoning with applications to expert systems*, IEEE Trans. Pattern Analysis and Machine Intelligence, vol.PAMI-7, 1985, May

[7] Michael A.S.Guth, *A Probability Foundation for Vagueness & Imprecision in Fault Tree Analysis*. IEEE Trans.on Reliability, vol.40, no.5, 1991, dec.

Daniela Elena Popescu (b. June 27, 1961) received her PhD in Computer Science (1998) from the "Politechnica" University Timisoara, Romania. Since 1990 she is working within the Department of Computer Science, Faculty of Electrical Engineering and Information Technology, University of Oradea, Romania, currently position occupied being professor. She is member in the Computer Architecture and Computer Testing research group and she has written more then 80 papers in international journals. Her current main research field of interest is in Computer Architecture and

Digital Circuits Design, Computers Networks and Computational Intelligent Methods. She is also member in Hungarian Technical Academy.

Lonea Alina-Madalina (b. July 16, 1984) is PhD Student at “Politehnica” University of Timisoara. She is studying “Optimisation in Grid Systems”. Her previous projects are in the following fields: Network Server Management, Network Systems, Web Application Development, Advanced Databases, Bash Scripting and Research Management Skills.

Doina Zmaranda (b. July 14, 1967) received her MSc in Computer Science (1990) and PhD in Computer Science (2001) from the "Politechnica" University Timisoara, Romania. Since 1990 she is working within the Department of Computer Science, Faculty of Electrical Engineering and Information Technology, University of Oradea, Romania, currently position occupied being professor. Her scientific research is focusing on real-time application development and programming. In addition, she also investigates issues related to reliability of complex control systems. She has (co-)authored 5 books and more than 20 papers in international journals in the last five years, participating also within several research projects.

Codruta Vancea (b. August 7, 1967) received her MSc in Informtics (1990) and PhD in Mathematics (2003) from the University "Babes Bolyai" of Cluj Napoca, Romania. Since 1991 she is working within the Department of Electrical Engineering, Electrical Measurement and Electric Power Use, Faculty of Electrical Engineering and Information Technology, University of Oradea, Romania. Her current position is assistant professor. Her scientific research is focusing on modeling of electromagnetic problems and parallel computing.

Cristian Tiurbe (b. October 22, 1979) received his MSc in Computer Science (2003) from the University of Oradea, Romania. Since 2002 he is working within the Department of Computer Science, Faculty of Electrical Engineering and Information Technology, University of Oradea, Romania, currently position occupied being assistant. His scientific research is focusing on computer networks security.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 4, pp. 567-577

A New Rymon Tree Based Procedure for Mining Statistically Significant Frequent Itemsets

P. Stanišić, S. Tomović

Predrag Stanisic, Savo Tomovic
University of Montenegro
Department of Mathematics and Computer Science
Dzordza Vasingtona bb, Podgorica, Montenegro
E-mail: pedjas@ac.me, savica@t-com.me

Abstract: In this paper we suggest a new method for frequent itemsets mining, which is more efficient than well known Apriori algorithm. The method is based on special structure called Rymon tree. For its implementation, we suggest modified sort-merge-join algorithm. Finally, we explain how support measure, which is used in Apriori algorithm, gives statistically significant frequent itemsets.
Keywords: frequent itemset mining, association analysis, Apriori algorithm, Rymon tree

1 Introduction

Finding frequent itemsets in databases is fundamental operation behind association rule mining. The problem of mining association rules over transactional databases was introduced in [1]. An example of such rule might be that "85% of customers who bought milk also bought bread". Discovering all such rules is important for planning marketing campaigns, designing catalogues, managing prices and stocks, customer relationships management etc.

The supermarket is interested in identifying associations between item sets; for example, it may be interested to know how many of customers who bought milk also bought bread. This knowledge is important because if it turns out that many of the customers who bought milk also bought bread, the supermarket will place bread physically close to milk in order to stimulate the sales of bread. Of course, such a piece of knowledge is especially interesting when there is a substantial number of customers who buy two items together and when large fraction of those individuals who buy milk also buy bread.

For example, the association rule $milk \Rightarrow bread$ [support=20%, confidence=85%] represents facts:

- 20% of all transactions under analysis contain milk and bread;
- 85% of the customers who purchased milk also purchased bread.

The result of association analysis is strong association rules, which are rules satisfying a minimal support and minimal confidence threshold. The minimal support and the minimal confidence are input parameters for association analysis.

The problem of association rules mining can be decomposed into two sub-problems [1]:

- Discovering frequent itemsets. Frequent itemsets have support higher than minimal support;
- Generating rules. The aim of this step is to derive rules with high confidence (strong rules) from frequent itemsets. For each frequent itemset l all nonempty subsets of l are found; for each $a \subset l \wedge a \neq \varnothing$ the rule $a \Rightarrow l - a$ is generated, if $\frac{support(l)}{support(a)} > minimal\ confidence$.

Overall performances of mining association rules are determined by the first step; we do not cosider the second step in this paper. Efficient algorithms for solving the second sub-problem are presented in [12].

The paper is organized as follows. Section 2 provides formalization of frequent itemsets mining

problem. Section 3 describes Apriori multiple_num algorithm which is a modification of well known Apriori algorithm [1]. Section 4 presents a new candidate generation procedure which is part of Apriori multiple_num. In section 5 we use hypothesis testing to validate generated frequent itemsets.

2 Preliminaries

Suppose that I is a finite set; we refer to the elements of I as items. We primarily use notions from [10].

Definition 1. A transaction dataset on I is a function T: {1, ...,n} → P(I), where P(I) is set of all subsets of I. The set T(k) is the k^{th} transaction of T. The numbers 1,...,n are the transaction identifiers (TIDs). [10]

Given a transaction data set T on the set I, we would like to determine those subsets of I that occur often enough as values of T. [10]

Definition 2. Let T: 1, ..., n→P(I) be a transaction data set on set of items I, where P(I) is set of all subsets of I. The support count of subset K of set of items I in T is the number $suppcount_T(K)$ given by:

$$suppcount_T(K) = |\{k|1 \leq k \leq n \wedge K \subseteq T(k)\}|. \quad (1)$$

The support of an item set K (in the following text instead of "item set K" we will use "itemset K") is the number:

$$support_T(K) = suppcount_T(K)/n. \quad (2)$$

[10]

The following rather straightforward statement is fundamental for the study of frequent itemsets. It is known as Apriori principle [1]. Proof is presented in order to introduce anti-monotone property.

Theorem 3. *Let T: 1, ..., n→ P(I) be a transaction data set on a set of items I, where P(I) is set of all subsets of I. If K and K' are two itemsets, then $K' \subseteq K$ implies $support_T(K') \geq support_T(K)$. [10]*

Proof: The previous theorem states that $support_T$ for an itemset has the anti-monotone property. It means that support for an itemset never exceeds the support for its subsets. For proof, it is sufficient to note that every transaction that contains K also contains K'. The statement from the theorem follows immediately. □

Definition 4. An itemset K is μ-frequent relative to the transaction data set T if $support_T(K) \geq \mu$. We denote by F_T^μ the collection of all μ-frequent itemsets relative to the transaction data set T and by $F_{T,r}^\mu$ the collection of μ-frequent itemsets that contain r items for $r \geq 1$ (in the following text we will use r-itemset to denote itemset that contains r items). [10]

Note that $F_T^\mu = \bigcup_{r\geq 1} F_{T,r}^\mu$. If it is clear what μ and T are, we can omit them.

In this paper we will propose new algorithm for frequent itemsets mining which is based on special structure: Rymon tree. The Rymon tree was introduced in [8] in order to provide a unified search-based framework for several problems in artificial intelligence; the Rymon tree is also useful for data mining algorithms. In Definition 5 and 6 we define necessary concepts and in Definition 7 we define the Rymon tree.

Definition 5. Let S be a set and let $d : S \to N$ be an injective function. The number $d(x)$ is the index of $x \in S$. If $P \subseteq S$, *view of* P is subset $view(d,P) = \{s \in S | d(s) > max_{p\in P} d(p)\}$.

Definition 6. A collection of sets C is *hereditary* if $U \in C$ and $W \subseteq U$ implies $W \in C$.

Definition 7. Let C be a hereditary collection of subsets of a set S. The graph $G = (C, E)$ is a Rymon tree for C and the indexing function d if:

- the root of the G is $\emptyset$
- the children of a node P are the sets of the form $P \cup \{s\}$, where $s \in view(d, P)$

If $S = \{s_1, ..., s_n\}$ and $d(s_i) = i$ for $1 \leq i \leq n$, we will omit the indexing function from the definition of the Rymon tree for $P(S)$.

Let $S = \{i_1, i_2, i_3, i_4\}$ and let C be $P(S)$, which is clearly a hereditary collection of sets. Finally, let d be injective mapping: $d(i_k) = k$ for $1 \leq k \leq 4$. The Rymon tree for C and d is shown in Fig. 1.

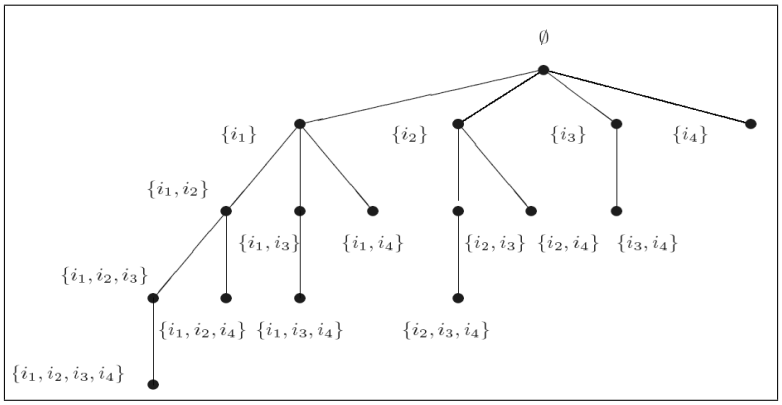


Figure 1: Example of Rymon tree

A key property of a Rymon tree is stated next.

Theorem 8. *Let* G *be a Rymon tree for a hereditary collection* C *of subsets of a set* S *and an indexing function* d*. Every set* P *of* C *occurs exactly once in the tree.*

Note that in the Rymon tree of a collection $P(S)$, the collection S_r, that consists of sets located at distance r from the root, denotes all subsets of the size r of S.

3 Apriori multiple_num Algorithm

Apriori multiple_num algorithm generates frequent itemsets starting with frequent 1-itemsets (itemsets consisted of just one item). Next, the algorithm iteratively generates frequent itemsets to the maximal length of frequent itemset. Each iteration of the algorithm consists of two phases: *candidate generation* and *support counting*.

In candidate generation phase potentially frequent itemsets or candidate itemsets are generated. The Apriori principle [1] is used in this phase. It is based on anti-monotone property of the itemset support (see Theorem 3) and provides elimination or pruning of some candidate itemsets without calculating its support. According to the Apriori principle, if X is frequent itemset, then all its subsets are also frequent. This fact is used in candidate generation phase in a way that the candidate containing at least one not frequent subset is being pruned immediately (before support counting phase).

Support counting phase consists of calculating support for all previously generated candidates (which are not pruned according to the Apriori principle in the candidate generation phase). Calculating candidate support requires one database scan and efficient determination if the candidates are contained in

particular transaction $t \in T$. For candidates contained in $t \in T$, it's support will be incremented. On account of that, the candidates are organized in hash tree. The candidates which have enough support are termed as frequent itemsets.

The main difference between iterations in original Apriori algorithm [1] and Apriori multiple_num algorithm is that iterations in later one are "longer", which is determined by *multiple_num* parameter. Actually, in original Apriori algorithm in the iteration k set F_k(containing all frequent itemsets with k items) is generated, while Apriori multiple_num algorithm in the iteration k generates sets $F_{k+i}, 0 \leq i \leq multiple\_num$. If $k_{Max} < multiple\_num$ is true, where k_{Max} is the maximal length of frequent itemset, Apriori multiple_num algorithm terminates in just two iterations, or just two database scans.

In addition, all candidate k-itemsets (itemsets containing k items) will be signed as C_k, and all frequent k-itemsets as F_k. Pseudocode for Apriori multiple_num algorithm is given bellow.

Apriori multiple_num Algorithm
Input: T-transactional database; μ -minimal support;
Output: F-frequent itemsets in T
Method:

1. $F_1 = all\_large\_1itemsets(T, \mu)$
2. $multiple\_num = maximal\_length\_of\_transactions$
3. $C_2 = apriori\_gen(F_1, F_1)$
4. FOR i=3 TO multiple_num
   $C_i = apriori\_gen(C_{i-1}, C_{i-2})$
   END FOR
5. FOR i=2 TO multiple_num
   $createCandidateHashtree(C_i)$
   END FOR
6. FOR EACH $t \in T$ DO
   FOR i=2 TO multiple_num
   $traverseHashtree(C_i, t)$
   END FOR
   END FOR
7. FOR i=2 TO multiple_num
   $F_i = \{c \in C_i | support(c) \geq \mu\}$
   END FOR
8. $F = \bigcup_k F_k$

Let us explain the most important steps briefly. Generating frequent 1-itemsets is done in the same way as in original Apriori algorithm [1]. This step requires one database scan. Then, parameter *multiple_num* is set to the length of the longest transaction from the database T, which ensures that the algorithm will need just one more database scan. Steps 3 and 4 are concerned with candidate generation: in step 3 set C_2 is generated by calling apriori_gen function, then loop in step 4 generates all other candidates $C_i, 3 \leq i \leq multiple\_num$, by calling apriori_gen function, but with the following difference. According to original Apriori algorithm [1] candidate itemsets C_{k+1} (candidate itemsets containing k+1 items) is formed from the set F_k(frequent itemsets containing k items) in iteration k+1. However, we want to generate all itemsets in just one loop in order to reduce number of iterations (database scans) to two, but we do not have the necessary frequent sets. As the solution, arguments are candidate sets C_{k-1} and C_{k-2}, which is known at this moment. The next section describes modification of apriori_gen function and its fast implementation.

The support counting phase comes next. All candidate itemsets $C_i, 2 \leq i \leq multiple\_num$ are orga-

nized in separate hash trees in order to make support counting process efficient. Then, we scan database and calculate support for candidates by traversing corresponding hash trees. At the end of support counting phase frequent itemsets $F_i, 1 \leq i \leq$ *multiple_num* are generated. As we stated earlier, we will not further consider support counting phase.

Apriori algorithm [1] performs $k_{max}+1$ iterations, where k_{max} is the maximal length of frequent itemsets, and in each iteration it scans whole database. Apriori multiple_num algorithm finishes after 2 iterations and performs 2 database scans.

4 New Procedure for Candidate Generation

We assume that any itemset K is kept sorted according to some relation $<$, where for all $x, y \in K$, $x < y$ means that object x is in front of object y. Also, we assume that all transactions in database T and all subsets of K are kept sorted in lexicographic order according to relation $<$.

For candidate generation we suggest an original method by which the set $C_{T,k}$ is calculated by joining $C^{\mu}_{T,k-1}$ with $C^{\mu}_{T,k-2}$, for $k \geq 3$. Candidate k-itemset is created from one candidate (k-1)-itemset and one candidate (k-2)-itemset in the following way. Let $X = \{x_1, ..., x_{k-1}\} \in C^{\mu}_{T,k-1}$ and $Y = \{y_1, ..., y_{k-2}\} \in C^{\mu}_{T,k-2}$. Itemsets X and Y are joined if and only if the following condition is satisfied:

$$x_i = y_i, (1 \leq i \leq k-3) \wedge x_{k-1} < y_{k-2} \tag{3}$$

producing the candidate k-itemset $\{x_1, ..., x_{k-2}, x_{k-1}, y_{k-2}\}$.

We will prove the correctness of the suggested method. In the following text we will denote this method by $C_{T,k} = C^{\mu}_{T,k-1} \times C^{\mu}_{T,k-2}$. Let $I = i_1, ..., i_n$ be a set of items that contains n elements. Denote by $G_I = (P(I), E)$ the Rymon tree of $P(I)$. The root of the tree is $\emptyset$. A vertex $K = \{i_{p_1}, ..., i_{p_k}\}$ with $i_{p_1} < i_{p_2} < ... < i_{p_k}$ has $n - i_{p_k}$ children $K \bigcup j$, where $i_{p_k} < j \leq n$. Let S_r be the collection of itemsets that have r elements. The next theorem suggest a technique for generating S_r starting from S_{r-1} and S_{r-2}. It is a modification of Theorem 7.8. from [10].

