INTERNATIONAL JOURNAL

of

COMPUTERS, COMMUNICATIONS & CONTROL

ISSN 1841-9836

ISSN-L 1841-9836

A Bimonthly Journal
With Emphasis on the Integration of Three Technologies

Year: 2014 Volume: 9 Issue: 1 (February)

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).

Agora University Editing House

http://univagora.ro/jour/index.php/ijccc/

International Journal of Computers, Communications & Control

EDITORIAL ADDRESS:
Agora University/ 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: http://univagora.ro/jour/index.php/ijccc/

International Journal of Computers, Communications & Control

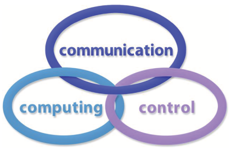
EDITORIAL BOARD

DATA FOR SUBSCRIBERS
Supplier: Cercetare Dezvoltare Agora Srl (Research & Development Agora Ltd.)
Fiscal code: 24747462
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

Short Description of IJCCC

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



Publication frequency: Bimonthly: Issue 1 (February); Issue 2 (April); Issue 3 (June); Issue 4 (August); Issue 5 (October); Issue 6 (December).

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(JCR)/Science Edition:
  - Impact factor (IF): JCR2009, IF=0.373; JCR2010, IF=0.650; JCR2011, IF=0.438; JCR2012, IF=0.441.
- 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: International Journal of Computers Communications & Control is directed to the international communities of scientific researchers in computer and control from the universities, research units and industry.

To differentiate from other similar journals, the editorial policy of IJCCC encourages the submission of scientific papers that focus on the integration of the 3 "C" (Computing, Communication, Control).

In particular the following topics are expected to be addressed by authors:

- Integrated solutions in computer-based control and communications;
- Computational intelligence methods (with particular emphasis on fuzzy logic-based methods, ANN, evolutionary computing, collective/swarm intelligence);
- Advanced decision support systems (with particular emphasis on the usage of combined solvers and/or web technologies).

International Journal of Computers, Communications & Control
INT J COMPUT COMMUN, ISSN 1841-9836
Volume 9, Issue 1, February, 2014.

Contents

INT J COMPUT COMMUN, ISSN 1841-9836
9(1):8-15, February, 2014.

Handwritten Documents Text Line Segmentation based on Information Energy

C.A. Boiangiu, M.C. Tanase, R. Ioanitescu

Costin-Anton Boiangiu*, Radu Ioanitescu
"Politehnica" University of Bucharest
Romania, 060042 Bucharest
*Corresponding author: icostin.boiangiu@cs.pub.ro

Mihai Cristian Tanase
VirtualMetrix Design
Romania, 060104 Bucharest
mihaicristian.tanase@gmail.com

Abstract: The first step in the text recognition process is represented by the text line segmentation procedures. Only after text lines are correctly identified can the process proceed to the recognition of individual characters. This paper proposes a line segmentation algorithm based on the computation of an information content level, called energy, for each pixel of the image and using it to execute the seam carving procedure. The algorithm proposes the identification of text lines which follow the text more accurately with the expected downside of the computational overhead.

Keywords: text line segmentation, text recognition, information energy, OCR.

1 Introduction

The identification of the boundaries of the lines of text represents an essential step in many algorithms like the ones for document structure extraction or text recognition. The research in this field has focused mostly on the development of such algorithms for printed documents. This limitation of the domain of application reduces the complexity of the problem as the printed documents are perfectly formed and the main problem that would need to be solved is the skew angle introduced in the process of printing or scanning, angle which is assumed to be the same for the entire document,. With such documents, the problem is reduced to the identification of the skew angle which is assumed to be constant for the entire page because the text lines are parallel with each other. Such methods are presented in [1]- [3].

However, when dealing with handwritten documents, the assumptions made by such algorithms do not hold anymore. There is no constant skew angle, the lines are not parallel and even the size and format similarity between the same characters found on different areas of the page cannot be assumed. Even worse, the separation between lines cannot be assumed as for printed documents because, in handwritten tex, the characters often overlap the line bellow as they are more compactly spaced on the vertical. Algorithms trying to address these increased complexities have also been attempted in [4] - [13].

The pixel of a given document stored as a digital image have different levels of information. A white pixel in a big white region or one that is found between two lines of text contains very little information while a character defining pixel would contain a lot of information. This paper proposes the association of an energy level for each pixel of the input image which tries to estimate the importance or the amount of information provided by that pixel. Although the concept can be applied to general images, printed or handwritten text, we will apply this concept for handwritten documents. The information level is then used by the seam carving algorithm for the segmentation of the text lines.

2 Related Work

The algorithms that try to locate the text lines in a document are divided mainly by the information they take from the input document. As the input documents are the results of a digitization process, they are acquired, generally through scanning, as grayscale images. This grayscale representation is converted into binary or black and white for algorithms which are designed to work with this type of documents only. The binary representation conversion is done using a previously defined threshold level. All pixels that have gray levels above the threshold are converted to black, the rest are converted to white. A too high threshold will results in an image containing too little information for text recognition to be possible and a too low threshold will result in too many artifacts, again making the text recognition process impossible.

Based on the observation that the body of the lines of text contains gray pixels and because the pixels that make the characters being of a darker shade of gray, some algorithms compute projection profiles representing the sum of all the pixels values in a given direction. The method works well for printed documents but fails to produce good results when applied to handwritten documents.

To address these problems, different approaches were used: identify the local skew of the handwritten text, calculate the accumulated space between characters, try to fill the space between characters or use attraction from the text pixels and repulsion from the previously detected line trying to estimate the text line boundaries closer.

3 Information Energy

The different pixels from an image carry different information content. A document can be viewed as a group of low information pixels representing the space between lines of text and respectively a group of high information pixels representing the actual lines of text. Each pixel in the energy map has a value associated with it that represents the amount of information that the given pixel stores in the image. If a high energy pixel is removed from the image, the resulting image has a significant drop in detail, whereas removing a low energy pixel results in a negligible information loss.

The information energy concept can be understood by trying to eliminate a continuous band of pixels from an image representing handwritten text. This concept is illustrated in 1.

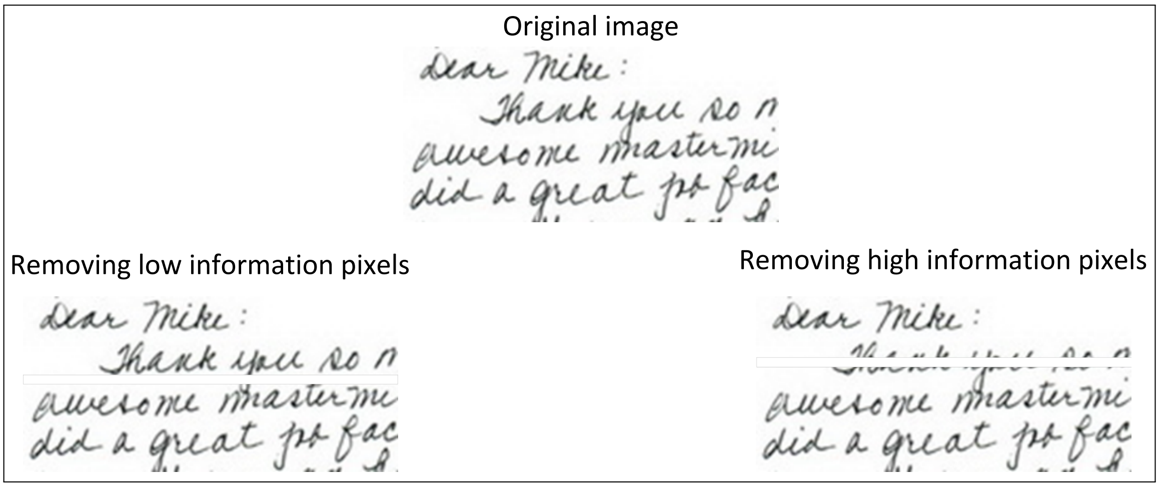


Figure 1: Information energy exemplification

A band with the same amount of pixels has been removed from a text document. It can be seen that when removing high information pixels there is not enough remaining information to detect the content, while when removing low information pixels, there is hardly any information loss and the content can be usually detected entirely.

An important observation is that for handwritten documents and in general text documents, the variation of pixel values represent information in itself. It can be seen that large continuous areas with similar pixel values constitute low energy areas, in particular spaces between lines of text, while high variation pixel values constitute high energy area. Computing the energy map of the entire document will thus provide information of the location of the high energy areas which represent the lines of text.

4 Accounting for the Text Direction

The algorithm begins with the calculation of the information energy for each pixel. Similarly to [4] and [11], for each pixel, the energy value is calculated using the next formula:

$$E(i,j) = 2 * e(i,j) + min\left(d(\frac{neighborsNumber}{2} + k) * E(i + direction, j + k)\right) \tag{1}$$

where:

- k fulfills the following condition: $-\frac{neighborsNumber}{2} < k < \frac{neighborsNumber}{2}$
- direction represents the direction of processing the energy map $+1$ representing left to right processing and -1 the opposite direction.

Using a neighborsNumber value of 3 and because k is a natural number since we work with discrete pixel positions, we get only the immediate neighbors. During the calculation the direction coefficient has been kept constant. As shown in images from the chapter “Test and Results” section, for documents with horizontal lines or very low skew angles we obtained good results. However, for larger skew angles the results quality decreased with wrong text line segmentation and as a result we considered accounting for the text direction during processing.

To accurately follow the text, the direction of the text lines should be taken into consideration at each pixel. For each pixel, the direction coefficient that accounts for the direction of the text line when that given pixel is reached, will be calculated.

The calculated information energy map is presented in 2 as a result from interpolating the two processing directions.

Computation algorithm

Input: image to be processed, processing window (height*width), skewing angle

For every pixel in the image to be processed:

```
1. For every processing window skewing angle
i. Sum up the pixels in the skewed window
ii. If the sum is less than the current minimum then
a) Update the current minimum
b) Save the window variation
2. With the minimum variation saved in step 1 ii:
i. Update the direction coefficient for the minimum variation
ii. Rescale the minimum window variation
```

Good results were obtained experimentally with processing window sizes that were at least the width and height of the average character in the processed text and for values of a few order

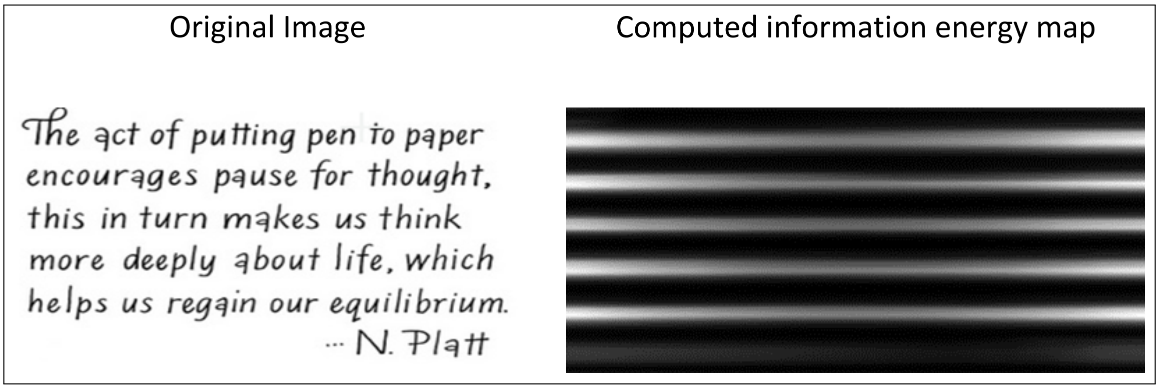

Original Image
Computed information energy map
The act of putting pen to paper
encourages pause for thought,
this in turn makes us think
more deeply about life, which
helps us regain our equilibrium.
... N. Platt

Figure 2: The information energy map of an image

of magnitude larger. If the processing windows has a smaller size then the algorithm attempts to segment the characters specific structures as separate lines of text leading to erroneous results.

Because the size of the processing text is usually roughly known and since the results are good with larger sizes for the processing window, this aspect does not represent a limitation. Similarly, the minimum and maximum variation angles can be set to 20 degrees. These variation angles represent the maximum skewing angles and combined cover a 40 degrees area that is sufficient for most documents.

The size of the processing window is fixed upfront, the calculations at each step are fixed. The total number of operations depends on the direction coefficient and height of the initial image since the algorithm is repeated for every pixel. For each pixel and for each variation of the skewing angle, the information energy level is computed.

Line identification algorithm

Input: computed information energy for each pixel in the image to be processing, neighborsNumber

Output: the number and layout of the identified lines

```
1. For each line (1..h)
2. For each column (1..w)
3. Select the minimum cost pixel on the right not found further away than neighborsNumber
i. If the selected pixel is not included in any line include it
ii. Else move to the next line
```

The algorithm looks through all the information energy levels to locate the minimal values. The values with the minimum energy level represent the blank pixel regions that separate the lines of text.

5 Tests and Results

To test the algorithm, a number of images of handwritten documents have been used. The test data files consisted of about 500 different types of documents representing old letters, library index files, patents, receipts and various printed documents. The documents showed pronounced skew angles and their layout was not trivial. The database was considered to be relevant, although the number of documents is not large, because of their variety which allowed the testing of the algorithm on a wide set of conditions.

Different methods for the computation of the information energy are used as examples to show how the algorithm depends on this type of variation and to show the general application of the concept. The results of the algorithm are discussed in the conclusions section and future possible solutions are presented with an explanation of the associated computational costs.