Theorem 9. *Let G be the Ryman tree of P(I), where $I = i_1, ..., i_n$. If $W \in S_r$, where $r \geq 3$, then there exists a unique pair of distinct sets $U \in S_{r-1}$ and $V \in S_{r-2}$ that has a common immediate ancestor $T \in S_{r-3}$ in G such that $U \bigcap V \in S_{r-3}$ and $W = U \bigcup V$.*

Proof: Let u and v and p be the three elements of W that have the largest, the second-largest and the third-largest subscripts, respectively. Consider the sets $U = W - \{u\}$ and $V = W - \{v, p\}$. Note that $U \in S_{r-1}$ and $V \in S_{r-2}$. Moreover, $Z = U \bigcup V$ belongs to S_{r-3} because it consists of the first r-3 elements of W. Note that both U and V are descendants of Z and that $U \bigcup V = W$(for r=3 we have $Z = \emptyset$).

The pair (U, V) is unique. Indeed, suppose that W can be obtained in the same manner from another pair of distinct sets $U_1 \in S_{r-1}$ and $V_1 \in S_{r-2}$ such that U_1 and V_1 are immediate descendants of a set $Z_1 \in S_{r-3}$. The definition of the Rymon tree G_I implies that $U_1 = Z_1 \bigcup \{i_m, i_q\}$ and $V_1 = Z_1 \bigcup \{i_y\}$, where the letters in Z_1 are indexed by a number smaller than $min\{m, q, y\}$. Then, Z_1 consists of the first r-3 symbols of W, so $Z_1 = Z$. If $m < q < y$, then m is the third-highest index of a symbol in W, q is the second-highest index of a symbol in W and y is the highest index of a symbol in W, so $U_1 = U$ and $V_1 = V$. □

The following theorem, together with the obvious fact $C^{\mu}_{T,k} \subset F^{\mu}_{T,k}$ for all k, directly proves correctness of our method $C_{T,k} = C^{\mu}_{T,k-1} \times C^{\mu}_{T,k-2}$. It is modification of Theorem 7.10. from [10].

Theorem 10. *Let T be a transaction data set on a set of items I and let $k \in N$ such that $k > 2$. If W is a μ-frequent itemset and $|W| = k$, then there exists a μ-frequent itemset Z and two itemsets $\{i_m, i_q\}$ and $\{i_y\}$ such that $|Z| = k-3$, $Z \subseteq W$, $W = Z \bigcup \{i_m, i_q, i_y\}$ and both $Z \bigcup \{i_m, i_q\}$ and $Z \bigcup \{i_y\}$ are μ-frequent itemsets.*

Proof: If W is an itemset such that $|W| = k$, than we already know that W is the union of two subsets U and V of I such that $|U| = k-1$, $|V| = k-2$ and that $Z = U \bigcap V$ has k-3 elements (it follows from Theorem 2). Since W is a μ-frequent itemset and Z, U, V are subsets of W, it follows that each of these sets is also a μ-frequent itemset (it follows from Theorem 1). □

Apriori algorithm [2] generates candidate k-itemset by joining two large (k-1)-itemsets, if and only if they have first (k-2) items in common. Because of that, each join operation requires (k-2) equality comparisons. If a candidate k-itemset is generated by the method $C_{T,k} = C^{\mu}_{T,k-1} \times C^{\mu}_{T,k-2}$ for $k \geq 3$, it is enough to process (k-3) equality comparisons.

The method $C_{T,k} = C^{\mu}_{T,k-1} \times C^{\mu}_{T,k-2}$ can be represented by the following SQL query:

INSERT INTO $C_{T,k}$
SELECT $R_1.item_1, ..., R_1.item_{k-1}, R_2.item_{k-2}$
FROM $C^{\mu}_{T,k-1}$ **AS** $R_1, C^{\mu}_{T,k-2}$ **AS** R_2
WHERE $R_1.item_1 = R_2.item_1 \wedge ... \wedge R_1.item_{k-3} = R_2.item_{k-3} \wedge R_1.item_{k-1} < R_2.item_{k-2}$

For the implementation of the join $C_{T,k} = C^{\mu}_{T,k-1} \times C^{\mu}_{T,k-2}$ we suggest a modification of sort-merge-join algorithm (note that $C^{\mu}_{T,k-1}$ and $C^{\mu}_{T,k-2}$ are sorted because of the way they are constructed and lexicographic order of itemsets).

By the original sort-merge-join algorithm [9], it is possible to compute natural joins and equi-joins. Let $r(R)$ and $s(S)$ be the relations and $R \bigcap S$ denote their common attributes. The algorithm keeps one pointer on the current position in relation $r(R)$ and another one pointer on the current position in relation $s(S)$. As the algorithm proceeds, the pointers move through the relations. It is supposed that the relations are sorted according to joining attributes, so tuples with the same values on the joining attributes are in consecutive order. Thereby, each tuple needs to be read only once, and, as a result, each relation is also read only once.

The number of blocks transfers is equal to the sum of the number of blocks in both sets $C^{\mu}_{T,k-1}$ and $C^{\mu}_{T,k-2}$, $n_{b1} + n_{b2}$.

The modification of sort-merge-join algorithm we suggest refers to the elimination of restrictions that join must be natural or equi-join. First, we separate the condition (3):

$$x_i = y_i, 1 \leq i \leq k-3 \tag{4}$$

$$x_{k-1} < y_{k-2}. \tag{5}$$

Joining $C_{T,k} = C^{\mu}_{T,k-1} \times C^{\mu}_{T,k-2}$ is calculated according to the condition (4), in other words we compute natural join. For this, the described sort-merge-join algorithm is used, and our modification is: before $X = \{x_1, ..., x_{k-1}\}$ and $Y = \{y_1, ..., y_{k-2}\}$, for which $X \in C^{\mu}_{T,k-1}$ and $Y \in C^{\mu}_{T,k-2}$ and $x_i = y_i, 1 \leq i \leq k-3$ is true, are joined, we check if condition (5) is satisfied, and after that we generate candidate k-itemset $\{x_1, ..., x_{k-2}, x_{k-1}, y_{k-2}\}$.

The pseudocode of **apriori_gen** function comes next.

```
FUNCTION apriori_gen(C^μ_{T,k-1}, C^μ_{T,k-2})
   1.  i = 0
   2.  j = 0
   3.  while i ≤ |C^μ_{T,k-1}| ∧ j ≤ |C^μ_{T,k-2}|
         iset_1 = C^μ_{T,k-1}[i++]
         S = {iset_1}
         done = false
         while done = false ∧ i ≤ C^μ_{T,k-1}
```

```
        iset_1a = C^μ_{T,k-1}[i++]
        if iset_1a[w] = iset_1[w], 1 ≤ w ≤ k-2 then
           S = S ⋃ {iset_1a}
           i++
        else
           done = true
        end if
     end while
     iset_2 = C^μ_{T,k-2}[j]
     while j ≤ |C^μ_{T,k-2}| ∧ iset_2[1,...,k-2] ≺ iset_1[1,...,k-2]
        iset_2 = C^μ_{T,k-2}[j++]
     end while
     while j ≤ |C^μ_{T,k-2}| ∧ iset_1[w] = iset_2[w], 1 ≤ w ≤ k-2
        for each s ∈ S
           if iset_1[k-1] ≺ iset_2[k-2] then
              c = {iset_1[1],...,iset_1[k-1], iset_2[k-2]}
              if c contains-not-frequent-subset then
                 DELETE c
              else
                 C_{T,k} = C_{T,k} ⋃ {c}
              end if
        end for
        j++
        iset_2 = C^μ_{T,k-2}[j]
     end while
  end while
```

5 Statistical Test for Validating Frequent Itemset

Frequent itemset mining algorithms have the potential to generate a large number of patterns. For example, even if we assume that no customer has more than five items in his shopping cart and that there are 10000 items, there are $\sum_{i=1}^{5} \begin{pmatrix} 10000 \\ 5 \end{pmatrix}$ possible contents of this cart, which corresponds to the subsets having no more than five items of a set that has 10,000 items, and this is indeed a large number. As the size and dimensionality of real commercial databases can be very large, we could easily end up with thousands or even millions of patterns, many of which might not be interesting. It is therefore important to establish a set of well-accepted criteria for evaluating the quality of patterns.

In Apriori multiple_num algorithm support measure is used to determine whether an itemset is frequent: an itemset X is considered frequent in the data set T, if $supp_T(X) > minsup$, where *minsup* is a user-specified threshold. Support measure is kind of *objective interestingness measure*, which is data-driven and domain-independent approach that uses statistics derived from data for evaluating the quality of association patterns [12].

Now, we will explain how statistical hypothesis testing can be applied to validate frequent itemsets generated with support measure.

Hypothesis testing is a statistical inference procedure to determine whether a hypothesis should be accepted or rejected based on the evidence gathered from data. Examples of hypothesis tests include verifying the quality of patterns extracted by many data mining algorithms and validating the significance

of the performance difference between two classification models.

In hypothesis testing, we are usually presented with two opposite hypothesis, which are known, respectively, as the null hypothesis and the alternative hypothesis. The general procedure for hypothesis testing consists of the following four steps [12]:

- Formulate the null and alternative hypotheses to be tested.
- Define a test statistic θ that determines whether the null hypothesis should be accepted or rejected. The probability distribution associated with the test statistic should be known.
- Compute the value of θ from the observed data. Use the knowledge of the probability distribution to determine a quantity known as p-value.
- Define a significance level a which controls the range of θ values in which the null hypothesis should be rejected. The range of values for θ is known as the rejection region.

Frequent itemsets mining problem can be formulated into the hypothesis testing framework in the following way. To validate if the itemset X is frequent in the data set T, we need to decide whether to accept the null hypothesis, $H_0 : supp_T(X) = minsup$, or the alternative hypothesis $H_1 : supp_T(X) > minsup$. If the null hypothesis is rejected, then X is considered as frequent itemset. To perform the test, the probability distribution for $supp_T(X)$ must also be known.

Theorem 11. *The measure $supp_T(X)$ for the itemset X in transaction data set T has the binomial distribution with mean $supp_T(X)$ and variance $\frac{supp_T(X)*(1-supp_T(X))}{n}$, where n is the number of transactions in T.*

Proof: We will use measure $suppcount_T(X)$ and calculate mean and variance for it and later derive mean and variance for the measure $supp_T(X)$. The measure $suppcount_T(X) = X_n$ presents the number of transactions in T that contain itemset X, and $supp_T(X) = suppcount_T(X)/n$ (Definition 2).

The measure X_n is analogous to determining the number of heads that shows up when tossing n coins. Let us calculate $E(X_n)$ and $D(X_n)$.

Mean is $E(X_n) = n * p$, where p is the probability of success, which means (in our case) the itemset X appears in one transaction. According to Bernoulli low, the following holds:

$$\forall \varepsilon > 0, lim_{N\to\infty} P\{|\frac{X_n}{n} - p| \leq \varepsilon\} = 1. \tag{6}$$

Freely speaking, for large n (we work with large databases so n can be considered large), we can use relative frequency instead of probability. So, we now have:

$$E(X_n) = np \approx n\frac{X_n}{n} = X_n \tag{7}$$

For variance we compute:

$$D(X_n) = np(1-p) \approx n\frac{X_n}{n}(1-\frac{X_n}{n}) = \frac{X_n(n-X_n)}{n} \tag{8}$$

Now we will compute $E(supp_T(X))$ and $D(supp_T(X))$. Recall that $supp_T(X) = \frac{X_n}{n}$. We have:

$$E(supp_T(X)) = E(\frac{X_n}{n}) = \frac{1}{n}E(X_n) = \frac{X_n}{n} = supp_T(X). \tag{9}$$

$$D(supp_T(X)) = D(\frac{X_n}{n}) = \frac{1}{n^2}D(X_n) = \frac{1}{n^2}\frac{X_n(n-X_n)}{n} = \frac{1}{n}supp_T(X)(1-supp_T(X)) \tag{10}$$

□

The binomial distribution can be further approximated using normal distribution if n is sufficiently large, which is typically the case in association analysis.

Regarding previous paragraph and Theorem 4, under the null hypothesis $supp_T(X)$ is assumed to be normally distributed with mean *minsup* and variance $\frac{minsup(1-minsup)}{n}$. To test whether the null hypothesis should be accepted or rejected, the following statistic can be used:

$$W_N = \frac{supp_T(X) - minsup}{\sqrt{\frac{minsup(1-minsup)}{n}}}. \tag{11}$$

The previous statistic, according to the Central Limit Theorem, has the distribution $N(0,1)$. The statistic essentially measures the difference between the observed support $supp_T(X)$ and the *minsup* threshold in units of standard deviation.

Let N=10000, $supp_T(X) = 0.11$, *minsup*=0.1 and $\alpha = 0.001$. The last parameter is the desired significance level. It controls Type 1 error which is rejecting the null hypothesis even though the hypothesis is true.

In the Apriori algorithm we compare $supp_T(X) = 0.11 > 0.1 = minsup$ and we declare X as frequent itemset. Is this validation procedure statistically correct?

Under the hypothesis H_1 statistics W_{10000} is positive and for rejection region we choose
$R = \{(x_1,...,x_{10000}) | w_{10000} > k\}, k > 0$.

Let us find k.

$0.001 = P_{H_0}\{W_{10000} > k\}$
$P_{H_0}\{W_{10000} > k\} = 0.499$
$k = 3.09$

Now we compute $w_{10000} = \frac{0.11-0.1}{\sqrt{\frac{0.1*(1-0.1)}{10000}}} = 3.33\cdots$

We can see that $w_{10000} > k$, so we are in rejection region and H_1 is accepted, which means the itemset X is considered statistically significant.

6 Conclusion

In this section we compare the proposed method with original Apriori [1] and with Apriori Multiple which we introduced in [11].

Sections 3 and 4 of the paper contain comparison with the original Apriori algorithm [1]. The main advantages of the new algorithm are: it finishes in just two database scans and it uses more efficient candidate generation procedure.

The algorithm from [11] also finishes in two database scans and it uses similar procedure for candidate generation as the Apriori multiple_num algorithm proposed here. But, the Apriori multiple_num algorithm is more efficient. The main advantage of the Apriori multiple_num algorithm in comparison with algorithm from [11] is in the following. The Apriori multiple_num uses Rymon tree structure for definition of candidate join procedure as it is explained in Section 4. Because of that, candidate sets are stored in Rymon tree structure before joining instead of storing candidates in array as it is done in [11]. The following experiment confirms that Rymon tree based implementation is more efficient.

We implemented the Apriori multiple_num, the original Apriori [1] and the Apriori Multiple [11] algorithms in C in order to evaluate its performances. Experiments are performed on PC with a CPU Intel(R) Core(TM)2 clock rate of 2.66GHz and with 2GB of RAM. Also, run time used here means the

total execution time, i.e., the period between input and output instead of CPU time measured in the experiments in some literature. In experiments dataset which can be found on www.cs.uregina.ca is used. It contains 10000 binary transactions. The average length of transactions is 8.

We did not compare number of I/O operations because the algorithm proposed here finishes in just two database scans, while the original Apriori requires at least $k_{max} + 1$ scans, where k_{max} is the length of the longest frequent itemset (as explained in Section 3).

Figure 1 shows that the original Apriori algorithm from [1] is outperformed by both the Apriori Multiple [11] and the Apriori multiple_num presented here. Also, it can be seen that Apriori multiple_num with Rymon tree based implementation is significantly better than the algorithm from [11].

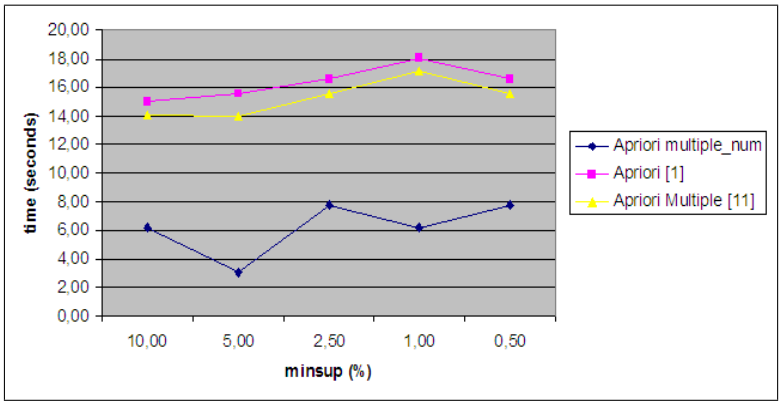


Figure 2: Execution times for different algorithms

Bibliography

[1] Agrawal, R., Srikant, R., Fast Algorithms for Mining Association Rules, Proceedings of VLDB-94, 487-499, Santiago, Chile (1994)

[2] Coenen, F.P., Leng, P., Ahmed, S., T-Trees, Vertical Partitioning and Distributed Association Rule Mining, Proceedings ICDM-2003, 513-516 (2003)

[3] Coenen, F.P., Leng, P., Ahmed, S., Data Structures for Association Rule Mining: T-trees and P-trees, IEEE Transactions on Data and Knowledge Engineering, Vol. 16, No 6, 774-778 (2004)

[4] Coenen, F.P., Leng, P., Goulbourne, G., Tree Structures for Mining Association Rules, Journal of Data Mining and Knowledge Discovery Vol. 8, No. 1, 25-51 (2004)

[5] Goulbourne, G., Coenen, F., Leng, P., Algorithms for Computing Association Rules Using a Partial-Support Tree, Journal of Knowledge-Based Systems Vol. 13, 141-149 (1999)

[6] Grahne, G., Zhu, J., Efficiently Using Prefix-trees in Mining Frequent Itemsets, Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (2003)

[7] Han, J., Pei, J., Yu, P.S., Mining Frequent Patterns without Candidate Generation, Proceedings of the ACM SIGMOD Conference on Management of Data, 1-12 (2000)

[8] Rymon, R., Search Through Systematic Set Enumeration, Proceedings of 3rd International Conference on Principles of Knowledge Representation and Reasoning, 539-550 (1992)

[9] Silberschatz, A., Korth, H. F., Sudarshan, S., Database System Concepts, Mc Graw Hill, New York (2006)

[10] Simovici, A. D., Djeraba, C., Mathematical Tools for Data Mining (Set Theory, Partial Orders, Combinatorics), Springer-Verlag London Limited (2008)

[11] Stanisic, P., Tomovic, S., Apriori Multiple Algorithm for Mining Association Rules, Information Technology and Control Vol. 37, No. 4, 311-320 (2008)

[12] Tan., P.N., Steinbach, M., Kumar, V., Introduction to Data Mining, Addicon Wesley (2006).

Dr. Predrag Stanisic is a professor in the Faculty of Science - Department of Mathematics and Computer Science at University of Montenegro. He received his B.Sc. degree in Mathematics and Computer Science from University of Montenegro in 1996, his M.Sc degree in Computer Science from University of Belgrade Serbia in 1998 and his Ph.D. degree in Computer Science from Moscow State University M.V. Lomonosov in 1999. He is currently the dean of Faculty of Science at University of Montenegro and he teaches a wide variety of undergraduate and graduate courses in several computer science disciplines, especially database systems, operating systems and programming.

M.Sc Savo Tomovic is a teaching assistant in the Faculty of Science - Department of Mathematics and Computer Science at University of Montenegro. He received his B.Sc. degree in Mathematics and Computer Science from University of Montenegro in 2006, his M.Sc degree in Computer Science from University of Montenegro in 2007. He is currently a Ph.D. student in Computer Science at University of Montenegro.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 4, pp. 578-585

An Ontology to Model e-portfolio and Social Relationship in Web 2.0 Informal Learning Environments

D. Taibi, M. Gentile, G. Fulantelli, M. Allegra

Davide Taibi, Manuel Gentile, Giovanni Fulantelli, Mario Allegra
Italian National Research Council
Institute for Educational Technologies
Via Ugo La Malfa 153
90146 Palermo, Italy
E-mail: {davide.taibi,manuel.gentile,giovanni.fulantelli,mario.allegra}@itd.cnr.it

Abstract: Web 2.0 applications and the increasingly use of social networks have been creating new informal learning opportunities. Students interact and collaborate using new learning environments which are structurally different from traditional e-learning environments. In these informal unstructured learning contexts the boundaries between the learning contexts and social spheres disappear, and the definition of the students competences appears more and more important. In this paper we propose a semantic web approach in order to create the basis for a software platform to model learner profiles.
In particular we propose to extend the FOAF ontology, used to describe people and their personal relationships, with an ontology related to the IMS Learning Portfolio used to model students' competencies. This ontology could be a fundamental layer for a new Web 2.0 learning environment in which students' informal learning activities carried out in social networks can be managed and evaluated.
Keywords: semantic web, e-portfolio, social networks, informal learning.

1 Informal learning and social communities

The significant changes in society that Castells in [1] sums up in what he calls "The Rise of the network society" also have considerable implications in the definition of learning activities. The Information Society dramatically increases the opportunities for knowledge acquisition. Beyond the structured training activities designed by specialists in the education field, we have to consider the large number of educational opportunities related to everyday activities that define the so-called "informal learning" [2]. In this perspective, the concept of networked learning is drastically changing. The informal learning opportunities created by information technologies, such as Web 2.0 applications and social networks, allow users to interact and collaborate in new ways thus leading to the definition of new learning environments; these are structurally different from traditional e-learning environments, since the boundaries between the learning contexts and other social spaces tend to disappear. In these unstructured learning contexts, the definition of the skills acquired by the users is a central objective. Consequently, the use of software environments that model learner profiles and can deal with them in a semantic way appears increasingly important.

In [3] the authors argue that learning is related to the activities, the environmental and cultural contexts in which it is developed and therefore social interaction is a critical factor. From this point of view learning can be described as a process: students are involved in a community of practice that represents knowledge and behavior in which students play a more active role in the cultural sphere. The concept of situated learning comes from Vygotsky's social development theory, which affirms that social interaction has a fundamental role in the knowledge development process [4].

This theory argues that situated learning is generally unintentional and for this reason learning is more effective if the student is a member of a community of practice he has chosen to join rather than

being assigned to a group by external actors such as teachers. The social aspect in learning activities is extremely important and leads to a further consideration. For example, Tinto claims that participation in a collaborative learning group allows students to develop a supporting network, that helps students to maintain relations with a wider social community [5].

A peer-to-peer community promotes participation in learning activities. Moreover, the communities of learners provide students with the opportunity to satisfy simultaneously both social and academic requirements. These unstructured learning contexts give rise to the need to measure and asses the acquired knowledge; the traditional competence based certification systems are not designed for this type of environment and for this reason are less suitable in this kind of educational context. The semantic web provides a technological substrate which can overcome the limits of current web technologies, setting the base for creating ontological systems in order to model competences in informal educational contexts that are being developed in web 2.0 environments.

In this paper we consider the problems connected to the description of competences in informal learning environments within social networks mediated by technologies. In particular, we propose the integration of the FOAF (Friend Of A Friend) ontology, which is used to model people and their personal contacts, with semantic ontology related to student e-portfolios used to model their competences. The use of ontologies and the surrounding semantic web technologies allow us to create relationships between the students ongoing educational experiences and the evolution of their social network. For this to happen, we integrate FOAF ontology with the IMS Learning Portfolio model in order to support the creation of a new Web 2.0 learning environment based on social networks and competences.

2 Social Semantic Web

At present a huge amount of shared contents such as bookmarks, images, videos and photos are being created within so called web 2.0 applications, very popular in social and personal spheres as well as in professional and organizational ones. They possess common features like the creation and sharing of contents (images, photos, papers), discussions (comments) and connections between users (group of friends, private messages, and so on). This scenario raises new considerations related to the sharing of social contributions between software applications and the interoperability of social networks.

Due to the heterogeneity of the nature of social contribution, sharing, searching, connecting and retrieving these kinds of contents has become more complex. The semantic web technologies provide standards and models which are useful for creating a network of data, with unified models which can represent data from different sources appropriately. The unification of semantic web technologies and social paradigms gives rise to "Social Semantic Information Spaces" in which information is socially created and managed, as well as being interconnected and available in a machine understandable format, promoting new methodologies to discover information present on the web [6].

Moreover, the semantic web offers a generic infrastructure to interchange, integrate and reuse structural data, in order to overcome the limits of Web 2.0 platforms. Currently, in fact, web 2.0 applications have search mechanisms based mainly on tags and few keywords. Adding semantics to the web would enable this kind of problem to be solved, by providing easier search mechanisms, supporting the reuse of contents and creating more connections between different types of contents. Moreover, the use of ontologies is useful to structure and elaborate information. Ontologies represents entity-relationship models related to a specific knowledge or practice domain. A typical web ontology contains the definition of classes, objects and their relationships, and a set of deduction rules that give inferential power about the concepts.

Through ontologies the semantic web provides the basis for enriching the resources description with a well defined meaning and in a comprehensible format which can be elaborated by software applications.

The strict relationship between documents produced in web 2.0 environments and the specific social network [7] bring us to consider the information objects as the result of the activities of the network;

consequently, we should also represent social relationships in a well structured way, using approaches based on the semantic web concepts.

FOAF, the acronym of Friend of a friend, uses semantic web technologies, in particular the Resource Description Framework (RDF) and the Ontology Web Language (OWL), to define a machine-readable ontology describing people, their activities and their relations to other people and objects. FOAF is useful for describing social networks and their relationships.

3 Learner model in social semantic networks

Educational activities adopt web 2.0 technologies and social networks more and more frequently. The interaction paradigms at the basis of web 2.0 technologies and social networks are different from those adopted by conventional e-learning tools.

FOAF is considered the most common vocabulary for constructing social networks, it has been very successful in the applications that use semantic web technologies, and it is useful to model learner in social networks [8].

Each student can be described through an FOAF file that can be extended and modified at any time. This is useful, for example, for publishing data regarding students using the URI that represent them. To facilitate the creation of the profiles it is possible to use an interface based on foaf-a-matic; in this way, it will be possible to describe social links in the community of practice. Defining ontologies it is possible to use a inferential engine like Jena and the SPARQL language to work with data and generate new knowledge about the domain. As reported by [9] there are several advantages in using the FOAF approach to model student profiles:

- the use of RDF facilitates extensibility and interoperability
- the presence of different extensions of the FOAF vocabulary, makes FOAF very flexible
- the creation of FOAF files is simplified by the use of foaf-a-matic
- FOAF simplifies the identification of people with common interests, which is essential for creating communities of practice

To use FOAF in a learning context, it is necessary to extend this model to include specific characteristics related to learning. Regarding this aspect we should take into consideration elements related to: the extension of the FOAF vocabulary to include specific information regarding students activities; the consideration of privacy problems in sharing personal information; the evaluation of the ties strength between students belonging to the group.

In conclusion, using FOAF as a basis for learning models makes it possible to exploit the benefits of the numerous existing tools, and also to use the extension of the model to define specific aspects and personal and group relationships, which are indispensable for creating and supporting social learning networks.

FOAF is used successfully to describe a student profile, in particular the profile can be extended to bring together information coming from other models containing student data, like, for example the competences described following the IMS e-portfolio standard. The model of students competences plays a key role in making the use of social networks to better support learning activities.

4 e-portfolio and ontologies in social learning environments

An e-portfolio is defined by the EDUCAUSE NLII (National Learning Infrastructure Initiative) as *"a collection of authentic and diverse evidence, drawn from a larger archive, that represents what a person*

or organization has learned over time, on which the person or organization has reflected, designed for presentation to one or more audiences for a particular rhetorical purpose."

As sustained by [10], an e-portfolio can be used to developmental, presentation, assessment purposes, and it can contain different information related to personal and professional achievements, competences, digital works. This relevant information about students can be stored and maintained by different institutions in different sites, so the management can be improved by the use of web-based e-portfolios. A key concept in this scenario is the interoperability between different institutional systems which requires a unified model describing students e-portfolios.

The pedagogical objectives of e-portfolios are various: they allow students to describe their learning path, increase awareness of their strengths and weaknesses, take responsibility and increase their autonomy and have a unified way of presenting their competences.

At present, the lack of common standards to describe e-portfolio information means that most e-portfolio systems are using different proprietary formats to store this type of information, and moreover, they don't provide features for importing and exporting e-portfolio information from other systems. In this scenario the interoperability between e-portfolio systems is hindered, and for example, it is difficult to integrate the e-portfolio information coming from a university system and from an enterprise.

For these reasons it is desirable to use a common standard in order to unify the description processes of competences in lifelong learning.

There are two main standards for describing student learning experiences. The IEEE Learner Model working group has defined the Public and Private Information for Learners [11] as a standard for a student model, with the aim of gathering information related to competences, personal data, learning style, and so on. This standard considers six types of data related to *Personal, Relations, Security, Preference, Performance and Portfolio information*; in addition, it is possible to extend and integrate the standard in order to enrich the student description.

In 2005 the IMS consortium released the IMS ePortfolio Practice and Implementation Guide [12]. This specification uses the XML language to define the characteristics of an e-portfolio. XML is at the basis of the semantic web layer cake, so this specification constitutes the first step towards a semantic description of student competences. The use of specific ontologies can enrich this description by considering also the relationships between the concepts that are at the basis of e-portfolio systems.

For example, figure 1 shows the e-Portfolio *Activity* concept, its properties and its relations reported in our ontology.

An e-Portfolio can bring together different kinds of information such as: digital and non digital works; activities in which the student has participated, is participating, or plans to participate; competences and skills of the student; students achievements, whether or not certificated; student's preferences; student's goals and plans; student's interests and values; any notes, reflections or assessments relevant to any other part; the results of any test or examination taken by the student; contextual information to help the interpretation of any results.

5 Semantic framework for e-portfolio management

Many educational approaches are based on groupwork, since peer learning promotes cognitive processes. There are many different kinds of collaborative work that allow students to learn in different modalities, such as group discussions, group problem solving and group study. The form of collaboration differs according to the duration, the complexity and the level of collaboration.

Social interactions can help students to share their experiences and to work collaboratively on relevant topics. In this sense social networks occupy a key role in the learning dynamics. The number of informal learning activities which take place in technology supported social networks is constantly increasing. Collaborative group activities are frequently used by teachers in the educational curriculum. In

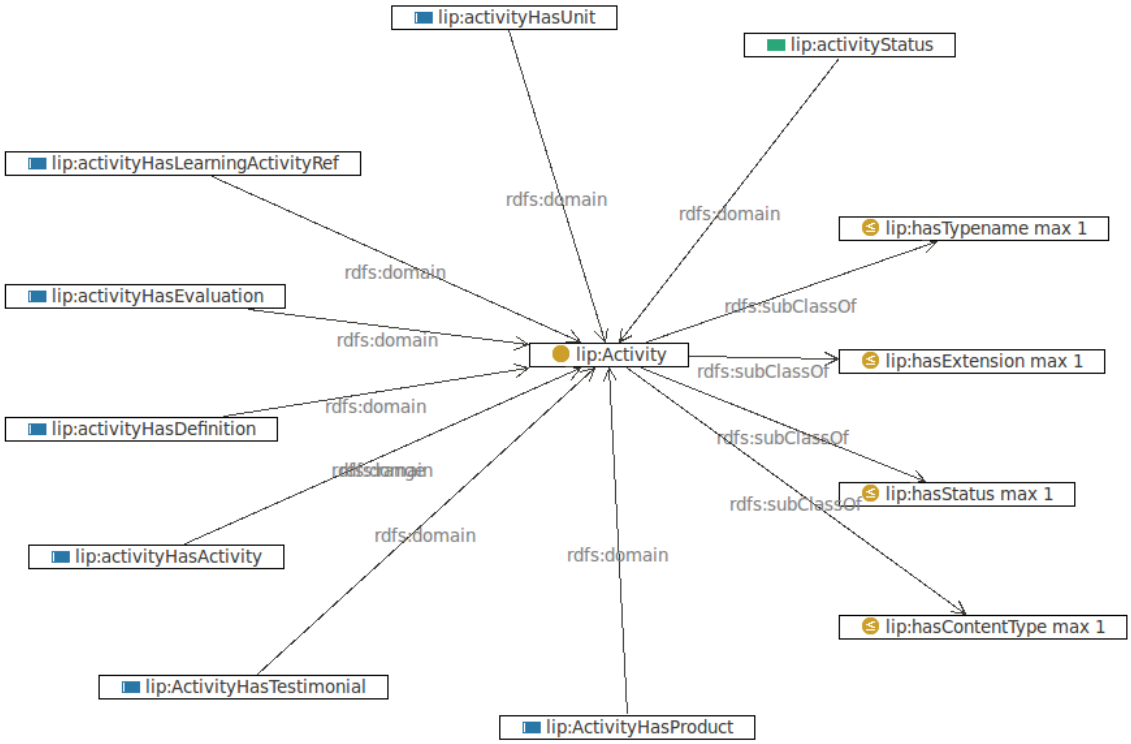


Figure 1: The activity Concept of the e-Portfolio Ontology

these activities it is necessary to create well balanced groups with the aim of maximizing the attainment of the learning objectives.