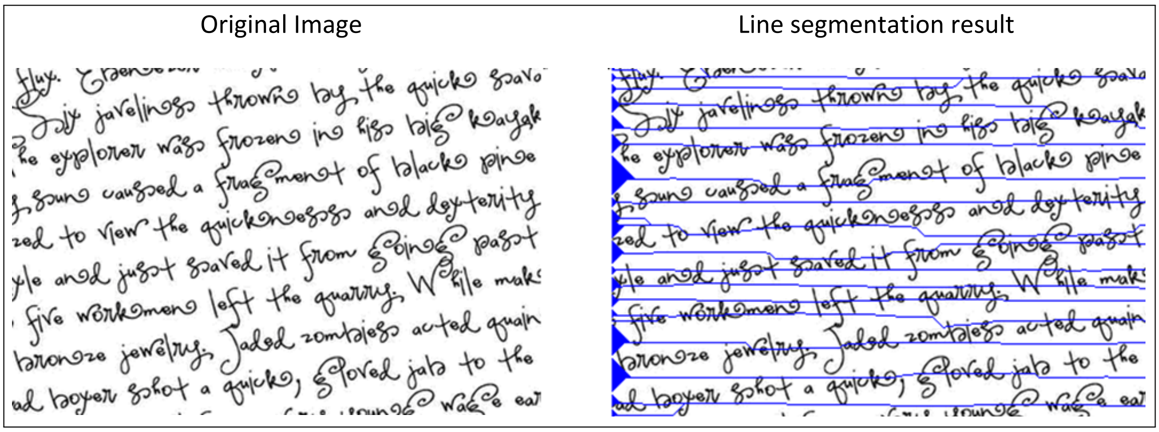


Figure 3: Gaussian first derivative energy computation Test 1

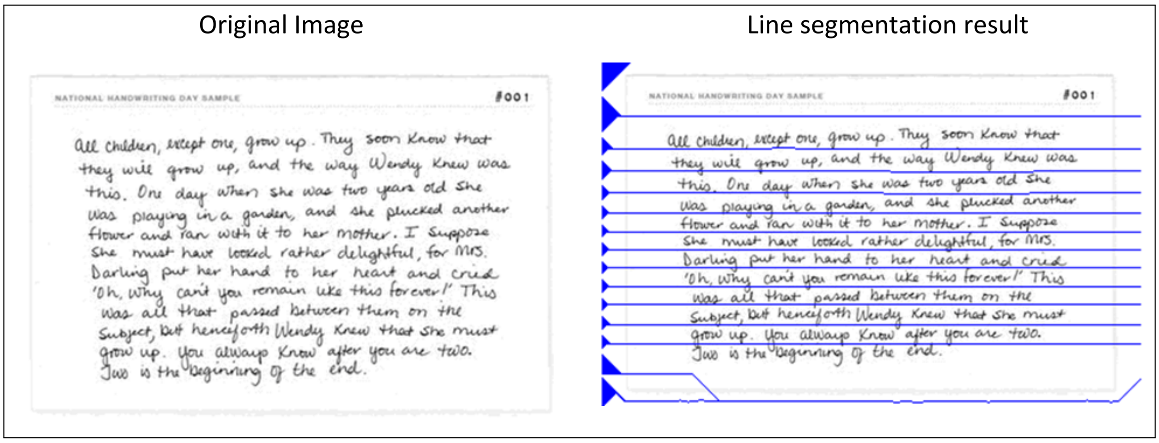


Figure 4: Gaussian first derivative energy computation Test 2

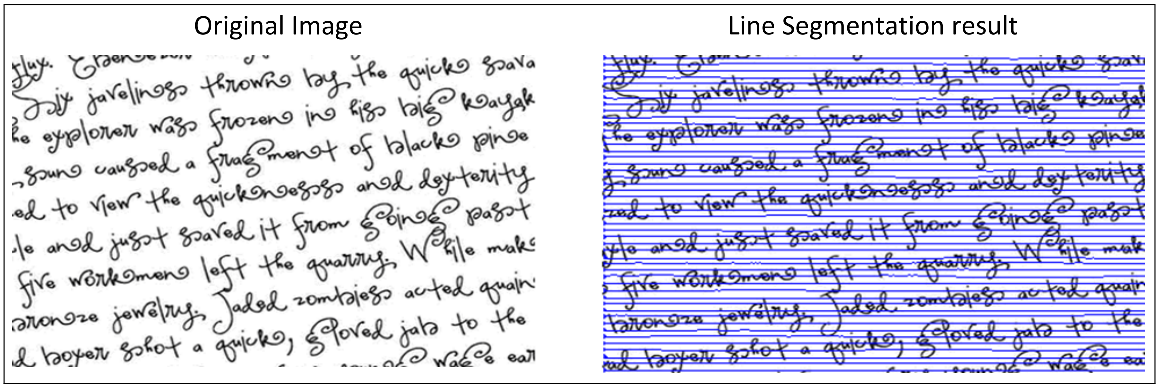


Figure 5: Magnitude of the gradient energy computation Test 1

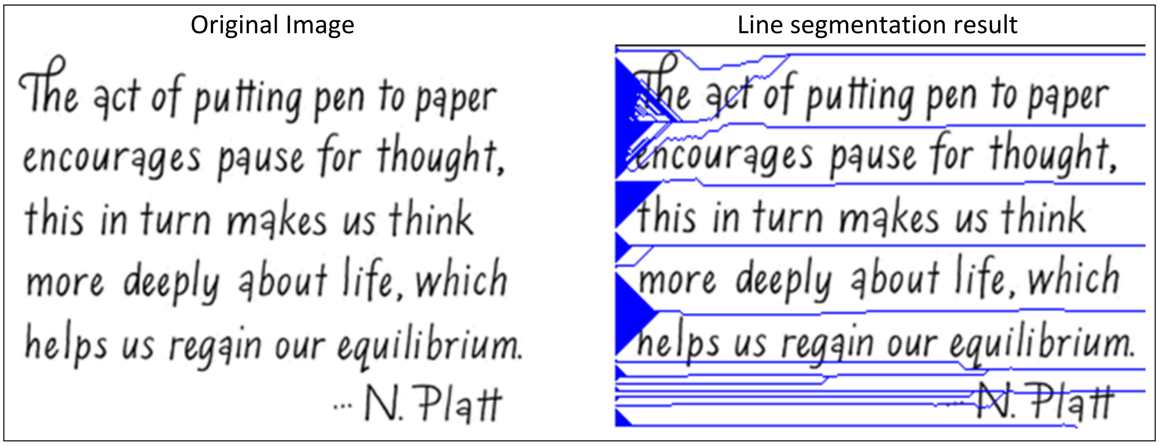

Original Image
Line segmentation result
The act of putting pen to paper
encourages pause for thought,
this in turn makes us think
more deeply about life, which
helps us regain our equilibrium.
... N. Platt

Figure 6: Magnitude of the gradient energy computation Test 2

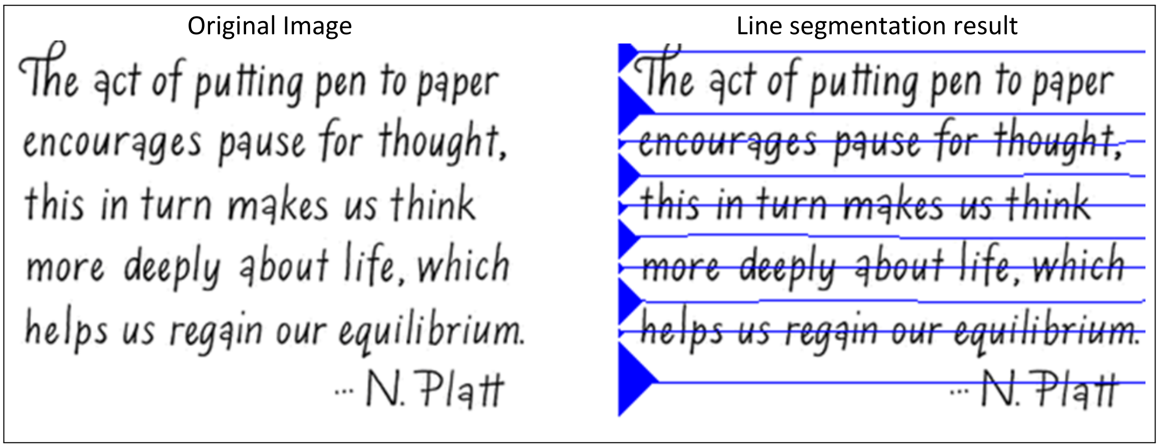

Original Image
Line segmentation result
The act of putting pen to paper
encourages pause for thought,
this in turn makes us think
more deeply about life, which
helps us regain our equilibrium.
... N. Platt

Figure 7: Inverse Distance Transform energy computation 1

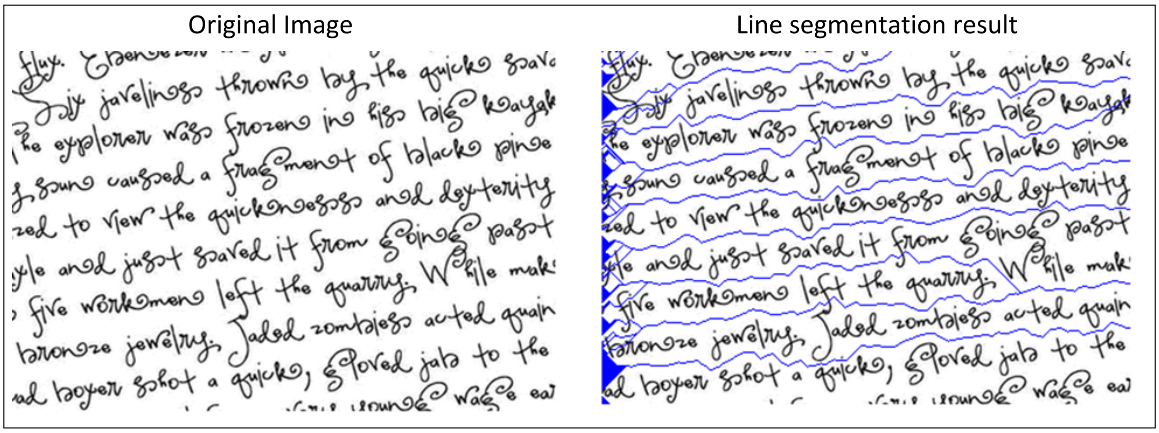


Figure 8: Inverse Distance Transform energy computation Test 2

6 Conclusions and Future Works

This paper describes a text line segmentation algorithm that shows good results even for handwritten documents. The algorithm is based on the concept of information energy which is used to estimate the text lines in the processed document.

Experimental results showed that using constant direction coefficients when calculating the information energy levels produces bad results for inputs that show high levels of skew. The algorithm addresses this problem by updating the direction coefficients with a window processing algorithm. These direction coefficients account for the local text direction and for variations of the skew angles that are not uncommon in handwritten documents. By using direction coefficients, pixels from the same line have higher probability of selection when calculating the minimum values in the information energy map.

The algorithm that computes the direction coefficient has a large computation complexity. Since the computation time depends on the size of the processing window, one solution is to use a smaller size. Experimental testing showed that at minimum, the height and width of the average character should be used for the size of this window. Alternatively the skew angle could be assumed to be smaller eliminating parts of the computation cases. Another change could be the limitation of the possible line curvature angle for the detected document lines to a lower interval. By constraining the line curvature angle. This limitation would also address cases in which a detected line would follow the space between two words and reach the previous or the next line which would represent a wrong line detection.

A possible future direction is the evaluation of the robustness of the algorithm on larger images datasets and with a variation of the document types which would allow to more extensively evaluate the accuracy of the handwritten text segmentation algorithm.

The work presented in this paper is a building block of a much bigger project: a complete, modular, fully automatic content conversion system developed for educational purposes. In the near future, with the completion of the system and the running in automatic batch processing of large image databases of all kind of skewed documents (containing handwriting or not) the algorithm will be fully evaluated in order to assess its real potential as a preprocessing phase for OCR applied on handwritten documents.

Acknowledgement

The work presented in this paper was funded by the Sectorial Operational Programme Human Resources Development 2007-2013 of the Romanian Ministry of Labour, Family and Social Protection through the financial agreement POSDRU/89/1.5/S/62557.

Bibliography

[1] dos Santos, R.P. et al, Text Line Segmentation Based on Morphology and Histogram Projection, *Document Analysis and Recognition (ICDAR)*, 651- 655, 2009.

[2] Saha, S. et al, A Hough Transform based Technique for Text Segmentation, *Journal of Computing*, 2(2):135-140, 2010.

[3] Arivazhagan, M. et al, A Statistical approach to line segmentation in handwritten documents, *Proceedings of SPIE*, 2007.

[4] Strand, L. et al, Minimal Cost-Path for Path-Based Distances, *Image and Signal Processing and Analysis*, 379-384, 2007.

[5] Avidan, S. et al, Seam Carving for Content-Aware Image Resizing, *ACM Siggraph*, article 10, 2007.

[6] Saabni, S. et al, Language-Independent Text Lines Extraction Using Seam Carving, *Document Analysis and Recognition (ICDAR)*, pp. 563-568, 2001.

[7] Papavassiliou, V. et al , Handwritten document image segmentation into text lines and words, *Pattern Recognition*, 43(1):369-377, 2010..

[8] Du, X. et al, Text Line Segmentation in Handwritten Documents Using Mumford-Shah Model, *Pattern Recognition*, 42(12):3136-3145, 2009.

[9] T ripathy, N.; Pal, U., Handwriting segmentation of unconstrained Oriya text, *Frontiers in Handwriting Recognition*, 306-311. 2004.

[10] Kennard, D.J., Barrett, W.A., Separating Lines of Text in Free-Form Handwritten Historical Documents, *Document Image Analysis for Libraries*, 12-23, 2006.

[11] Asi, A. et al, Text Line Segmentation for Gray Scale Historical Document Images, *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, pp. 120-126, 2011.

[12] Bar-Yosef, I., Input sensitive thresholding for ancient Hebrew manuscript, *Pattern Recognition Letters*, 26(8):1168-1173, 2005.

[13] Bar-Yosef, I. et al, Line segmentation for degraded handwritten historical documents, *Document Analysis and Recognition*, 1161-1165, 2009.

INT J COMPUT COMMUN, ISSN 1841-9836
9(1):16-23, February, 2014.

Non-Negative Factorization for Clustering of Microarray Data

L. Morgos

Lucian Morgos
Dept. of Electronics and Telecommunications
Faculty of Electrical Engineering and Information Technology
University of Oradea
Romania, 419987 Oradea, Universitatii, 1
lmorgos@uoradea.ro

Abstract: Typically, gene expression data are formed by thousands of genes associated to tens or hundreds of samples. Gene expression data comprise relevant (discriminant) information as well as irrelevant information often interpreted as noise. The irrelevant information usually affects the efficiency of discovering and grouping meaningful latent information correlated to biological significance, process closely related to data clustering. Class discovery through clustering may help in identifying latent features that reflect molecular signatures, ultimately leading to class forming. One solution for improving the class discovery efficiency is provided by data dimensionality reduction, where data is decomposed into lower dimensional factors, so that those factors approximate original data.

Keywords: computational intelligence, microarray data analysis, clustering, recognition.

1 Introduction

When analyzing high dimensional data the analysis process often becomes prohibitive and computational intractable if the data is analyzed in its receiving form. Fortunately, in many cases, data comprises redundant information, irrelevant or insignificant entries which can be discarded without significantly affecting the final result or decision, process known as *dimensionality reduction*. Discovering latent and meaningful low-dimensional factors from high-dimensional data is of great importance in many research fields. Basically, dimensionality reduction may be accomplished by either feature selection or data transformation. The latter is a factorization process that decomposes data into meaningful factors so that their product approximates the initial data as good as possible. By factorization, some initial information is lost; however, the factorization of gene expression should be performed so that the factors provide a biologically meaningful decomposition of the data, which is a major goal for this particular task. The cluster statistics should match the known sample labels when statistics are available from the full data.

Let us consider our data containing m samples of n - dimensional vectors stored into a $n \times m$ matrix $\mathbf{V}$, where the vector elements indexed as $i = 1 \dots n$ correspond to the receiving information values. Dimensionality reduction refers to transforming the n - dimensional data into a lower r - dimensional data. This further relates to decomposing the matrix $\mathbf{V}$ into factors $\mathbf{W}$ of dimension $n \times r$ and $\mathbf{H}$ of dimension $r \times m$, where $r < (n, m)$ is often named the factorization rank, so that their product approximates $\mathbf{V}$, i.e. $\mathbf{WH} \cong \widetilde{\mathbf{V}} \approx \mathbf{V}$. For analyzing the receiving information the n - dimensional space is replaced by the lower r - dimensional space where the new information resides. Typically, $\mathbf{W}$ represents the *basis* factors and $\mathbf{H}$ the *encoding* (also know as the decomposition *coefficients*). Thus, each $j = 1 \dots m$ information embedded into an n - dimensional column vector $\mathbf{v}_j$ can be represented as linear (or nonlinear) combination of the basis matrix $\mathbf{W}$ and the corresponding low dimensional coefficients vector h_j, more precisely, $\mathbf{v}_j = \mathbf{W}\mathbf{h}_j$. When the factorization process is applied to gene expression data from a set of microarray experiments, we typically deal with gene expression profile for a single gene across many samples or experimental conditions. More precisely, the matrix $\mathbf{V}$ comprises in its rows the expression levels of genes, while the columns correspond to samples. Usually, n is of order of thousands while m is often less than 100. In this context NMF seeks for finding a small number of metagenes. Therefore, each column of $\mathbf{W}$ represents a metagene and each column of $\mathbf{H}$ denotes the expression profile.

2 Approach

One of the frequently used dimensionality reduction approaches is given by singular values decomposition (SVD). SVD has been proposed [1] to reduce the gene expression space where the columns of $\mathbf{W}$ form orthonormal basis called eigenarrays, while $\mathbf{H}$ forms eigengenes. Sorting the genes according to the descending values of eigengenes gives a global picture of the dynamics of the gene expression, in which individual genes and arrays appear to be clustered into groups of similar functions or cellular states. A robust SVD variant [2] has been further proposed to cope with outliers or missing values often found in gene expression. Other approaches refer to novel methods for analyzing SVD for obtaining gene groups along with a confidence measure [3], or SVD based biclustering through incorporating simultaneous normalization of rows and columns [4].

A relatively recent approach for data factorization is provided by nonnegative matrix factorization, where both factors are constrained to have nonnegative entries. The work of Lee and Seung published in *Nature* [5] can be considered as a starting point from where the interest in this approach almost exploded due to the algorithm's simplicity and its compliance with physiological principles. Ever since, new NMF variants have been developed in the attempt to improve its convergence, numerical stability, processing speed or to extend the approach to temporal domain. Original NMF has been applied to gene clustering [6] and, recently a sparse NMF approach has been developed [7], that appears to yield better clustering performance compared to the standard NMF.

Although the original NMF has been successfully applied to gene expression clustering, its application was rather limited to three data sets. We have extended the previous work of [6] by employing three more expression data sets in order to validate the efficiency of this decomposition approach. In addition, two more NMF variants were applied for a systematic comparison. The extension also includes the use of k-means clustering strategy along with the expression intensity provided by the rows of $\mathbf{H}$.

Moreover, the suitability of NMF approaches was investigated for the recognition task. Finally, we rank the algorithms according to their clustering and recognition performance as well as with respect to their stability, as defined later.

3 Methods

3.1 Non-negative matrix factorization

In the NMF formulation the difference between the initial data matrix and the product of its decomposition factors can be expressed in different ways leading to various cost functions f_{NMF} which quantify the decomposition quality. One frequently cost function is described by the Kullback-Leibler (KL) divergence based cost function:

$$f_{NMF}(\mathbf{v},\mathbf{w},\mathbf{h}) \triangleq \sum_{i,j}\left(v_{ij}\ln\frac{v_{ij}}{\sum_l w_{il}h_{lj}} + \sum_l w_{il}h_{lj} - v_{ij}\right), \tag{1}$$

for any $i = 1 \dots n$, $j = 1 \dots m$, and $l = 1 \dots r$.