To ensure the success of a learning activity, teachers must consider the constraints that can affect the entire group or an individual performance, such as previous experiences by students in similar educational contexts, cultural background or interests and competences.

The importance of a system based on competences in informal learning environments such as those developed using social software is undeniable. For example, there are clear benefits in involving members with different levels of experience within the group in order to improve the dynamics of collaborative work in problem solving activities.

From this point of view it is increasingly important to have software applications that can store data related to the user profiles and process these data semantically.

The approach proposed in this paper consists in integrating and extending FOAF ontology, used for modeling contacts and personal relationship, with semantic data related to students competences.

Enriching the description of social networks using semantics can provide precious support for a more effective use of the network for educational purposes. Social learning experiences must consider the competences and the e-portfolio of the participants, so web semantic technologies are an essential substrate for merging models related to the social network description with models used to define and structure competences.

In particular, the use of ontologies and semantic web technologies makes it possible to relate the evolution of educational activities experienced by the students with their relationships.

The description of a students social network using FOAF, integrated with the definition of competences by means of the IMS model, is the basis for the creation of a competence based ontological system for virtual learning environments using social networks and web 2.0 technologies. The result is a learning environment which is no longer based on the transmission of information from teacher to student but rather is focused on the ability of the students to play an active role in their learning activities.

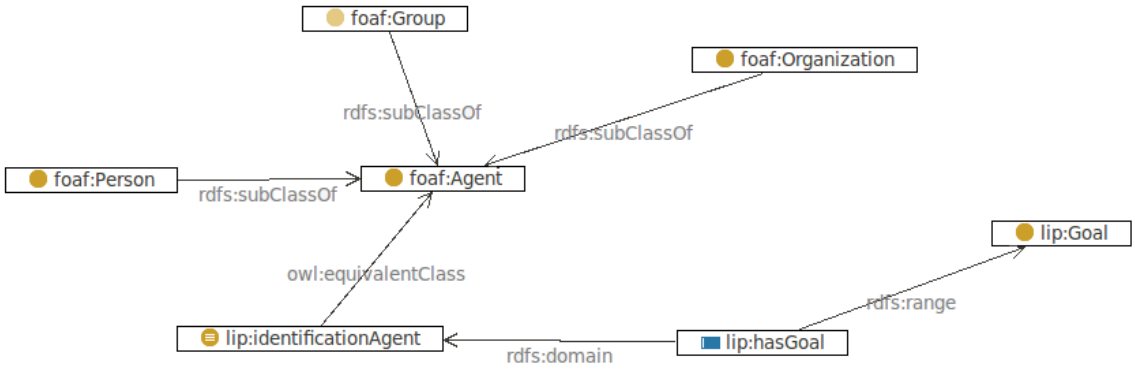


Figure 2: An example of the relation between FOAF and e-Portfolio in the proposed ontology

The figure 2 shows an example of how our ontology describes the relation between a concept of the e-Portfolio and a concept in FOAF.

As an example, we consider how our approach can be used to increase learning effectivness in informal learning activities that take place within an on-line social network environment. Using the ontology proposed in our work, it is possibile to describe both social relationships between students and the e-portfolio for each student. By social relationships, we mean the friendship links explicitly declared by students within the on-line social network software, while the e-portfolio allows us to describe the competences acquired by students, their learning goals and so on. All this information can be available in a software platform that uses our ontology to:

- create a sub-group from a student's friendship group, which includes the friends that have common learning interests and objectives
- suggest new friends to a student by selecting the people from within the on-line social network who have specific competences in their portfolios that can help the student to achieve his learning objectives

In a social network it is important not only to be connected to other people but to be connected to the right people, depending on your goals. This is true not only for business or work experiences but it is also important for informal learning activities that take place in an on-line social network. The results of learning activities are highly influenced by the group in which students participate. Several studies have been conducted analizyng the impact of positive interdependence on the effectiveness of cooperation.

As stated by Johnson and Johnson one of the essential elements for efficacy in cooperation is positive interdependence [13]. The authors affirm that positive interdependence is structured in three categories: outcome, boundary and means. Our approach provides helpful conditions to support the first two categories of positive interdependence. In particular, the outcome categories include goals and rewards; in our case we create a sub-group of friends in which learning objectives and competences coincide, thus facilitating a structuring positive outcome interdependence in order to increase achievement and productivity [14], [15], [16].

Johnson and Johnson also state that *"the boundary category includes: outside enemy (or negative interdependence with another group), identity (which binds members together as an entity), and environmental (such as a specific work area) interdependence"* [13]. From this point of view, creating the sub-group from the existing friendship network increases some factors of boundary interdependence.

In conclusion, our ontological approach is useful for supporting the creation of learning groups, promoting positive interdependence that has a positive influence to produce higher achievement and

productivity more then the membership in and of itself [17] and the interpersonal interaction itself [18], [19].

6 Conclusions

Collaborative group activities are frequently used by teachers in the educational curriculum. In these activities it is necessary to create well balanced groups with the aim of maximizing the attainment of the learning objectives.

To ensure the success of a learning activity, teachers must consider the constraints that can affect the entire group or an individual performance, such as previous experiences by students in similar educational contexts, cultural background or interests. The greater the number of constraints to consider, the more complex becomes the management of the learning experience.

Semantic web technologies offer the substrate needed to overcome the problems of social network with large groups of students. The versatility of these technologies means that they can be successfully applied for describing social networks and competences in learning experiences.

An interesting approach for creating an ontological system based on semantic web technologies that makes it possible to define a social network considering the competences of participants, the quality of the group and its robustness, is based on the use of an ontology as a result of an extension of the FOAF vocabulary, to create a semantic data base including specific references to educational paths.

In particular, the approach proposed in this work is based on the creation of a specifically designed ontology that extends FOAF ontology, in order to describe the domain of competences as defined by the IMS e-portoflio standard.

Bibliography

[1] M. Castells, *The Information Age, Economy, Society and Culture Volume I: The Rise of Network Society,* Blackwell, Oxford 1996.

[2] A. Andreatos, Virtual Communities and their Importance for Informal Learning *International Journal of Computers, Communications & Control*, 2(1):39-47, 2007.

[3] J. Lave and E. Wenger, *Situated Learning: Legitimate Peripheral Participation*, Cambridge University Press, 1991.

[4] L.S. Vigotsky, *Mind in Society*. Harvard University Press. Cambridge, 1978.

[5] V. Tinto, Classrooms as communities: Exploring the educational character of student persistence, *Journal of Higher Education*, 68(6):599-622, 1997.

[6] J.G. Breslin, *Social Semantic Information Spaces*. In S.R. Kruk and B. McDaniel, Semantic Digital Libraries. Springer, 2008.

[7] J. Jung, J. Euzenat, *Towards Semantic Social Networks*. In Proceedings of the 4th European Semantic Web Conference, Innsbruck, Austria, pp. 267-280, 2007.

[8] A. Ounnas, I. Liccardi, H.C. Davis, D.E. Millard and S.A. White *Towards a Semantic Modeling of Learners for Social Networks*. In Proceedings of the International Workshop on Applications of Semantic Web Technologies for E-Learning (SW-EL) at the AH2006 Conference, Dublin, Ireland, 2006.

[9] I. Liccardi, A. Ounnas, R. Pau, E. Massey, P. Kinnunen, S. Lewthwaite, M. Midy, C. Sarkar, The role of social networks in students' learning experiences. *SIGCSE Bulletin* 39(4):224-237, 2007.

[10] R. Mason, C. Pegler and M. Weller, *E-portfolios: An assessment tool for online courses*. British Journal of Educational Technology, 35,6, pp. 717-727, 2004.

[11] IEEE P1484.2.1/D8 : Draft Standard for Learning Technology - Public and Private Information (PAPI) for Learners (PAPI Learner) Core Features Sponsored by the Learning Technology Standards Committee of the IEEE Computer Society, 2001.

[12] IMS ePortfolio Practice and Implementation Guide, IMS Global Learning Consortium, 2005.

[13] D. W. Johnson, R. T. Johnson. An educational psychology success story: social interdependence theory and cooperative learning. *Educational Researcher*, 38(5), 365-379, 2009.

[14] M. Jensen, Cooperative quizzes in the anatomy and physiology laboratory: A description and evaluation. *Advances in Physiology Education*, 16(1), S48-S54, (1996).

[15] M. Jensen, D.W.Johnson, R. Johnson, Impact of positive interdependence during electronic quizzes on discourse and achievement. *Journal of Educational Research*, 95, 161-166, (2002).

[16] T. Matsui, T. Kakuyama, M. Onglatco, Effects of goals and feedback on performance in groups. *Journal of Applied Psychology*, 72, 407-415, (1987).

[17] N. Hwong, A. Caswell, D.W. Johnson, R. Johnson, Effects of cooperative and individualistic learning on prospective elementary teachers' music achievement and attitudes. Journal of Social Psychology, 133, 53-64, (1993).

[18] D. Mesch, M. Lew, D.W. Johnson, R. Johnson, Isolated teenagers, cooperative learning and the training of social skills. *Journal of Psychology*, 120, 323-334, (1986).

[19] D. Mesch, D.W. Johnson, R. Johnson, Impact of positive interdependence and academic group contingencies on achievement. *Journal of Social Psychology*, 128, 345-352, (1988).

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 4, pp. 586-591

Cryptanalysis on Two Certificateless Signature Schemes

F. Zhang, S. Li, S. Miao, Y. Mu, W. Susilo, X. Huang

Futai Zhang, Songqin Miao
1. School of Computer Science and technology
Nanjing Normal University,
Nanjing 210046, P.R. China, and
2. Jiangsu Engineering Research Center on Information Security and Privacy Protection Technology
Nanjing 210046, P.R. China
E-mail: zhangfutai@njnu.edu.cn, miaosongqin@163.com

Sujuan Li
1. School of Computer Science and technology
Nanjing Normal University,
Nanjing 210046, P.R. China, and
2. Nanjing University of Technology
Nanjing 210037, P.R. China
E-mail: lisujuan1978@126.com

Yi Mu, Willy Susilo, Xinyi Huang
Centre for Computer and Information Security Research
School of Computer Science and Software Engineering
University of Wollongong
NSW 2522, Australia
E-mail: ymu@uow.edu.au, wsusilo@uow.edu.au, xyhuang81@gmail.com

Abstract:
Certificateless cryptography has attracted a lot of attention from the research community, due to its applicability in information security. In this paper, we analyze two recently proposed certificateless signature schemes and point out their security flaws. In particular, we demonstrate universal forgeries against these schemes with known message attacks.

Keywords: certificateless cryptography, certificateless signature, public key replacement, universal forgery.

1 Introduction

Certificateless cryptography [1] is a new paradigm that not only removes the inherent key escrow problem of identity based public cryptography [2] (ID-PKC for short), but also eliminates the cumbersome certificate management in traditional PKI. In CL-PKC, the actual private key of a user is comprised of two secrets: a secret value and a partial private key. The user generates a secret value by himself, while the partial private key is generated by a third party called Key Generating Center (KGC), who makes use of a system wide master key and the user's identity information. In this way, the key escrow problem in identity-based public key cryptosystems is removed. A user's public key is derived from his/her actual private key, identity and system parameters. It could be available to other entities by transmitting along with signatures or by placing in a public directory. Unlike the traditional PKI, there is no certificate in certificateless public key cryptography to ensure the authenticity of the entity's public key. A number of certificateless signature schemes [3–14] have been proposed. Some of them are analysed under reasonable security models with elaborate security proofs [8, 11, 13, 14], while some others are subsequently broken due to flawed security proof or unreasonable model [3, 6–8, 12].

Recently two certificateless signature schemes were proposed in [4] and [5] respectively. They were claimed to provide high efficiency and provable security. In this short note, unfortunately, we show that these two schemes [4, 5] are insecure even in a very weak security model. Namely these two schemes are suffering from universal forgeries under known message attacks.

2 Review of the Original Schemes

We omit the preliminaries, basic notions, and security models about certificateless signature schemes. Please refer to [1, 8, 11, 13, 14] for details. The two original schemes [4, 5] are based on bilinear maps. They were both called McCLS scheme. To distinguish them, we call the one in [4] as McCLS1, and the other one in [5] as McCLS2.

2.1 Description of McCLS1

We first describe McCLS1. It consists of the following five algorithms.

- **Setup.** On input a security parameter, it generates a list of system parameters $\{ p, G_1, G_2, \hat{e}, P, P_{pub}, H_1, H_2\}$ and a system master private key $s \in Z_p^*$, where p is a large prime, G_1, G_2 are groups of order p with an admissible bilinear map $\hat{e}: G_1 \times G_1 \rightarrow G_2$, $H_1 : \{0,1\}^* \rightarrow G_1$ and $H_2 : \{0,1\}^* \rightarrow Z_p^*$ are cryptographical Hash functions, P is a generator of G_1, and $P_{pub} = sP$.

- **Extract Partial Private Key.** On input a user identity ID, it computes $Q_{ID} = H_1(ID)$, and outputs $D_{ID} = sQ_{ID}$ as the user's partial private key.

- **Generate Key Pair.** A user with identity ID selects a random $x \in Z_p^*$ as its secret value S_{ID}, and publish its public key $P_{ID} = xP_{pub}$.

- **CL-Sign.** Given a user's private keys (D_{ID}, S_{ID}) and a message M, the user randomly picks an element $r \in Z_p^*$, computes $S = {S_{ID}}^{-1} D_{ID}$, $R = (r - S_{ID})P$, $V = H_2(M, R, P_{ID})r$, and outputs $\sigma = (S, R, V)$ as his/her signature on message M under the public key P_{ID}.

- **CL-Verify.** Given a signature (S, R, V) on a message M of a user ID with public key P_{ID}, a verifier computes $h = H_2(M, R, P_{ID})$ and checks whether $(P_{pub}, VP - hR, h^{-1}S, Q_{ID})$ is a valid Diffie-Hellman tuple, namely whether the equation $\hat{e}(VP - hR, h^{-1}S) = \hat{e}(P_{pub}, Q_{ID})$ holds.

2.2 Description of McCLS2

The first three algorithms of McCLS2 in [5] are exactly the same as those of McCLS1 in [4]. There are slight differences in the **CL-Sign** and **CL-Verify** algorithms. We just depict the differences here.

- **CL-Sign.** Given a user's private keys (D_{ID}, S_{ID}) and a message M, the user randomly picks an element $r \in Z_p^*$, computes $S = {S_{ID}}^{-1} D_{ID}$, $R = (r - S_{ID})P$, $V = H_2(M, R, P_{ID})rP$ and outputs $\sigma = (S, R, V)$ as his/her signature on message M under public key P_{ID}.

- **CL-Verify.** Given a signature (S, R, V) on a message M of a user ID with public key P_{ID}, a verifier computes $h = H_2(M, R, P_{ID})$ and checks whether $(P_{pub}, V - hR, h^{-1}S, Q_{ID})$ is a valid Diffie-Hellman tuple, namely whether the equation $\hat{e}(V - hR, h^{-1}S) = \hat{e}(P_{pub}, Q_{ID})$ holds.

3 Universal forgery

As we can see, in the McCLS schemes, a signature on a message M of a user ID with public key P_{ID} consists of three components S, R and V. Note that for a user ID with public key P_{ID}, S remains unchanged for all messages, R and V are irrelevant to the partial private key D_{ID}. Here we give two kinds of universal forgery under known message attacks.