One solution for minimizing the above cost functions is given by adopting an Expectation - Maximization (EM) like strategy [5] for iteratively updating the factors. Starting with random values for $\mathbf{W}$ and $\mathbf{H}$, at each iteration, the following updating rules are used to guarantee the minimization (towards a local minimum) of the KL-based cost function:

$$h_{lj} \longleftarrow h_{lj} \frac{\sum_i w_{li} \frac{v_{ij}}{\sum_l w_{il} h_{lj}}}{\sum_i w_{il}} \tag{2}$$

$$w_{il} \longleftarrow w_{il} \frac{\sum_j h_{jl} \frac{v_{ij}}{\sum_k w_{il} h_{lj}}}{\sum_j h_{lj}} \tag{3}$$

3.2 Local non-negative matrix factorization

To enhance the decomposition sparseness, Li et al [8] have developed the Local Non-negative Matrix Factorization (LNMF) algorithm, imposing more constraints to the KL cost function to get more localized basis factors. The associated cost function is then given by:

$$h_{lj} \longleftarrow \sqrt{h_{lj} \sum_i v_{ij} \frac{w_{il}}{\sum_l w_{ik} h_{lj}}} \tag{4}$$

$$w_{il} \longleftarrow \frac{w_{il} \sum_j v_{ij} \frac{h_{lj}}{\sum_l w_{il} h_{lj}}}{\sum_j h_{lj}} \tag{5}$$

However, we should note that, here, although sparseness issue refers to the basis factors, this approach does not guarantee a high degree of sparseness for the expression profiles. As we shall see in the experimental section, the coefficient sparseness indeed seems to positively correlate with the accuracy performance.

3.3 Polynomial non-negative matrix factorization

Unlike NMF, the polynomial NMF [9] [10] was developed as alternative to the standard NMF for decomposing data in some nonlinear fashion. PNMF assumes that the input data $\mathbf{V} \in \mathcal{V} \subseteq \mathbb{R}^{n \times m}$ are transformed to a higher dimensional space $\mathcal{F} \subseteq \mathbb{R}^{l \times m}$, $l \gg n$. By denoting the set of the transformed input data with $\mathbf{F} = [\phi(\mathbf{v}_1), \phi(\mathbf{v}_2), \dots, \phi(\mathbf{v}_m)]$, with the l - dimensional vector expressed as $\phi(\mathbf{v}_j) = [\phi(\mathbf{v})_1, \phi(\mathbf{v})_2, \dots, \phi(\mathbf{v})_s, \dots, \phi(\mathbf{v})_l]^T \in \mathcal{F}$, a matrix $\mathbf{Y} = [\phi(\mathbf{w}_1), \phi(\mathbf{w}_2), \dots, \phi(\mathbf{w}_r)]$, $\mathbf{Y} \in \mathcal{F}$, that approximates the transformed data set, such that $r < n$, can be found. Therefore, each vector $\phi(\mathbf{v})$ can be written as a linear combination as $\phi(\mathbf{v}) \approx \mathbf{Yh}$. Defining a kernel function κ that satisfies $\kappa^{(\mathbf{vw})} \triangleq \kappa(\mathbf{v}, \mathbf{w}) = \langle \phi(\mathbf{v}), \phi(\mathbf{w}) \rangle$, for all $\mathbf{v}, \mathbf{w} \in \mathcal{V}$, where ϕ is a mapping from $\mathcal{V}$ to an (inner product) feature space $\mathcal{F}$, $\phi : \mathbf{v} \longrightarrow \phi(\mathbf{v}) \in \mathcal{F}$, the following squared Euclidean distance based cost function was proposed [9]:

$$\begin{aligned} f_{PNMF}(\mathbf{v}, \mathbf{w}, \mathbf{h}) &= \frac{1}{2} \|\phi(\mathbf{v}_j) - \mathbf{Yh}_j\|^2 \\ &= \mathbf{k}(\mathbf{v}, \mathbf{v}) - 2\mathbf{k}(\mathbf{v}, \mathbf{w}_l)\mathbf{h} + \mathbf{h}^T \mathbf{k}(\mathbf{w}_o, \mathbf{w}_l)\mathbf{h}, \end{aligned} \tag{6}$$

The cost function f_{PNMF} should be minimized subject to $h_l, w_{il} \geq 0$, and $\sum_{i=1}^{n} w_{il} = 1$, $l = 1 \dots r$.

4 Results

The experiments were conducted with respect to the following aspects: (i) *clustering error* defined as the number of misclustered samples expressed in percentage, and (ii) *model stability* defined thorough the standard deviation of the clustering error for several NMF runs, and (iii) *recognition error* defined as the percent of misclassified samples from the test set.

4.1 Data set description

We have applied NMF, LNM, and PNMF to six publicly available microarray data sets: *leukemia* [11], *medulloblastoma* [12] , *colon* [13], *DLBCL* [14], *SRBCT* [15], and *prostate* [16] data set, respectively. The first two data sets are exactly the ones used by [6]. We should also note that the original *DLBCL* data set initially contained 7126 genes. By setting an intensity thresholds at 20 - 16000 units, then filtering out genes with max/min ≤ 3 or (max - min) ≤ 100, the final number if genes was 6285. The filtering procedure was similar to the one described in [17]. For the *prostate*, the conditions for gene selection was a threshold at 100 - 16000 units, with max/min ≤ 5 or (max - min) ≤ 50, conducting from a total of initial 12600 to 5966 genes.

4.2 Clustering

Starting from random values for both factors **W** and **H**, NMF iteratively modifies them until the algorithm converges to a local minimum. More precisely, the running stops when the difference between **V** and the product of its decomposition factors reaches a minimum imposed threshold. Due to its stochastic nature, the algorithm does not always converge to the same solution with different initial random values for the factors. Therefore, NMF should run several times. In our experiments we ran NMF 30 times, and the minimum (*Min.*) and average (*Av.*) clustering error along with its standard deviation (*Std.*) are reported. Standard deviation is a good indicator for the model stability. A small value reveals better stability, i.e. the algorithm leads to approximately the same performance for multiple runs.

Once NMF converged and found its final decomposition factors, it is the **H** that contains clustering information. There are two approaches to assign sample labels. One approach suggested by [6] refers to the label assignment according only to the relative values in each column of H, i.e. the class labels are determined by the index of the highest metagene expression. When the number of rows of **H** equals the number of known clusters we count the number of maximum values for each column of **H** with respect to the known class range. Clustering errors are simply the misassignments, i.e. the mismatch between the index and associated value. Table 1 depicts the clustering errors, where the table column denoted by H indicates the first clustering criteria. It may happen that this procedure does not lead to the correct number of clustering for a particular run. Even worse, whenever this approach fails to discover the correct number of clusters for any of those 30 runs, we report this case as N/A in the table.

The first clustering approach only applies when the number of rows of **H** (viewed as dimension of **H**, or the decomposition *rank*) r equals the number of known clusters. The second clustering approach deals with varying ranks. The NMF algorithms were run for $r = \{2, 3, 4, 5, 6, 7, 8\}$. Each r - dimensional column vector of **H** was clustered with the help of K - means clustering strategy. We have also carried out experiments with SVD, as baseline. The SVD method always fails to discover the correct number of clusters for the first clustering approach for all data sets. To emphasis the importance of dimensionality reduction task, experiments with the initial data dimension were also conducted and shown in Table 1 corresponding to the row termed “No processing”. Indeed, the high dimension technically halves the performance of the K - means clustering.

Table 1: Clustering results for the NMF methods. Clustering results for SVD as dimensionality reduction approach is also tabulated. For comparison purpose, the clustering errors are also shown when K means is applied for the initial data dimension. The first two lowest errors are emphasized in bold.

Method	Approach	Clust. Err. (%)	Data					
			Leukemia	Medulloblastoma	Colon	DLBCL	SRBCT	Prostate
No processing	K means	Min.	13.15	14.70	20.96	28.57	34.93	36.27
SVD	K means	Min.	5.26	11.76	**11.29**	18.18	N/A	**18.62**
		Av.	23.97	34.11	29.26	27.27	N/A	38.67
		Std.	17.42	9.61	9.27	6.94	N/A	10.86
NMF	H	Min.	2.63	38.23	25.80	28.57	N/A	N/A
		Av.	4.94	39.64	26.25	34.41	N/A	N/A
		Std.	1.15	1.49	0.73	4.76	N/A	N/A
	K means	Min.	**0**	**5.88**	**11.29**	**12.98**	**21.68**	24.50
		Av.	9.29	8.08	25.00	23.11	26.80	40.68
		Std.	10.75	6.23	9.96	5.60	8.65	6.69
LNMF	H	Min.	10.52	38.23	32.25	29.87	N/A	N/A
		Av.	17.57	38.23	34.38	32.20	N/A	N/A
		Std.	2.24	0	2.45	0.99	N/A	N/A
	K means	Min.	2.63	35.29	46.77	27.27	33.73	N/A
		Av.	12.65	35.29	46.77	29.71	50.96	N/A
		Std.	9.90	0	0	2.06	4.50	N/A
PNMF	H	Min.	2.63	41.17	33.87	28.57	N/A	N/A
		Av.	2.63	41.17	33.87	28.57	N/A	N/A
		Std.	0	0	0	0	N/A	N/A
	K means	Min.	**0**	11.76	14.51	23.37	32.53	**16.66**
		Av.	15.41	29.41	21.58	30.99	43.01	33.23
		Std.	9.05	9.03	4.72	6.00	5.90	7.98

As far as the clustering ability is concerned, NMF seems to yield the lowest clustering error for all data sets but *prostate*. Comparing the two clustering approaches, K - means clearly outperforms the first approach for all data sets and all the decomposition methods. Moreover, except the *prostate* case for LNMF, this clustering approach always finds the correct number of classes for several NMF runs. For the *leukemia* data set, K - means conducted to perfect clustering (zero clustering error) for both NMF and LNMF. Brunet et al. [6] as well as other authors, including Kim and Park [7] reported that two ALL samples were consistently misclassified by most methods, including NMF. The same tendency has been noticed by us when applying the first clustering strategy but not with the second one. However, we should stress that this data set is probably an easy one compared to the others, in terms of method's clustering ability.

It is worth mentioning that, amongst all data sets, the *prostate* data seems to be the most difficult one to be correctly clustered. This indicates hard overlaps within samples from both tumor and non-tumor classes. Surprisingly, although PNMF performs modest on the other data sets, in this case it had conducted to the lowest clustering error. This behaviors could be due to the nonlinear nature of the PNMF algorithm that enables to find discriminant non-linear biological structures for this data set. For the PNMF algorithm the best result corresponds to the polynomial degree of 2.

Analyzing the model stability, LNMF followed by the PNMF algorithm are the most stable models and tend to make the same errors for various numbers of algorithm runs, as indicated by low or even zero standard deviation value. However, the LNMF was outperformed by all other methods.

The stability of the algorithm was also measured through the cophenetic correlation coefficient [6] computed for various decomposition ranks and displayed in Figure 1. All algorithms tend to be instable as rank increases. However, the behavior differs from one data set to another.

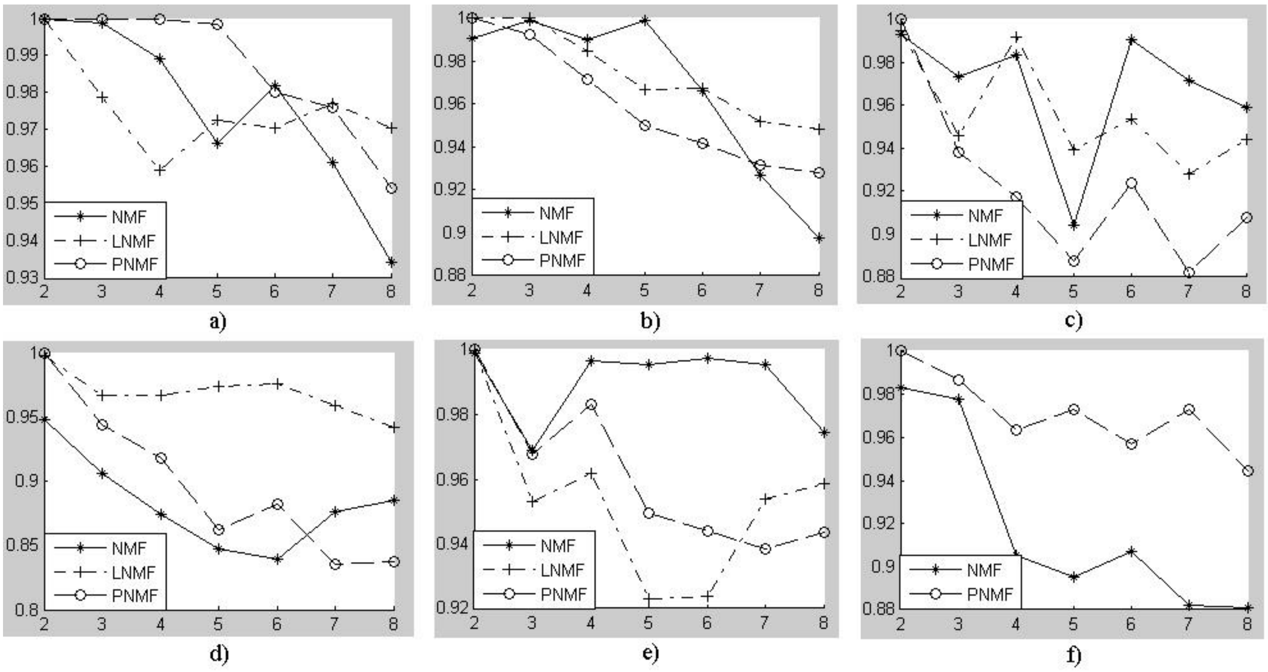


Figure 1: The cophenetic correlation coefficient versus decomposition rank for the *leukemia*, *medulloblastoma*, *colon*, *DLBCL*, *SRBCT*, and *prostate* data set, respectively.

4.3 Sample recognition

The second major experimental task involves the recognition of samples previously misclassified by the clustering K - means clustering approach. The correctly assigned samples were all gathered to form a training set $\mathbf{V}_{tr}$, while a test set $\mathbf{V}_{te}$ comprises all wrongly clustered samples associated to the minimum clustering error. The NMF methods were only applied to the training set. Thus, there is a big difference between clustering and recognition tasks in terms of training. While in the latter case the NMF uses all samples to extract relevant information, for the recognition task, NMF employs only a subset of this statistics, more precisely, those associated to the correctly clustered samples. Let us denote the final decomposition factors corresponding to the training set by $\mathbf{W}_{tr}$ and $\mathbf{H}_{tr}$, respectively. To form a training feature vector, the training data were projected into the inverse of $\mathbf{W}_{tr}$, i.e., $\mathbf{F}_{tr} = \mathbf{W}_{tr}^{-1} * \mathbf{V}_{tr}$. Similarly, the test feature vectors are formed as $\mathbf{F}_{te} = \mathbf{W}_{tr}^{-1} * \mathbf{V}_{te}$. We should note that, unlike the clustering procedure where the clustering results depend on $\mathbf{H}$, the sample recognition is solely affected by $\mathbf{W}$. By projecting the data into the inverse of $\mathbf{W}$, $\mathbf{F}_{tr}$ comprises mixed positive and negative values. To keep the non-negativity constraints we have also ran experiments by projecting the data into the $\mathbf{W}$. However, the classification results were much lower in this case.

A simple maximum correlation classifier (MCC) relying on the minimum Euclidean distance between vectors was chosen as distance measurement. The distance from any $\mathbf{h}_{te}$ to any $\mathbf{h}_{tr}$ is expressed as $\|\mathbf{h}_{te} - \mathbf{h}_{tr}\|^2 = -2g_c(\mathbf{h}_{te} + \mathbf{h}_{te})^T\mathbf{h}_{te}$, where $g_c(\mathbf{h}_{te}) = \mathbf{h}_{tr}^T\mathbf{h}_{te} - 1/2\|\mathbf{h}_{tr}\|^2$ is a linear discriminant function of $\mathbf{h}_{te}$. By computing $\mathcal{Q}$ linear discriminant functions, a test sample is assigned to the label according to the

$$MCC = \text{argmax}_c\{g_c(\mathbf{h}_{te}\} \tag{7}$$

In other words, the label of the training feature vector corresponding to the maximum correlation is associated to the test feature vector (test sample). The recognition results are tabulated

Table 2: Sample recognition results with MCC.

Method	Data					
	Leukemia	Medulloblastoma	Colon	DLBCL	SRBCT	Prostate
NMF	-	50	57.14	40	**55.55**	**28**
LNMF	0	66.66	86.20	71.42	64.28	-
PNMF	-	**25**	**44.44**	**33.33**	66.66	47.05

in Table 2. A recognition error of 100% indicates no improvement over the misclustered samples. LNMF wrongly clustered only one *leukemia* sample that have been finally correctly recognized by the recognition procedure. Four misclustered *medulloblastoma* samples are provided by the PNMF, while three of them were also further correctly classified by the recognition procedure. An impressive improvement was obtained in the case of NMF for the *prostate* data set. While the clustering strategy leaded to 25 mislabelled samples out of 102, the recognition procedure had further reduced those samples to 7.