3.1 Attacks Against McCLS1

1. **Universal forgery by replacing public key**

   The scheme McCLS1 cannot resist public key replacement attacks of a type I adversary $\mathcal{A}$. For the definition of type I and type II adversaries, please refer to [1, 4, 5, 8, 11, 13, 14]. Let $\sigma = (S,R,V)$ be ID's valid signature on a message M, where

   $$S = {S_{ID}}^{-1} D_{ID},\ R = (r - S_{ID})P,\ V = H_2(M,R,P_{ID})r, \text{ and } r \in_R Z_p^*.$$

   Given R and V, the random number r can be easily derived as $r = VH_2(M,R,P_{ID})^{-1}$. And then $S_{ID}P$ is known as $S_{ID}P = rP - R$. Now $\mathcal{A}$ is able to forge a user ID's valid signature on any message m as follows:

   (a) Choose a random $c \in Z_p^*$ and let $r' = cr \in Z_p^*$;

   (b) Replace ID's public key as $P'_{ID} = cP_{ID}$ (the new secret value corresponding to the public key P'_{ID} is $S'_{ID} = cS_{ID}$);

   (c) Compute $S' = c^{-1}S, R' = cR, V' = H_2(m,R',P'_{ID})r'$;

   (d) Set $\sigma' = (S',R',V')$ as ID's signature on message m under the public key P'_{ID}. We can see that $(P_{pub}, V'P - H_2(m,R',P'_{ID})R', H_2(m,R',P'_{ID})^{-1}S', Q_{ID})$ is a valid Diffie-Hellman tuple since

   $$\begin{aligned}
   & \hat{e}(V'P - H_2(m,R',P'_{ID})R', H_2(m,R',P'_{ID})^{-1}S') \\
   = \ & \hat{e}((H_2(m,R',P'_{ID})crP) - H_2(m,R',P'_{ID})(crP - cS_{ID}P)), H_2(m,R',P'_{ID})^{-1}c^{-1}S) \\
   = \ & \hat{e}(H_2(m,R',P'_{ID})cS_{ID}P, H_2(m,R',P'_{ID})^{-1}c^{-1}S) \\
   = \ & \hat{e}(S_{ID}P, S) \\
   = \ & \hat{e}(P, D_{ID}) \\
   = \ & \hat{e}(P_{pub}, Q_{ID})
   \end{aligned}$$

2. **Universal forgery without replacing public key**

   From ID's valid signature $\sigma = (S,R,V)$ on a message M, the adversary can get

   $$r = VH_2(M,R,P_{ID})^{-1}, S_{ID}P = rP - R.$$

   With these he can forage a signature $\sigma' = (S',R',V')$ on any message m without replacing ID's public key as follows:

   Pick $r' \in_R Z_p^*$, and compute $S' = S$, $R' = r'P - s_{ID}P$, $V' = H_2(m,R',P_{ID})r'$.

   The verification will always output "accept" since

   $$(P_{pub}, V'P - H_2(m,R',P_{ID})R', H_2(m,R',P_{ID})^{-1}S', Q_{ID})$$

is really a valid Diffie-Hellman tuple. The reason is

$$\begin{aligned} & \hat{e}(V'P - H_2(m,R',P_{ID})R', H_2(m,R',P_{ID})^{-1}S') \\ = \; & \hat{e}(H_2(m,R',P_{ID})r'P - H_2(m,R',P_{ID})(r'P - S_{ID}P), H_2(m,R',P_{ID})^{-1}S) \\ = \; & \hat{e}(H_2(m,R',P_{ID})S_{ID}P, H_2(m,R',P_{ID})^{-1}S) \\ = \; & \hat{e}(S_{ID}P,S) \\ = \; & \hat{e}(P_{pub},Q_{ID}) \end{aligned}$$

3.2 Attacks Against McCLS2

1. **Universal forgery by replacing public key**

   Let $\sigma = (S,R,V)$ be ID's valid signature on a message M. It is obvious that

   $$rP = H_2(M,P,P_{ID})^{-1}V,\ S_{ID}P = rp - R.$$

   A type I adversary $\mathcal{A}$ may forge ID's valid signature on any message m as follows:

   (a) Choose a random $c \in Z_p^*$ and let $r' = cr \in Z_p^*$.

   (b) Replace ID's public key as $P'_{ID} = cP_{ID}$ (this implies the new secret value corresponding to the new public key P'_{ID} is $S'_{ID} = cS_{ID}$).

   (c) Compute
   $S' = c^{-1}S, R' = (r' - S'_{ID})P = cR, V' = H_2(m,R',P'_{ID})r'P = cH_2(m,R',P'_{ID})rP.$

   (d) Set $\sigma' = (S',R',V')$ as ID's signature on message m using public key P'_{ID}.
   We can see $(P_{pub}, V' - H_2(m,R',P'_{ID})R', H_2(m,R',P'_{ID})^{-1}S', Q_{ID})$ is a valid Diffie-Hellman tuple since

   $$\begin{aligned} & \hat{e}(V' - H_2(m,R',P'_{ID})R', H_2(m,R',P'_{ID})^{-1}S') \\ = \; & \hat{e}(H_2(m,R',P'_{ID})crP - H_2(m,R',P'_{ID})(crP - cS_{ID}P), H_2(m,R',P'_{ID})^{-1}c^{-1}S) \\ = \; & \hat{e}(H_2(m,R',P'_{ID})cS_{ID}P, H_2(m,R',P'_{ID})^{-1}c^{-1}S) \\ = \; & \hat{e}(S_{ID}P,S) \\ = \; & \hat{e}(P,D_{ID}) \\ = \; & \hat{e}(P_{pub},Q_{ID}) \end{aligned}$$

2. **Universal forgery without replacing public key**

   The adversary can get

   $$rP = H_2(M,R,P_{ID})^{-1}V,\ S_{ID}P = rP - R = H_2(M,R,P_{ID})^{-1}V - R,$$

   from ID's valid signature $\sigma = (S,R,V)$ on a message M. Then it (may be type I or type II) can forge a signature $\sigma' = (S',R',V')$ on any message m without replacing ID's public key as follows:

   Pick $r' \in_R Z_p^*$, and compute $S' = S$, $R' = r'P - s_{ID}P$, $V' = H_2(m,R',P_{ID})r'P$.

   The verification will always output "accept" since

   $$(P_{pub}, V' - H_2(m,R',P_{ID})R', H_2(m,R',P_{ID})^{-1}S', Q_{ID})$$

is really a valid Diffie-Hellman tuple. This is because

$$\begin{aligned} & \hat{e}(V' - H_2(m,R',P_{ID})R', H_2(m,R',P_{ID})^{-1}S') \\ = \; & \hat{e}(H_2(m,R',P_{ID})r'P - H_2(m,R',P_{ID})(r'P - S_{ID}P), H_2(m,R',P_{ID})^{-1}S) \\ = \; & \hat{e}(H_2(m,R',P_{ID})S_{ID}P, H_2(m,R',P_{ID})^{-1}S) \\ = \; & \hat{e}(S_{ID}P, S) \\ = \; & \hat{e}(P_{pub}, Q_{ID}) \end{aligned}$$

From these attacks, one can see McCLS1 and McCLS2 are insecure even in the weakest security model.

4 Conclusion

Recently, two certificateless signature schemes McCLS1 and McCLS2 were proposed for Mobile Wireless Cyber-Physical Systems. They only require two scalar multiplications in signing phase and two scalar multiplications and one pairing in verification phase. So they are efficient with respect to computational cost. Although the authors claimed and proved that McCLS1 and McCLS2 were secure, as we have shown in this paper they are in fact insecure. Universal forgeries against those two schemes have been presented under known message attacks.

5 Acknowledgment

This research is supported by the Natural Science Foundation of China under grant number 60673070 and Academic Discipline Fund of NJUT.

Bibliography

[1] S. Al-Riyami, K. Paterson. Certificateless public key cryptography. *Proceedings of Asiacrypt 2003*, Lecture Notes in Computer Science 2894, Springer-Verlag, 452-473, 2003.

[2] A. Shamir. Identity based cryptosystems and signature schemes. *Proceedings of Crypto'84*, 47-53, 1984.

[3] X. Huang, W. Susilo, Y. Mu, F. Zhang. On the security of a certificateless signature scheme. *Proceedings of ACISP 2005*, 13-25, 2005.

[4] Z. Xu, X. Liu, G. Zhang, W. He, G. Dai, W. Shu. A Certificateless Signature Scheme for MobileWireless Cyber-Physical Systems. *The 28th International Conference on Distributed Computing Systems Workshops*, 489-494, 2009.

[5] Z. Xu, X. Liu, G. Zhang, W. He. McCLS: Certificateless Signature Scheme for Emergency Mobile Wireless Cyber-Physical Systems. *International Journal of Computers, Communications & Control (IJCCC)*, 3(4): 395-411, 2008.

[6] W. Yap, S. Heng, B. Goi1. An efficient certificateless signature scheme. *Proceedings of EUC Workshops 2006*, Lecture Notes in Computer Science 4097, Springer-Verlag, 322-331, 2006.

[7] J. Park. An attack on the certificateless signature scheme from EUC Workshops 2006. *Cryptology ePrint Archive*, Report 442, 2006.

[8] Z. Zhang, D. Feng. Key replacement attack on a certificateless signature scheme. *Cryptology ePrint Archive*, Report 453, 2006.

[9] K. Choi, J. Park, J. Hwang, D. Lee. Efficient certificateless signature schemes. *Proceedings of ACNS 2007*, Lecture Notes in Computer Science 4521, Springer-Verlag, 443-458, 2007.

[10] R. Castro, R. Dahab. Two notes on the security of certificateless signatures. *Proceedings of ProvSec 2007*, Lecture Notes in Computer Science 4784, Springer-Verlag, 85-102, 2007.

[11] Z. Zhang, D. Wong, J. Xu, D. Feng. Certificateless public-key signature: security model and efficient construction. *Proceedings of ACNS 2006*, Lecture Notes in Computer Science 3989, Springer-Verlag, 293-308, 2006.

[12] B. Hu, D. Wong, Z. Zhang, X. Deng. Key replacement attack against a generic construction of certificateless signature. *Proceedings of ACISP 2006*, Lecture Notes in Computer Science 4058, Springer-Verlag, 235-346, 2006.

[13] X. Huang, Y. Mu, W. Susilo, D. Wong, W. Wu. Certificateless signature revisited. *Proceedings of ACISP 2007*, Lecture Notes in Computer Science 4586, Springer-Verlag, 308-322, 2007.

[14] L. Zhang, F. Zhang, F. Zhang. New efficient certificateless signature scheme. *Proceedings of EUC Workshops 2007*, Lecture Notes in Computer Science 4809, Springer-Verlag, 692-703, 2007.

Futai Zhang (b. Augest 28, 1965) received his M.Sc. in Mathematics (1990) from Shaanxi Normal University, China, and PhD in Cryptology (2001) from Xidian University, China. Now he is a professor at Nanjing Normal University, China. His current research interest is public key cryptography.

Sujuan Li is now a PhD candidate in Nanjing Normal University, China. Her current research interests include information security and cryptography.

Songqin Miao is now a M.Sc candidate in Nanjing Normal University, China. Her main research interest is cryptography.

Yi Mu received his PhD from the Australian National University in 1994. He is an associate Professor at the School of Computer Science and Software Engineering at the University of Wollongong. He is the co-director of Centre for Computer and Information Security Research (CCISR) at the University of Wollongong. His current research interests include network security, electronic payment, cryptography, access control, and computer security.

Willy Susilo received a Ph.D. in Computer Science from University of Wollongong, Australia. He is a Professor at the School of Computer Science and Software Engineering at the University of Wollongong. He is the co-director of Centre for Computer and Information Security Research (CCISR) at the University of Wollongong.His current research interests include information security and cryptography.

Xinyi Huang received his Ph.D. in Computer Science from University of Wollongong, Australia in 2009. His current research mainly focuses on cryptography and its applications.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 4, pp. 592-602

H_∞ Robust T-S Fuzzy Design for Uncertain Nonlinear Systems with State Delays Based on Sliding Mode Control

X.Z. Zhang, Y.N. Wang, X.F. Yuan

Xizheng Zhang
Hunan University
P.R.China, 410082 Changsha, Yuelushan, and
Hunan Institute of Engineering
P.R.China, 411104 Xiangtan, 88 Fuxing East Road
E-mail: z_x_z2000@163.com

Yaonan Wang, Xiaofang Yuan
Hunan University
P.R.China, 410082 Changsha, Yuelushan
E-mail: yaonan@hnu.cn, yuanxiaof@21cn.com

Abstract: This paper presents the fuzzy design of sliding mode control (SMC) for nonlinear systems with state delay, which can be represented by a Takagi-Sugeno (T-S) model with uncertainties. There exist the parameter uncertainties in both the state and input matrices, as well as the unmatched external disturbance. The key feature of this work is the integration of SMC method with H_∞ technique such that the robust asymptotically stability with a prescribed disturbance attenuation level γ can be achieved. A sufficient condition for the existence of the desired SMC is obtained by solving a set of linear matrix inequalities (LMIs). The reachability of the specified switching surface is proven. Simulation results show the validity of the proposed method.
Keywords: sliding mode control, T-S fuzzy model, time-delayed system, H_∞ control.

1 Introduction

Recently, the dynamic T-S fuzzy model has become a popular tool and has been employed in most model-based fuzzy analysis approaches [1]. Moreover, the ordinary T-S fuzzy model has been further extended to deal with nonlinear uncertain systems with time-delays [2]. The stability analysis and stabilization controller design for fuzzy time-delayed systems have attracted much attention over the past few decades due to their extensive applications in mechanical systems, economics, and other areas. A large number of results on this topic have been reported in the literature, see, e.g. [3–5]. Note that the uncertainties may exist in the real systems, or come from the fuzzy modeling procedure. Hence, the robust stabilization problems have recently been investigated in [6] for nonlinear uncertain fuzzy systems.

In practice, the inevitable uncertainties may enter a nonlinear system in a much more complex way. The uncertainty may include modeling error, parameter perturbations, fuzzy approximation errors, and external disturbances. In such circumstances, especially in the existence of external disturbances, the above established methods to control fuzzy time-delay systems could not work well any more. However, it is well known that the sliding mode control (SMC) is a reasonable approach to take effect if the lumped uncertainties are known to be bounded by smooth functions. In a more detail, the SMC system could drive the trajectories onto the so-called switching surface in a finite time and maintain on it thereafter, and on the switching surface the system is insensitive to internal parameter perturbations and external disturbances [7]. SMC approach has been successfully adopted in the control of time-delay systems these years. Quite recently, SMC approach has been also applied to solve the stabilization and tracking problems for fuzzy systems with matched uncertainties [8]. However, the sliding motion cannot be detached

from the effect of unmatched parameter uncertainties, especially, unmatched external disturbances [9]. This means that the unmatched external disturbances make the design of SMC complex and challenging.

On the other hand, the H_∞ control, in the past decades, has been widely employed to deal with the uncertain systems with external disturbance [10, 11]. The goal of this problem is to design a controller to stabilize a given system while satisfying a prescribed level of disturbance attenuation. The H_∞ control for uncertain time-delayed systems has been considered by some researchers [12, 13]. In [14], Peng and Yue investigated the H_∞ controller design for uncertain T-S fuzzy systems with time-varying interval delay by using a new Lyapunov-Krasovskii functionals and an innovative integral inequality. The output feedback controller in [15] was designed for uncertain fuzzy systems such that the closed-loop systems are robustly asymptotically stable and satisfy a prescribed H_∞ performance. Lin et al. [16] further presented the mixed H_2/H_∞ filter design for nonlinear discrete-time systems with state-dependent noise.

Motivated by the above discussion, it is certain that the integration of the SMC method with H_∞ technique would have a great potential in extending the SMC to the systems with unmatched uncertainties and obtain a better dynamic performance. Therefore, in this paper, by utilizing the H_∞ technique to attenuate the effect of unmatched external disturbance, we proposes a novel SMC controller that can ensure the robust stability with a prescribed disturbance attenuation level γ for the fuzzy time-delayed system, irrespective of parameter uncertainties and unmatched external disturbance. The controller design method is presented in terms of LMIs.

The notations used in this paper are quite standard: $\Re^n$ denotes the n-dimensional real Euclidean space; $\boldsymbol{I}$ is the identity matrix with appropriate dimensions; $\boldsymbol{W} < 0$ ($\boldsymbol{W} > 0$) means that $\boldsymbol{W}$ is symmetric and negative (positive) definite; $L_2[0,\infty]$denotes the space of square-integrable vector functions over $[0,\infty]$; The superscript "T" represents the transpose of a matrix, and the notation "$*$" is used as an ellipsis for terms that are induced by symmetry; $\|\cdot\|$ denotes the spectral norm; Matrices, if they are not explicitly stated, are assumed to have compatible dimensions.