5 Conclusions

Microarray analysis typically involves tens of samples where each sample is formed of thousands of values (genes). This recasts into a high-dimensional matrix. The analysis of such matrix is associated in the computer science community with the small size problem negatively affecting both the clustering and recognition accuracy. Wisely reducing data dimension is one solution to this issue. This paper addressed data dimensionality reduction applied to clustering and recognition tasks by three non-negative decomposition variants, NMF, LNM, and PNMF. The novelty of this work resides in the following: (a) we have carried out a systematic comparison in terms of clustering of these NMF algorithms involving six gene expression data sets publicly available; (b) we have shown that, unlike the clustering approach where the sample assignment depends only on the relative (maximum) values in each column of $\mathbf{H}$, strategy that failed to discover the correct class for some certain methods and data sets, by employing a K - means clustering approach the correct class discovery is always guaranteed for sufficient number of runs.

Concluding, the experiments indicate superior performance for the standard NMF over the other two, leading to the lowest clustering errors. The standard NMF was closely followed by SVD and PNMF. The nonlinear NMF approach (PNMF) outperforms NMF for only one data set, i.e. *prostate.*

Bibliography

[1] Alter, O. et al.; Singular value decomposition for genome-wide expression data processing and modeling, *Proc. Nat. Acad. Sci. USA*, 97: 10101-10106, 2000.

[2] Liu L. et al; Robust singular value decomposition of microarray data, *Proc. Nat. Acad. Sci. USA*, 100: 13167-13172, 2003.

[3] Wall, M.E. et al; SVDMAN singular value decomposition analysis of microarray data, *Bioinformatics*, 17: 566-568, 2001.

[4] Kluger, Y. et al; Spectral biclustering of microarray data: Coclustering genes and conditions, *Genome Research*, 13: 703-716, 2003.

[5] Lee, D.D.; and Seung, H.S.; Learning the parts of the objects by non-negative matrix factorization, *Nature*, 401: 788-791, 1999.

[6] Brunet, J.P. et al; Metagenes and molecular pattern discovery using matrix factorization, *Proc. Nat. Acad. Sci. USA*, 101: 4164-4169, 2004.

[7] Kim, H.; and Park, H. (2007); Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis, *Bioinformatics*, 23: 1495-1502, 2007.

[8] Li, S.Z. et al; Learning spatially localized, parts-based representation, *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 207-212, 2001.

[9] Buciu, I. et al; Non-negative matrix factorization in polynomial feature space, *IEEE Trans. on Neural Nerworks*, 19: 1090-1100, 2008.

[10] Buciu, I., Non-negative Matrix Factorization, A New Tool for Feature Extraction: Theory and Applications, *Int J Comput Commun*, ISSN 1841-9836, Vol. 3, Supplement: Suppl. S, 3(S): 67-74, 2008.

[11] Golub, T.R. et al; Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, 286:531-537, 1999.

[12] Pomeroy, S. L. et al; Prediction of central nervous system embryonal tumour outcome based on gene expression, *Nature*, 415:436-442, 2002.

[13] Alon, U. et al; Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Nat. Acad. Sci. USA*, 96: 6745-6750, 1999.

[14] Shipp, M.A. et al; Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning, *Nature Medicine*, 8: 68-74, 2002.

[15] Khan, J. et al; Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature Medicine*, 7: 673-679, 2001.

[16] Singh, D. et al; Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell*, 1: 203-209, 2002.

[17] Yang, K. et al; A stable gene selection in microarray data analysis, *BMC Bioinformatics*, 7: 228, 2006.

INT J COMPUT COMMUN, ISSN 1841-9836
9(1):24-36, February, 2014.

The Parallel One-way Hash Function Based on Chebyshev-Halley Methods with Variable Parameter

M. Nouri, M. Safarinia, P. Pourmahdi, M.H. Garshasebi

Mahdi Nouri
Department of Electrical Engineering,
Iran University of Science and Technology (IUST), Tehran, Iran
Mnouri@mtu.edu

Mahyar Safarinia, Payam Pourmahdi
Electrical Engineering Department
A.B.A Institude of Higher Education, Qazvin, Iran
Mahyar1986@gmail.com, payamict67pm@gmail.com

Mohammad Hossein Garshasebi
Communication Engineering Department
Basir Institude of Higher Education, Qazvin, Iran
E-mail: Mh.garshasebi@chmail.ir

Abstract: In this paper a parallel Hash algorithm construction based on the Chebyshev Halley methods with variable parameters is proposed and analyzed. The two core characteristics of the recommended algorithm are parallel processing mode and chaotic behaviors. Moreover in this paper, an algorithm for one way hash function construction based on chaos theory is introduced. The proposed algorithm contains variable parameters dynamically obtained from the position index of the corresponding message blocks. Theoretical analysis and computer simulation indicate that the algorithm can assure all performance requirements of hash function in an efficient and flexible style and secure against birthday attacks or meet-in-the-middle attacks, which is good choice for data integrity or authentication.

Keywords: Hash function; Chebyshev-Halley methods; Two-dimensional coupled map lattices; Spatiotemporal chaos; Chaotic nonlinear map; variable parameter.

1 Introduction

By the rapid development of Internet, the security is essential for modern communication [1][2].

Hash functions play an essential role in many areas of cryptographic applications such as digital signature, random number generation, data source authentication, key update and derivation, message authentication code, integrity protection and malicious code recognition and, etc. Generally, hash functions can be classified into two categories [3,4]: unkeyed hash functions with a single input parameter (a message) for data integrity, and keyed hash functions, usually known as message authentication code (MAC), with two distinct inputs. Conventional hash functions, such as MD5 and SHA, involve logical operations or multi-round iterations of some available ciphers. Although each step of the former is simple, the number of processing rounds could be massive even if the message is very short. Recently, spatiotemporal chaos has been attracting more and more interests among researchers engineering. After the conventional Hash function such as MD5, SHA was successfully attacked, the research on the design of secure and efficient Hash function remains a research hotspot. Compared with simple chaotic maps, spatiotemporal chaos possesses two additional advantages for cryptographic purpose. Due to the finite computing precision, chaotic orbits will eventually become periodic.

In particular, the period of chaotic orbits generated by a system with a large number of chaotic coupled oscillators is too long to be reached in practical communications. The period of spatiotemporal chaos is longer than that of simple chaotic maps [7]. Therefore, periodicity can be practically avoided in spatiotemporal chaotic systems [8,29,30]. Chaotic systems have vital characteristics like ergodic, mixed and sensitive which are so important in this area [3,5]. Some algorithms for one-way Hash function based on chaotic map have already been brought forward [11,12]. Most algorithms for chaos-based hash function proposed by existing articles are based on dissipative chaotic systems. But the dissipative chaotic systems can lead to many hidden threats in the practical application for secure communication because of their potential warning. When the dissipative chaotic systems were used in the application of encryption, if the attacker gets a continuous of plaintext-ciphertext pairs, and the length of ciphertext meets certain conditions, the attacker cannot predict by reconstructing through ciphertext without attractor structure for the conservative chaotic systems [5,13,14,31,32]. So the conservative chaotic systems with high security are ideal model in cryptography application.

In this paper the aim is to design an unkeyed hash function based on spatiotemporal chaos, which has high efficiency. In this algorithm the entire message blocks perform some rounds of iteration through Standard map. One output of the iterations of each round determines one of the initial values of the next round is the output of previous block. The method of this design enhances the diffusion of Hash function, and overcomes the inherent defects of dissipative chaotic systems. So the Hash algorithm has a higher security. Theoretical analysis and computer simulation show that this algorithm has good effect in Anti-conflict and Avalanche effect. This algorithm is easy to express and to satisfy all the performance requirements of Hash function in an efficient and flexible manner.