2 Problem Formulation

As stated in Introduction, T-S fuzzy models can provide an effective representation of complex nonlinear systems in terms of fuzzy sets and fuzzy reasoning applied to a set of linear input-output submodels. Hence, in this work, a class of nonlinear time-delay systems is represented by a T-S model. As in [2], the T-S fuzzy time-delay system with uncertainties is described by fuzzy IF-THEN rules, which locally represent linear input-output relations of nonlinear systems. The i-th rule of the fuzzy model is formulated in the following equation:

Plant rule i: IF θ_1 is η_1^i and θ_2 is $\eta_2^i \cdots$ and θ_p is η_1^p, THEN

$$\begin{cases} \dot{\boldsymbol{x}}(t) = [(\boldsymbol{A}_i + \Delta\boldsymbol{A}_i(t))\boldsymbol{x} + (\boldsymbol{A}_{di} + \Delta\boldsymbol{A}_{di}(t))\boldsymbol{x}(t-\tau)] + (\boldsymbol{B}_i + \Delta\boldsymbol{B}_i)\boldsymbol{u}(t) + \boldsymbol{B}_{wi}\boldsymbol{w}(t) \\ \boldsymbol{z}(t) = \boldsymbol{C}_i\boldsymbol{x}(t) \\ \boldsymbol{x}(t) = \boldsymbol{\varphi}(t), t \in [-\tau(t), 0], i = 1, 2, ., r \end{cases} \tag{1}$$

where η_j^i is the fuzzy set, $\boldsymbol{\theta} = [\theta_1(t), \theta_2(t), \cdots, \theta_p(t)]^T$ is the premise variable vector, r is the number of rules of this T-S fuzzy. $\boldsymbol{x}(t) \in \Re^n$ is the state vector, $\boldsymbol{u}(t) \in \Re^m$ is the control input vector, $\boldsymbol{z}(t) \in \Re^l$ is the controlled output, $\boldsymbol{w}_i(t) \in \Re^p$ denotes the unknown external disturbances or modeling error. $\boldsymbol{A}_i$, $\boldsymbol{A}_{di}$, $\boldsymbol{B}_i$, $\boldsymbol{B}_{wi}$ and $\boldsymbol{C}_i$ are known real constant matrices with appropriate dimensions. $\Delta\boldsymbol{A}_i(t)$, $\Delta\boldsymbol{A}_{di}(t)$, $\Delta\boldsymbol{B}_i$ are unknown time-varying matrices representing parameter uncertainties. τ is the time-varying delay for the state vector satisfying $0 < \tau(t) < d < \infty$, $\dot{\tau(t)} < h < 1$, where d and h are known real constant scalars. $\varphi(t)$ is a continuous vector-valued initial function.

The overall fuzzy model achieved by fuzzy blending of each plant rule is represented as follows:

$$\dot{\boldsymbol{x}}(t) = \sum_{i=1}^{r} h_i(\boldsymbol{\theta})[(\boldsymbol{A}_i + \Delta\boldsymbol{A}_i)\boldsymbol{x}(t) + (\boldsymbol{A}_{di} + \Delta\boldsymbol{A}_{di})\boldsymbol{x}(t-\tau) + (\boldsymbol{B}_i + \Delta\boldsymbol{B}_i)\boldsymbol{u}(t) + \boldsymbol{B}_{wi}\boldsymbol{w}(t)] \tag{2}$$

where $h_i(\boldsymbol{\theta}) = \frac{\alpha_i(\boldsymbol{\theta})}{\sum_{j=1}^{r}\alpha_i(\boldsymbol{\theta})}$,$\alpha_i(\boldsymbol{\theta}) = \prod_{j=1}^{p}\eta_j^i(\boldsymbol{\theta})$, in which $\eta_j^i(\boldsymbol{\theta})$ is the membership grade of θ_j in η_j^i. According to the theory of fuzzy sets, we have $\boldsymbol{\theta} \geq 0$ and $\sum_{i=1}^{r}\boldsymbol{\theta} \geq 0$. Therefore, it implies that $h_i(\boldsymbol{\theta}) \geq 0$ and $\sum_{i=1}^{r} h_i(\boldsymbol{\theta}) = 1$. In this work, the following assumptions are introduced.

(Assumption.1) The time-varying uncertainties $\Delta\boldsymbol{A}_i$ and $\Delta\boldsymbol{A}_{di}$ are assumed to be norm-bounded, that is,

$$[\Delta\boldsymbol{A}_i, \Delta\boldsymbol{A}_{di}] = \boldsymbol{H}_i\boldsymbol{F}_i(t)[\boldsymbol{E}_{1i}, \boldsymbol{E}_{2i}] \tag{3}$$

where $\boldsymbol{H}_i$,$\boldsymbol{E}_{1i}$, $\boldsymbol{E}_{2i}$ are known constant matrices, and $\boldsymbol{F}_i(t)$ is an unknown matrix function with Lebesgue-measurable elements and satisfies $\boldsymbol{F}_i^T(t)\boldsymbol{F}_i(t) \leq \boldsymbol{I}$,$\forall t$.

(Assumption.2) It is assumed that the matrices $\boldsymbol{B}_i$ satisfy $\boldsymbol{B}_1 = \boldsymbol{B}_1 = \cdots = \boldsymbol{B}_r = \boldsymbol{B}$. Moreover, the pair $(\boldsymbol{A}_i, \boldsymbol{B})$ is controllable and the input matrix $\boldsymbol{B}$ has full-column rank m and $m < n$.

(Assumption.3) The uncertainty matrix $\Delta\boldsymbol{B}_i$ is assumed to be matched, i.e., there exists a matrix $\boldsymbol{\delta}_i(t) \in \Re^{m\times m}$ such that $\Delta\boldsymbol{B}_i = \boldsymbol{B}_i\boldsymbol{\delta}_i(t)$ with $\|\boldsymbol{\delta}_i(t)\| \leq \rho_B < 1$, where ρ_B is a positive constant.

(Assumption.4) The upper bound for $\boldsymbol{w}_i(t)$ is known.

It is noted that there exists parameter uncertainties in both the state and control input matrices and unmatched external disturbance $\boldsymbol{w}_i(t)$ in the systems under consideration.

Remark 1: Assumptions 1 4 are standard assumptions in the study of variable structure control.

Before proceeding, some standard concepts and lemma are given as follows, which are useful for the development of our result.

Definition 1. The uncertain fuzzy time-delayed systems in (2) is said to be robustly asymptotically stable if the system with $\boldsymbol{u}(t) = 0$ and $\boldsymbol{w}_i(t) = 0$ is asymptotically stable for all admissible parameter uncertainties.

Definition 2. Given a scalar $\gamma > 0$, the unforced fuzzy system in (2) with $\boldsymbol{u}(t) = 0$ is said to be robustly stable with disturbance attenuation γ if it is robustly stable and and under zero initial condition, $\|\boldsymbol{z}(t)\|_{E2} \leq \gamma\|\boldsymbol{w}(t)\|_2$ for all non-zero and all admissible uncertainties, where

$$\|\boldsymbol{z}(t)\|_{E2} = \sqrt{\int_0^t |\boldsymbol{z}(t)|^2 dt} \tag{4}$$

(Lemma.1 Choi [10]): Let $\boldsymbol{E}$, $\boldsymbol{H}$, and $\boldsymbol{F}(\boldsymbol{t})$ be real matrices of appropriate dimensions with $\boldsymbol{F}(\boldsymbol{t})$ satisfying $\boldsymbol{F}_i^T(t)\boldsymbol{F}_i(t) \leq \boldsymbol{I}$. Then, we have

(i) For any scalar $\varepsilon \leq 0$, $\boldsymbol{E}\boldsymbol{F}(t)\boldsymbol{H} + \boldsymbol{H}^T\boldsymbol{F}^T(t)\boldsymbol{E}^T \leq \varepsilon^{-1}\boldsymbol{E}\boldsymbol{E}^T + \varepsilon\boldsymbol{H}^T\boldsymbol{H}$

(ii) For any matrix $P > 0$, $-2\boldsymbol{E}^T\boldsymbol{H} \leq \boldsymbol{E}^T\boldsymbol{P}\boldsymbol{E} + \boldsymbol{H}^T P^{-1}\boldsymbol{H}$.

3 Controller design

The objective of this work is to design a SMC law such that the desired control performance for the resulting closed-loop system is obtained despite of parameter uncertainties and unmatched external disturbance. In this section, a SMC law is first synthesized such that the closed-loop systems are robustly asymptotically stable with disturbance attenuation γ. It is further proven that the reachability of the specified switching (sliding) surface $s(t) = 0$ can be ensured by the proposed SMC law. Thus, it is concluded that the synthesized SMC law can guarantee the state trajectories of uncertain systems (2) to be driven onto the sliding surface, and asymptotically tend to zero along the specified sliding surface.

3.1 Sliding mode controller design

Essentially, a SMC design is composed of two phases: hyperplane design and controller design. There are various methods for designing hyperplane, however in this paper the switching surface is

defined as

$$s(t) = \sum_{i=1}^{r} h_i(\boldsymbol{\theta}(t))\boldsymbol{G}_i\boldsymbol{x}(t) \tag{5}$$

where $\boldsymbol{G}_i \in \Re^{m\times n}$ is designed so that $\boldsymbol{G}_i\boldsymbol{B}_i$ is not singular. Furthermore, we design the VSC control law as follows

$$\begin{cases} \boldsymbol{u}(t) = \boldsymbol{u}_s(t) + \boldsymbol{u}_r(t) \\ \boldsymbol{u}_s(t) = -\sum_{i=1}^{r} \boldsymbol{K}_i\boldsymbol{x}(t) \\ \boldsymbol{u}_r(t) = -\sum_{i=1}^{r} h_i(\boldsymbol{\theta}(t))\boldsymbol{G}_i\left[\boldsymbol{A}_i\boldsymbol{x}(t) + \boldsymbol{A}_{di}\boldsymbol{x}(t-\tau(t))\right] - \sum_{i=1}^{r} h_i(\boldsymbol{\theta}(t))\rho_i(\boldsymbol{x},t)sgn(s(t)) \end{cases} \tag{6}$$

where $\boldsymbol{K}_i \in \Re^{m\times n}$ is chosen such that $(\boldsymbol{A}_i - \boldsymbol{B}_i\boldsymbol{K}_i)$ is Hurwitz, $sgn()$is a sign function and $\rho_i(\boldsymbol{x},t)$is a positive scalar function given as

$$\begin{aligned} \rho_i(\boldsymbol{x},t) \quad \geq \quad & \frac{2}{1-\rho_B^2}\Big\{ \left[\|\Phi(\boldsymbol{A}_i - \boldsymbol{B}_i\boldsymbol{K}_i)\| + \|\Phi\boldsymbol{H}_i\boldsymbol{E}_{1i}\| + \rho_B\|\boldsymbol{K}_i\| + (1+\rho_B)\|\boldsymbol{G}\boldsymbol{A}_i\|\right]\|\boldsymbol{x}(t)\| \\ & + \left[\|\Phi\boldsymbol{A}_{di}\| + \|\Phi\boldsymbol{H}_i\boldsymbol{E}_{2i}\| + (1+\rho_B)\|\boldsymbol{G}\boldsymbol{A}_{di}\|\right]\|\boldsymbol{x}(t-\tau)\| \\ & + \|s\|\|\Phi\boldsymbol{B}_{wi}\|\|\boldsymbol{w}\| + \beta \Big\} \end{aligned} \tag{7}$$

with $\Phi = (\boldsymbol{G}_i\boldsymbol{B}_i)^{-1}\boldsymbol{G}_i$ and $\beta > 0$ is a small known scalar.

Thus, substituting (6) into (2), we obtain the closed-loop system as follows

$$\begin{aligned} \dot{\boldsymbol{x}}(t) \quad = \quad & \sum_{i=1}^{r} h_i(\boldsymbol{\theta})\Big\{ \left[\boldsymbol{A}_i - \boldsymbol{B}_i\boldsymbol{K}_i + \Delta\boldsymbol{A}_i(t) - \Delta\boldsymbol{B}_i(t)\boldsymbol{K}_i\right]\boldsymbol{x}(t) + \left[\boldsymbol{A}_{di} + \Delta\boldsymbol{A}_{di}(t)\right]\boldsymbol{x}(t-\tau(t)) \\ & + \left[\boldsymbol{B}_i + \Delta\boldsymbol{B}_i(t)\right]\boldsymbol{u}_r(t) + \boldsymbol{B}_{wi}\boldsymbol{w}(t) \Big\} \end{aligned} \tag{8}$$

The above expression Eq.(8) is the sliding-mode dynamics of the fuzzy uncertain system (2) in the specified sliding surface $s(t) = 0$.

3.2 Stability of the sliding mode motion

In this subsection, we analyze the dynamic performance of the closed-loop system described by (8), and derives some sufficient conditions for the asymptotically stability of the sliding dynamics via LMI method. The following theorem shows that system (2) in the defined switching surface is robustly stabilizable with disturbance attenuation level γ.

Theorem 3. *Consider the fuzzy uncertain systems (2) with Assumptions 1 4, with the prescribed switching function, if there exist matrices $\boldsymbol{P} > 0$, $\boldsymbol{Q} > 0$, and positive scalars ε_1, ε_2 and ε_3 such that the LMI shown in (11) holds, with*

$$\Theta_1 = \boldsymbol{P}(\boldsymbol{A}_i - \boldsymbol{B}_i\boldsymbol{K}_i) + (\boldsymbol{A}_i - \boldsymbol{B}_i\boldsymbol{K}_i)^T\boldsymbol{P} + \boldsymbol{Q} + \varepsilon_1\boldsymbol{E}_{1i}^T\boldsymbol{E}_{1i} + \varepsilon_3\rho_B^2\boldsymbol{K}_i^T\boldsymbol{K}_i + \boldsymbol{C}_i^T\boldsymbol{C}_i \tag{9}$$

$$\Theta_2 = -(1-h)\boldsymbol{Q} + \varepsilon_2\boldsymbol{E}_{2i}^T\boldsymbol{E}_{2i} \tag{10}$$

for $i = 1,2,\cdots,r$, then, by choosing $\boldsymbol{G}_i = \boldsymbol{B}_i^T\boldsymbol{P}$, the sliding-mode dynamics (8) is robust asymptotically stable with disturbance attenuation γ.

$$\begin{pmatrix} \Theta_1 & * & * & * & * & * \\ \boldsymbol{A}_{di}^T\boldsymbol{P} & \Theta_2 & * & * & * & * \\ \boldsymbol{B}_{wi}^T\boldsymbol{P} & 0 & -\gamma^2\boldsymbol{I} & * & * & * \\ \boldsymbol{H}_i^T\boldsymbol{P} & 0 & 0 & \varepsilon_1\boldsymbol{I} & * & * \\ \boldsymbol{H}_i^T\boldsymbol{P} & 0 & 0 & 0 & \varepsilon_2\boldsymbol{I} & * \\ \Theta_1 & 0 & 0 & 0 & 0 & \varepsilon_3\boldsymbol{I} \end{pmatrix} < 0 \tag{11}$$

Proof: To analyze the stability of the sliding-mode dynamics (8), we consider the fuzzy uncertain system (2) with $\boldsymbol{w}(t) = 0$ and choose the following Lyapunov functional candidate

$$V(\boldsymbol{x},t) = \boldsymbol{x}^T(t)\boldsymbol{P}\boldsymbol{x}(t) + \int_{t-\tau}^{t} \boldsymbol{x}(m)^T\boldsymbol{P}\boldsymbol{x}(m)dm \tag{12}$$

By differentiating the given Lyapunov function, we obtain the differential along the trajectories as

$$\begin{aligned} \dot{V} &= \sum_{i=1}^{r} h_i(\boldsymbol{\theta})\Big\{\boldsymbol{x}^T(t)\big[\boldsymbol{P}(\boldsymbol{A}_i-\boldsymbol{B}_i\boldsymbol{K}_i)+(\boldsymbol{A}_i-\boldsymbol{B}_i\boldsymbol{K}_i)^T\boldsymbol{P}+\boldsymbol{Q}+2\boldsymbol{P}(\Delta\boldsymbol{A}_i+\Delta\boldsymbol{B}_i\boldsymbol{K}_i)\big]\boldsymbol{x}(t) \\ &\quad +2\boldsymbol{x}^T(t)\boldsymbol{P}(\boldsymbol{A}_{di}+\Delta\boldsymbol{A}_{di})\boldsymbol{P}\boldsymbol{x}(t-\tau)-2s^T[\boldsymbol{I}+\delta(t)]\{\rho_i sgn(s))+\boldsymbol{B}_i^T\boldsymbol{P}[\boldsymbol{A}_i\boldsymbol{x}+\boldsymbol{A}_{di}\boldsymbol{x}(t-\tau)]\}\Big\} \\ &\quad -(1-\dot{\tau})\boldsymbol{x}^T(t-\tau)\boldsymbol{Q}\boldsymbol{x}(t-\tau) \end{aligned} \tag{13}$$

.

Noting the definition of switching function $s(t)$ and the control law (6), we have

$$\begin{aligned} \dot{V} &= \sum_{i=1}^{r} h_i(\boldsymbol{\theta})\Big\{\boldsymbol{x}^T(t)\big[\boldsymbol{P}(\boldsymbol{A}_i-\boldsymbol{B}_i\boldsymbol{K}_i)+(\boldsymbol{A}_i-\boldsymbol{B}_i\boldsymbol{K}_i)^T\boldsymbol{P}+\boldsymbol{Q}\big]\boldsymbol{x}(t) \\ &\quad +2\boldsymbol{x}^T(t)\boldsymbol{P}(\boldsymbol{A}_{di}+\Delta\boldsymbol{A}_{di})\boldsymbol{P}\boldsymbol{x}(t-\tau)+2\boldsymbol{x}^T(t)\boldsymbol{P}(\Delta\boldsymbol{A}_i-\Delta\boldsymbol{B}_i\boldsymbol{K}_i)\boldsymbol{P}\boldsymbol{x}(t) \\ &\quad -2s^T(t)[\boldsymbol{I}+\delta(t)]\boldsymbol{B}_i\boldsymbol{P}\big[\boldsymbol{A}_i\boldsymbol{x}+\boldsymbol{A}_{di}\boldsymbol{x}(t-\tau)\big] \\ &\quad -2s^T(t)[\boldsymbol{I}+\delta(t)]\rho_i sgn(s)\Big\}-(1-\dot{\tau})\boldsymbol{x}^T(t-\tau)\boldsymbol{Q}\boldsymbol{x}(t-\tau) \end{aligned} \tag{14}$$

.

By Lemma 1, we obtain that for $\varepsilon_i > 0$, the following inequalities hold.

$$2\boldsymbol{x}^T\boldsymbol{P}\Delta\boldsymbol{A}_i\boldsymbol{x}(t) \le \varepsilon_1^{-1}\boldsymbol{x}^T(t)\boldsymbol{P}\boldsymbol{H}_i\boldsymbol{H}_i^T\boldsymbol{P}\boldsymbol{x}(t)+\varepsilon_1\boldsymbol{x}^T(t)\boldsymbol{E}_{1i}^T\boldsymbol{E}_{1i}\boldsymbol{x}(t) \tag{15}$$

$$2\boldsymbol{x}^T(t)\boldsymbol{P}\Delta\boldsymbol{A}_{di}\boldsymbol{x}(t-\tau) \le \varepsilon_2^{-1}\boldsymbol{x}^T(t)\boldsymbol{P}\boldsymbol{H}_i\boldsymbol{H}_i^T\boldsymbol{P}\boldsymbol{x}(t)+\varepsilon_2\boldsymbol{x}^T(t-\tau)\boldsymbol{E}_{2i}^T\boldsymbol{E}_{2i}\boldsymbol{x}(t-\tau) \tag{16}$$

$$2\boldsymbol{x}^T(t)\boldsymbol{P}\Delta\boldsymbol{B}_i\boldsymbol{K}_i\boldsymbol{x}(t) \le \varepsilon_3^{-1}\boldsymbol{x}^T(t)\boldsymbol{P}\boldsymbol{B}_i\boldsymbol{B}_i^T\boldsymbol{P}\boldsymbol{x}(t)+\varepsilon_3\rho_B^2\boldsymbol{x}^T(t)\boldsymbol{K}_i^T\boldsymbol{K}_i\boldsymbol{x}(t) \tag{17}$$

$$-2s^T[\boldsymbol{I}+\delta(t)]\rho_i sgn(s) \le -2\rho_i\|s\|+\rho_i[s^T\delta\delta^T s+s^T s]\|s\|^{-1} \le \rho_i(\rho_B^2-1)\|s\| \tag{18}$$

Noting that (3) and $\sum_{i=1}^{r} h_i(\boldsymbol{\theta}(\boldsymbol{t})) = 1$, and substituting the above inequalities into (14) results in

$$\dot{V} \le \sum_{i=1}^{r} h_i(\boldsymbol{\theta}) \begin{bmatrix} \boldsymbol{x}^T(t) & \boldsymbol{x}^T(t-\tau) \end{bmatrix} \times \Pi \times \begin{bmatrix} \boldsymbol{x}(t) \\ \boldsymbol{x}(t-\tau) \end{bmatrix} \tag{19}$$

where $\Pi = \begin{pmatrix} \Xi_1 & \boldsymbol{P}\boldsymbol{A}_{di} \\ \boldsymbol{A}_{di}^T\boldsymbol{P} & \Xi_2 \end{pmatrix}$, with

$$\begin{aligned}\Xi_1 &= \boldsymbol{P}(\boldsymbol{A}_i-\boldsymbol{B}_i\boldsymbol{K}_i)+(\boldsymbol{A}_i-\boldsymbol{B}_i\boldsymbol{K}_i)^T\boldsymbol{P}+\boldsymbol{Q}+\varepsilon_1^{-1}\boldsymbol{P}\boldsymbol{H}_i\boldsymbol{H}_i^T\boldsymbol{P}+\varepsilon_1\boldsymbol{E}_{1i}^T\boldsymbol{E}_{1i}\\ &\quad+\varepsilon_2^{-1}\boldsymbol{P}\boldsymbol{H}_i\boldsymbol{H}_i^T\boldsymbol{P}+\varepsilon_3^{-1}\boldsymbol{P}\boldsymbol{B}_i\boldsymbol{B}_i^T\boldsymbol{P}+\varepsilon_3\rho_B^2\boldsymbol{K}_i^T\boldsymbol{K}_i \qquad (20)\end{aligned}$$

$$\Xi_2=-(1-h)\boldsymbol{Q}+\varepsilon_2\boldsymbol{E}_{2i}^T\boldsymbol{E}_{2i} \qquad (21)$$

In the following, it will be shown that the LMI (11) implies $\Pi<0$. By Schur's complement, $\Pi<0$ is equivalent to the LMI shown in (24), with

$$\Xi_3=\boldsymbol{P}(\boldsymbol{A}_i-\boldsymbol{B}_i\boldsymbol{K}_i)+(\boldsymbol{A}_i-\boldsymbol{B}_i\boldsymbol{K}_i)^T\boldsymbol{P}+\boldsymbol{Q}+\varepsilon_1^{-1}\boldsymbol{P}\boldsymbol{H}_i\boldsymbol{H}_i^T\boldsymbol{P}+\varepsilon_1\boldsymbol{E}_{1i}^T\boldsymbol{E}_{1i}+\varepsilon_3\rho_B^2\boldsymbol{K}_i^T\boldsymbol{K}_i \qquad (22)$$

$$\Xi_4=\Xi_2 \qquad (23)$$

$$\begin{pmatrix}\Theta_3 & * & * & * & *\\ \boldsymbol{A}_{di}^T\boldsymbol{P} & \Theta_4 & * & * & *\\ \boldsymbol{H}_i^T\boldsymbol{P} & 0 & -\varepsilon_1\boldsymbol{I} & * & *\\ \boldsymbol{H}_i^T\boldsymbol{P} & 0 & 0 & -\varepsilon_2\boldsymbol{I} & *\\ \boldsymbol{B}_i^T\boldsymbol{P} & 0 & 0 & 0 & -\varepsilon_3\boldsymbol{I}\end{pmatrix}<0 \qquad (24)$$

It is shown that the LMI (11) implies the above matrix inequality (24). Together with (19) implies that for all $\begin{bmatrix}\boldsymbol{x}^T(t) & \boldsymbol{x}^T(t-\tau)\end{bmatrix}\neq 0$, we have

$$\dot{V}(\boldsymbol{x}(t),t)\leq 0 \qquad (25)$$

This means that the closed-loop fuzzy system (8) with $\boldsymbol{w}(t)=0$ is robustly asymptotically stable. Next, we shall show that the fuzzy uncertain system (2) satisfies

$$\|\boldsymbol{z}(t)\|_{E2}\leq\gamma\|\boldsymbol{w}(t)\|_2 \qquad (26)$$

for all non-zero $\boldsymbol{w}(t)\in L_2[0,\infty]$. To this end, we assume zero initial condition, that is, with $\boldsymbol{x}(t)=0$ for all $t\in[-d,0]$. Then, we can rewritten the Lyapunov function candidate as follows:

$$\begin{aligned}\dot{V} &= \sum_{i=1}^{r}h_i(\boldsymbol{\theta})\Big\{\boldsymbol{x}^T(t)\big[\boldsymbol{P}(\boldsymbol{A}_i-\boldsymbol{B}_i\boldsymbol{K}_i)+(\boldsymbol{A}_i-\boldsymbol{B}_i\boldsymbol{K}_i)^T\boldsymbol{P}+\boldsymbol{Q}\big]\boldsymbol{x}(t)\\ &\quad+2\boldsymbol{x}^T(t)\boldsymbol{P}(\boldsymbol{A}_{di}+\Delta\boldsymbol{A}_{di})\boldsymbol{P}\boldsymbol{x}(t-\tau)+2\boldsymbol{x}^T(t)\boldsymbol{P}(\Delta\boldsymbol{A}_i-\Delta\boldsymbol{B}_i\boldsymbol{K}_i)\boldsymbol{P}\boldsymbol{x}(t)\\ &\quad+2\boldsymbol{x}^T(t)\boldsymbol{P}\boldsymbol{B}_{wi}\boldsymbol{w}(t)-2s^T(t)[\boldsymbol{I}+\delta(t)]\boldsymbol{B}_i^T\boldsymbol{P}\big[\boldsymbol{A}_i\boldsymbol{x}(t)+\boldsymbol{A}_{di}\boldsymbol{x}(t-\tau)\big]\\ &\quad-2s^T(t)[\boldsymbol{I}+\delta(t)]\rho_i sgn(s)\Big\}-(1-h)\boldsymbol{x}^T(t-\tau)\boldsymbol{Q}\boldsymbol{x}(t-\tau) \qquad (27)\end{aligned}$$

.

Now, set

$$J(t)=\int_0^t[\boldsymbol{z}^T(m)\boldsymbol{z}(m)-\gamma^2\boldsymbol{w}^T(m)\boldsymbol{w}(m)]dm \qquad (28)$$

with $t>0$. It is easy to show that

$$\begin{aligned}J(t) &= \int_0^t[\boldsymbol{z}^T(m)\boldsymbol{z}(m)-\gamma^2\boldsymbol{w}^T(m)\boldsymbol{w}(m)+\dot{V}(\boldsymbol{x},t)]dm-V(\boldsymbol{x},t)\\ &\leq\int_0^t[\boldsymbol{z}^T(m)\boldsymbol{z}(m)-\gamma^2\boldsymbol{w}^T(m)\boldsymbol{w}(m)+\dot{V}(\boldsymbol{x},t)]dm \qquad (29)\end{aligned}$$

Hence, noting (15)-(19), it follows from (29) that

$$J(t) \leq \int_0^t \left[\begin{array}{ccc} \boldsymbol{x}^T(m) & \boldsymbol{x}^T(m-\tau) & \boldsymbol{w}^T(m) \end{array} \right] \times \Omega \times \left[\begin{array}{ccc} \boldsymbol{x}(m) & \boldsymbol{x}(m-\tau) & \boldsymbol{w}(m) \end{array} \right] dm \tag{30}$$

with $\Omega = \begin{pmatrix} \Theta_1 & \boldsymbol{PA}_{di} & 0 \\ * & \Theta_2 & 0 \\ * & * & -\gamma^2 \boldsymbol{I} \end{pmatrix}$, where Θ_1 and Θ_2 are given as in (9) and (10). By Schur's complement, it can be shown that $\Omega < 0$ is ensured by LMI (11). This together with (30) implies that $J(t) < 0$ for all $t > 0$. Hence, we obtain (26) from (30). □

Remark 2: It is noted that the condition in Theorem 1 is delay independent, which might be conservative when the time delay is known and small. Hence, it would be appropriate to extend the current study to delay-dependent issues in future research.

3.3 Reachability of the sliding-mode

As the last step of design procedure, we will further prove that the VSC controller in (6) ensures the reachability of the specified switching surface. It is known from [17] that the solution of the system (2) is given by

$$\begin{aligned} J(t) \quad = \quad & \boldsymbol{x}(t) = \varphi(0) + \int_0^t \sum_{i=1}^r h_i(\boldsymbol{\theta}) \big[(\boldsymbol{A}_i + \Delta \boldsymbol{A}_i)\boldsymbol{x} + (\boldsymbol{A}_{di} + \Delta \boldsymbol{A}_{di})\boldsymbol{x}(m-\tau) \\ & + (\boldsymbol{B}_i + \Delta \boldsymbol{B}_i)\boldsymbol{u}(m) + \boldsymbol{B}_{wi}\boldsymbol{w} \big] dm \end{aligned} \tag{31}$$

Hence, the switching function $s(t)$ can be expressed as

$$\begin{aligned} s(t) \quad = \quad & \sum_{i=1}^r h_i(\boldsymbol{\theta}) \boldsymbol{B}_i^T \boldsymbol{P} \varphi(0) + \int_0^t \sum_{i=1}^r h_i(\boldsymbol{\theta}) \boldsymbol{B}_i^T \boldsymbol{P} \big[(\boldsymbol{A}_i + \Delta \boldsymbol{A}_i)\boldsymbol{x}(m) + (\boldsymbol{A}_{di} + \Delta \boldsymbol{A}_{di})\boldsymbol{x}(m-\tau) \\ & + (\boldsymbol{B}_i + \Delta \boldsymbol{B}_i)\boldsymbol{u}(m) + \boldsymbol{B}_{wi}\boldsymbol{w} \big] dm. \end{aligned} \tag{32}$$

This means that $s(t)$ varies finitely. That is, it is rational to take the time derivation of $s(t)$. Hence, we have

$$\dot{s}(t) = \sum_{i=1}^r h_i(\boldsymbol{\theta}) \boldsymbol{B}_i^T \boldsymbol{P} \big[(\boldsymbol{A}_i + \Delta \boldsymbol{A}_i)\boldsymbol{x}(t) + (\boldsymbol{A}_{di} + \Delta \boldsymbol{A}_{di})\boldsymbol{x}(t-\tau) + (\boldsymbol{B}_i + \Delta \boldsymbol{B}_i)\boldsymbol{u}(t) + \boldsymbol{B}_{wi}\boldsymbol{w} \big] \tag{33}$$

and then, the reachability of the specified sliding surface $s(t) = 0$ can be obtained in the following theorem.

Theorem 4. *For the uncertain fuzzy time-delay systems (2) with the given switching function (5) where $\boldsymbol{G}_i = \boldsymbol{B}_i^T \boldsymbol{P}$ and $\boldsymbol{P}, \boldsymbol{Q}$, $\varepsilon_i (i = 1,2,3)$ is the solution of LMIs (11). Then, it can be shown that the state trajectories of the system (2) will be driven onto the switching surface $s(t) = 0$ for all $\boldsymbol{w}(t) \in L_2[0, \infty]$ by the above VSC law (6).*

Proof: For purpose of design integrity, a simple stability analysis based on Lyapunov direct method is carried out. Define the Lyapunov function

$$V(t) = \frac{1}{2} s^T (\boldsymbol{G}_i \boldsymbol{B}_i)^{-1} s \tag{34}$$

Noting that (6), the expressions of ρ and $\dot{s}$, thus, we have

$$\begin{aligned}\dot{\boldsymbol{V}}(t) &\leq s^T(t)\sum_{i=1}^{r}h_i(\boldsymbol{\theta})(\boldsymbol{G}_i\boldsymbol{B}_i)^{-1}\boldsymbol{G}_i\Big\{(\boldsymbol{A}_i+\Delta\boldsymbol{A}_i)\boldsymbol{x}(t)+(\boldsymbol{A}_{di}+\Delta\boldsymbol{A}_{di})\boldsymbol{x}(t-d)\\ &\quad +s^T(t)[\boldsymbol{I}+\delta(t)]\boldsymbol{u}(t)+s^T(t)(\boldsymbol{G}_i\boldsymbol{B}_i)^{-1}\boldsymbol{G}_i\boldsymbol{B}_{wi}\boldsymbol{w}(t)\}\\ &\leq s^T(t)\sum_{i=1}^{r}h_i(\boldsymbol{\theta})(\boldsymbol{G}_i\boldsymbol{B}_i)^{-1}\boldsymbol{G}_i\Big\{(\boldsymbol{A}_i-\boldsymbol{B}_i\boldsymbol{K}_i+\Delta\boldsymbol{A}_i)\boldsymbol{x}(t)-s^T(t)\delta(t)\boldsymbol{K}_i\boldsymbol{x}(t)\\ &\quad +s^T(\boldsymbol{G}_i\boldsymbol{B}_i)^{-1}\boldsymbol{G}_i(\boldsymbol{A}_{di}+\Delta\boldsymbol{A}_{di})\boldsymbol{x}(t-d)-s^T[\boldsymbol{I}+\delta(t)]\boldsymbol{G}_i[\boldsymbol{A}_i\boldsymbol{x}(t)+\boldsymbol{A}_{di}\boldsymbol{x}(t-d)]\\ &\quad -s^T(t)[\boldsymbol{I}+\delta(t)]\rho_i(\boldsymbol{x},t)sgn(s(t))+s^T(t)(\boldsymbol{G}_i\boldsymbol{B}_i)^{-1}\boldsymbol{G}_i\boldsymbol{B}_{wi}\boldsymbol{w}(t)\Big\}\end{aligned} \tag{35}$$

By (18), we have

$$\begin{aligned}\dot{\boldsymbol{V}}(t) &\leq \|s(t)\|\sum_{i=1}^{r}h_i(\boldsymbol{\theta})\Big\{\big[\|\Phi(\boldsymbol{A}_i-\boldsymbol{B}_i\boldsymbol{K}_i)\|+\|\Phi\boldsymbol{H}_i\|\|\boldsymbol{E}_{1i}\|+\rho_B\|\boldsymbol{K}_i\|\big]\|\boldsymbol{x}(t)\|\\ &\quad +\big[\|\Phi\boldsymbol{A}_{di}\|+\|\Phi\boldsymbol{H}_i\|\|\boldsymbol{E}_{2i}\|\big]\|\boldsymbol{x}(t-d)\|\\ &\quad +(1+\rho_B)\big[\|\boldsymbol{B}_i^T\boldsymbol{P}\boldsymbol{A}_i)\|\|\boldsymbol{x}(t)\|+\|\boldsymbol{B}_i^T\boldsymbol{P}\boldsymbol{A}_{di}\|\|\boldsymbol{x}(t-d)\|\big]\\ &\quad +\|\Phi\boldsymbol{B}_{wi}\|\|\boldsymbol{w}(t)\|-0.5\rho(\boldsymbol{x},t)(1-\rho_B^2)\Big\}\end{aligned} \tag{36}$$

Then, it follows from (36) that for $s(t)\neq 0$

$$\dot{V}(t)\leq -\beta\|s(t)\|<0 \tag{37}$$

which implies that the reachability of the specified switching surface is guaranteed, and the trajectories of the fuzzy uncertain system (2) are globally driven onto the specified switching surface $s(t)=0$ for all $\boldsymbol{w}(t)\in L_2[0,\infty]$. Moreover, it is seen that the existence domain of the sliding mode is the whole switching surface. □

Remark 3: In fact, the design strategy of the sliding-mode controller (6) accords with the so-called parallel distributed compensation (PDC) scheme [3, 5, 6, 10, 12]. This idea is that the overall controller is a fuzzy blending of each individual controller for each local linear model. The PDC method has been widely utilized in fuzzy control, and is proven to be a very appealing approach.

4 Simulation Studies

In this section, a simple design example is used to illustrate the approach proposed in this paper. Consider a T-S fuzzy uncertain stated-delay system with the following model

Plant rule i: IF $x_2(t)$ is η_{ii}, THEN

$$\begin{cases}\dot{\boldsymbol{x}}(t)=[(\boldsymbol{A}_i+\Delta\boldsymbol{A}_i)\boldsymbol{x}+(\boldsymbol{A}_{di}+\Delta\boldsymbol{A}_{di})\boldsymbol{x}(t-\tau)]+(\boldsymbol{B}_i+\Delta\boldsymbol{B}_i)\boldsymbol{u}(t)+\boldsymbol{B}_{wi}\boldsymbol{w}(t)\\ \boldsymbol{z}(t)=\boldsymbol{C}_i\boldsymbol{x}(t)\end{cases}$$

where $i=1,2$. The model parameters are given as $\boldsymbol{A}_1=\begin{bmatrix}0.1 & 0\\ 0 & -2\end{bmatrix}$,$\boldsymbol{A}_2=\begin{bmatrix}-0.3 & 0\\ 1 & -3\end{bmatrix}$,$\boldsymbol{A}_{d1}=\begin{bmatrix}0.1 & 0.1\\ 0 & 0.1\end{bmatrix}$,$\boldsymbol{A}_{d2}=\begin{bmatrix}0.1 & 0\\ 0 & 0.2\end{bmatrix}$,$\boldsymbol{B}_1=\boldsymbol{B}_2=\begin{bmatrix}1\\ 1\end{bmatrix}$,$\boldsymbol{B}_{w1}=\begin{bmatrix}2\\ 0\end{bmatrix}$,,$\boldsymbol{B}_{w2}=\begin{bmatrix}1\\ 0\end{bmatrix}$,$\boldsymbol{C}_1=\boldsymbol{C}_2=\begin{bmatrix}2 & 0\\ 0 & 1.5\end{bmatrix}$.
The uncertainties are set to be $\Delta\boldsymbol{A}_1=\begin{bmatrix}0 & 0.08\sin t\\ 0 & 0.06\sin t\end{bmatrix}$, $\Delta\boldsymbol{A}_2=\begin{bmatrix}0.06\sin t & 0\\ 0.02\sin t & 0.01\sin t\end{bmatrix}$, $\Delta\boldsymbol{A}_{d1}=\begin{bmatrix}0 & 0.06\sin t\\ 0 & 0.06\sin t\end{bmatrix}$,

$\Delta \boldsymbol{A}_{d2} = \begin{bmatrix} 0.01\cos t & 0 \\ 0 & 0.06\sin t \end{bmatrix}$, $\Delta \boldsymbol{B}_1 = \begin{bmatrix} 0.1\cos t \\ 0.1\cos t \end{bmatrix}$, $\Delta \boldsymbol{B}_2 = \begin{bmatrix} 0.1\sin t \\ 0.1\sin t \end{bmatrix}$, and the time-varying delay $\tau(t) = 0.5 + 0.5\sin t$ with $d = 1$ and $h = 0.5$. When choosing the matrix function as $F_i(t) = 0.02\sin t$, one can easily obtain the real constant matrices $\boldsymbol{H}_i$, $\boldsymbol{E}_{1i}$ and $\boldsymbol{E}_{2i}$ from Assumption 1. It is also obviously that $\rho_B = 0.1$ with $\delta(t) \leq \rho_B$. The membership functions are selected as $\eta_{11} = \sin^2(x_2)$ and $\eta_{22} = \cos^2(x_2)$. Let the initial state $\boldsymbol{x} = [0.9, 0.9]^T$,$t \in [-1, 0]$.

The problem at hand is to design a sliding mode controller such that the sliding motion in the specified switching surface is robustly stable, and the state trajectories can be driven onto the switching surface. To this end, we select the attenuation level $\gamma = 0.5$ and the matrices as follows: $\boldsymbol{K}_1 = [1.5, 2.3]$,$\boldsymbol{K}_2 = [0.2, 3.4]$.

By solving LMIs (11), we obtain:$\boldsymbol{P} = \begin{bmatrix} 1.5757 & -0.0144 \\ -0.0144 & 1.6211 \end{bmatrix}$,$\boldsymbol{Q} = \begin{bmatrix} 0.0729 & 0.127 \\ 0.127 & 0.5216 \end{bmatrix}$.

Hence, the switching surface can be obtained as $s = [0.6405, 0.6223]\boldsymbol{x}(t)$. It following from Theorem 2 that the desired VSC law can be obtained. The simulation results are given in Figures 1-3. Since it is well known that the chattering phenomenon is undesirable as it may incite high-frequency un-modeled dynamics and even leads to the instability of controlled system, we replace $sgn(\cdot)$ by $s/(\|s\| + \varepsilon)$(ε is the thickness of boundary layer) in the previous VSC law so as to prevent the control signals from chattering. However, it should also be pointed out that such an approach may lead to delay or make the controller less robust. Recently, to avoid chattering the use of high order and adaptive sliding mode is receiving more attentions; see, e.g., [10] for more details. It is seen that the reachability of the sliding motion can be guaranteed. Furthermore, the simulation results also show that our present design effectively attenuates the effect of both parameter uncertainties and external disturbances.

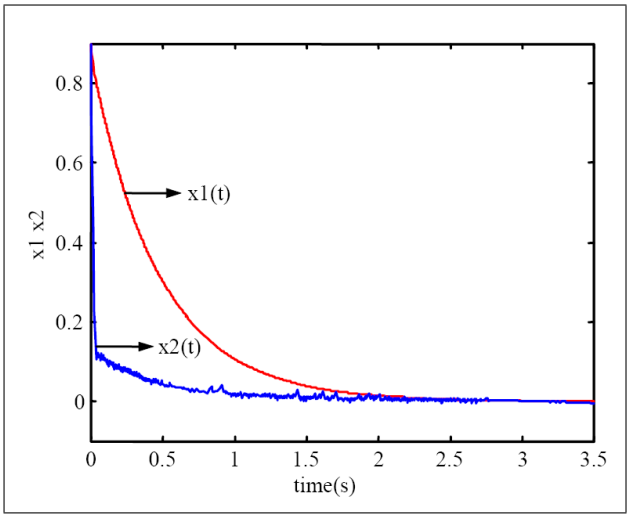


Figure 1: Trajectories of state x_1, x_2

5 Conclusions

This paper has firstly generalized the T-S model to represent a class of nonlinear uncertain systems. Then, a novel robust VSC method integrated with H_∞ technique, has been proposed for the fuzzy time-delayed system with parameters uncertainties and unmatched external disturbances. Moreover, by means of LMIs, a sufficient condition for the robustly stability of sliding motion with H_∞ disturbance attenuation level γ has been derived. It has been shown that both the switching surface and the VSC controller have been obtained by means of the feasibility of LMIs.

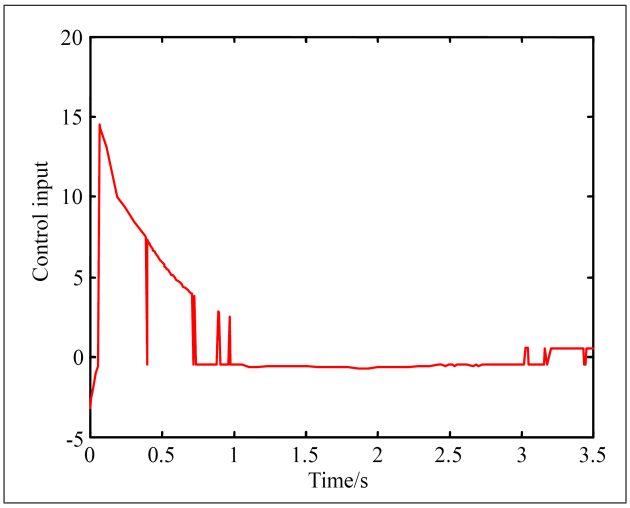


Figure 2: Switching surface $s(t)$

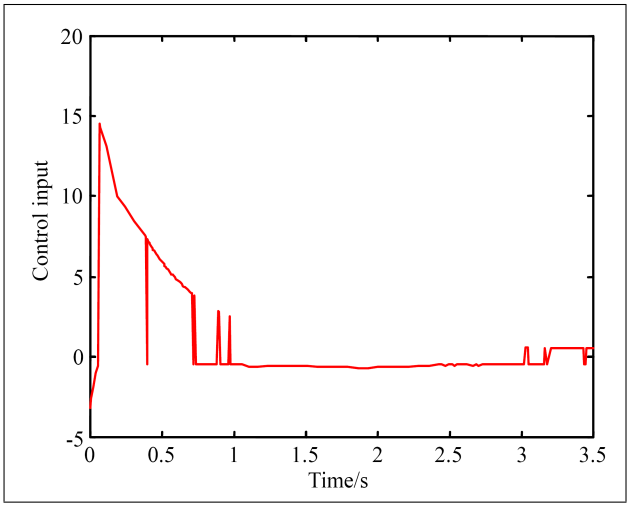


Figure 3: Control effort $\boldsymbol{u}(t)$

Bibliography

[1] T. Takagi, M. Sugeno, Fuzzy identification of systems and its applications to modeling and control, *IEEE Trans. Syst., Man, Cybern.*, 15(1):116-132, 1985.

[2] Y. Y. Cao, P. M. Frank, Stability analysis and synthesis of nonlinear time-delay systems via linear Takagi-Sugeno fuzzy models, *Fuzzy Sets Syst.*, Vol.124, No.2, pp.213-229, 2001.

[3] B. Chen, X. P. Liu, S. C. Tong, New delay-dependent stabilization conditions of T-S fuzzy systems with constant delay, *Fuzzy Sets Syst.*, Vol.158, No.20, pp.2209-2224, 2007.

[4] E. G. Tian, C. Peng, Delay dependent stability analysis and synthesis of uncertain T-S fuzzy systems with time-varying delay, *Fuzzy Sets Syst.*, Vol.157, No.4, pp.544-559, 2006.

[5] Y. Zhong, Y. P. Yang, New delay-dependent stability analysis and synthesis of T-S fuzzy systems with time-varying delay, *Int. J. Robust Nonlinear Control*, Vol.20, No.3, pp.313-322, 2009.

[6] P. Chen, Y. C. Tian, Improved delay-dependent robust stabilization conditions of uncertain T-S fuzzy systems with time-varying delay, *Fuzzy Sets Syst.*, Vol.159, No.20, pp.2713-2729, 2008.

[7] T. Z. Wu, Design of adaptive variable structure controllers for T-S fuzzy time-delay system, *Int. J. Adapt. Control Signal Process.*, Vol.24, No.2, pp.106-116, 2009.

[8] F. Gouaisbaut, M. Dambrine, J. P. Richard, Robust control of delay systems: a sliding mode control design via LMI, *Syst. Control Lett.*, Vol.46, No.4, pp.219-230, 2002.

[9] W. J. Cao , J. X. Xu, Nonlinear integral-type sliding surface for both matched and unmatched uncertain systems, *IEEE Trans. on Autom. Control*, Vol.49, No.8, pp.1355-1360, 2004.

[10] H. H. Choi, LMI-Based sliding surface design for integral sliding mode control of mismatched uncertain Systems, *IEEE Trans. on Autom. Control*, vol.52, no. 4, 736-742, 2007.

[11] B. S. Chen, C. H. Lee, Y. C. Chang, H_∞ tracking design of linear systems: Adaptive fuzzy approach, *IEEE Trans. Fuzzy Syst.*, Vol.4, No.1, pp.32-43, 1996.

[12] Y. He, Q. G. Wang, C. Lin, An improved H_∞ filter design for systems with time-varying interval delay, *IEEE Trans. Circuits Syst. II: Exp. Briefs*, Vol.53, No.11, pp.1235-1239, 2006.

[13] B. Chen, X. P. Liu, Delay-dependent robust H_∞ control for T-S fuzzy systems with time delay, *IEEE Trans. Fuzzy Syst.*, Vol.13, No.26, pp.544-556, 2005.

[14] P. Chen, Y. Dong, Y. C. Tian, New approach on robust delay-dependent H_∞ control for uncertain T-S fuzzy systems with interval time-varying delay, *IEEE Trans. Fuzzy Syst.*, Vol.17, No.4, pp.890-990, 2009.

[15] S. Xu, J. Lam, Robust H_∞ control for uncertain discrete-time-delay fuzzy systems via output feedback controllers, *IEEE Trans. Fuzzy Syst.*, Vol.13, No.1, pp.82-93, 2005.

[16] Y.C. Lin, J.C. Lo, Robust mixed H_2/H_∞ filtering for discrete-time delay fuzzy systems, *Int J. Syst. Science*, Vol.36, No.15, pp.993-1006, 2005.

[17] P. Gahinet, A. Nemirovski, A. J. Laub and M. Chilali, *LMI Control Toolbox*, Natick, MA: The MathWorks, 1995.

Author index