2 Feasibility of Hash Function Construction Based on Proposed

2.1 Nonlinear Equations Based on Newtons Method

Non-linear equations are an important and basic method [15], which converges quadratically. A family of third-order methods, called Chebyshev Halley methods [16], is written as

$$x_{n+1} = x_n - \frac{f(x_n)}{f\prime(x_n)} \tag{1}$$

where

$$L_f(x_n) = \frac{f\prime\prime(x_n)\, f(x_n)}{f\prime(x_n)^2} \tag{2}$$

This family includes the classical Chebyshevs method (CM), ($\alpha = 0$) , Halleys method (HM) ($\alpha = 1/2$) and Super-Halley method (SHM) ($\alpha = 1$)) (for the details of these methods, see [17-18] or a recent review [19]).In order to improve the local order of convergence, Grau and Daz-Barrero[20] propose an improvement of Chebyshevs method with fifth-order convergence

$$\tilde{x}_{n+1} = x_n - \left(1 + \frac{f\prime\prime(x_n)\left(f(x_n) + f(x_{n+1})\right)}{2f\prime(x_n)^2}\right)\frac{f(x_n) + f(x_{n+1})}{f\prime(x_n)} \tag{3}$$

Analysis of Convergence shows that the error convergence can be modeled with sixth order of error [10].

$$\tilde{e}_{n+1} = \left[(6-4\beta)\,c_2^2 - 3c_3\right]\left[2(1-\alpha)\,c_2^2 - c_3\right] e_n^5 + \mathrm{O}\left(e_n^6\right) \tag{4}$$

2.2 The Chebyshev-Halley Method

The family includes the classical Chebyshevs method (CM) ($\alpha = 0$) , Halleys method (HM) ($\alpha = 1/2$) and Super-Halley method (SHM) ($\alpha = 1$) (for the details of these methods, see [9,10] or a recent review [11]).Here the logistic map is taken as the local map, given as

$$f(x) = \mu x(1-x) \tag{5}$$

Where $x_i, y_i \in [0,1]$ and μ , α are the controlled variables. The map is nonlinear and the parameter ensures that the map runs in a chaotic status .The chaotic nonlinear map also has the same belongings to proposed map that are fit for composing Hash function. The form of the map is complicated and the equation involved is nonlinear. Figure 1 shows the simulation of the chaotic nonlinear map iterating 64 times with the initial values $x_0, y_0 = 0.3423$ and parameter $\mu = 60.5$. The map has some properties, which are appropriate for constructing the Hash function, such as initial value sensitivity and parameter sensitivity, with variable parameter valued in the interval (0,120). Figure 1 & 2 displays the chaotic iteration property with variable parameter first Iteration valued in the interval (0,64), which initial values are $x_0, y_0 = 0.3423$.

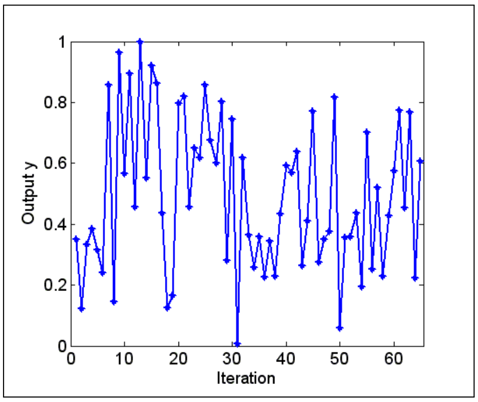


Figure 1: b

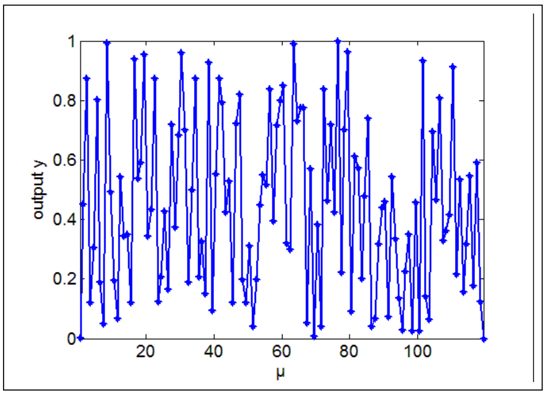


Figure 2: b

Figure 1 & 2 : Iteration property with changeable parameter First Iteration and μ and $x_0, y_0 = 0.3423$

3 Hash Function Based on Standard

Message expansion is significant and necessary; because it greatly improves the sensitivity of each bit in original message to the final Hash value [14]. The plaintext is an arbitrary message that is conveyed in a matrix M, for simple enlightenment of the extended message. Assume that M is a $N \times 16$ plain message matrix, each element with a size of 32 bits.

1) Padding the message: The purpose of padding is to ensure the padded message being a multiple of 128 bits. Suppose the total length of message is W bytes, compute d = (W mod 64), $0 \leq d \leq 12$. Pad as follows: if $12 \leq d < 16$ then pad 12d bytes, otherwise pad 28d bytes, the bytes been padded come from the head of message. The last four bytes are padded with message length. This method ensure at least one byte head of message been padded.

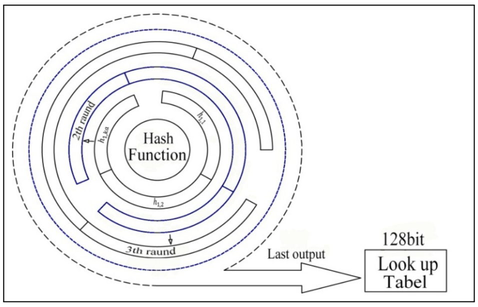


Figure 3: a

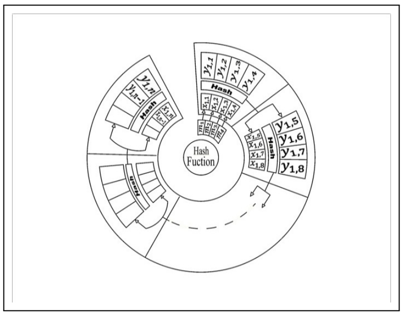


Figure 4: b

2) Parsing the padded message : The padded message is parsed into N 128-bit blocks, $M_1, M_2, \ldots, M_N$, $M_j (1 \leq j \leq N)$ is parsed into four 32-bit words

$$M_j = [m_{j,1}, m_{j,2}, m_{j,3}, m_{j,4}] \tag{6}$$

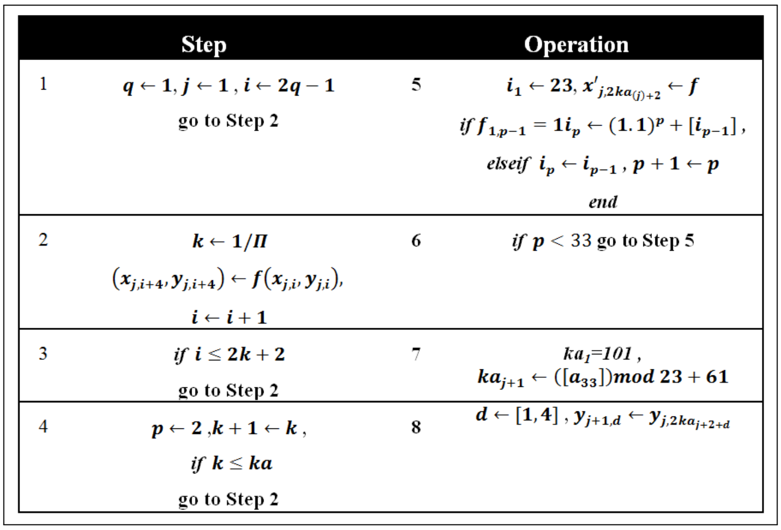
Figure 3 & 4 ; The whole algorithm ; a) All iterations of hash function structure . b) First iteration of hash function .

Input of the algorithm can have an input of arbitrary length output of the algorithm has a fixed length of 128 bits. Give a message M with length L. Take each letter of M as a plaintext block. Transform each plaintext block to the corresponding ASCII numbers; the ASCII numbers create the x_j, y_j which are the inputs of chaotic nonlinear map. Compression function inputs consider 16 lattice spaces. Let the initial iterative value of these inputs are:

$$\begin{cases} x_{j,1} = \left(\frac{m_{j,1}}{10^3}\right) + \left(\frac{m_{j,2}}{10^6}\right) + \left(\frac{m_{j,3}}{10^9}\right) + \left(\frac{m_{j,4}}{10^{12}}\right) \\ x_{j,2} = \left(\frac{m_{j,5}}{10^3}\right) + \left(\frac{m_{j,6}}{10^6}\right) + \left(\frac{m_{j,7}}{10^9}\right) + \left(\frac{m_{j,8}}{10^{12}}\right) \\ x_{j,3} = \left(\frac{m_{j,9}}{10^3}\right) + \left(\frac{m_{j,10}}{10^6}\right) + \left(\frac{m_{j,11}}{10^9}\right) + \left(\frac{m_{j,12}}{10^{12}}\right) \\ x_{j,4} = \left(\frac{m_{j,13}}{10^3}\right) + \left(\frac{m_{j,14}}{10^6}\right) + \left(\frac{m_{j,15}}{10^9}\right) + \left(\frac{m_{j,16}}{10^{12}}\right) \end{cases}$$

From $\{x_i\}$ and $\{y_i\}$, 4 groups of (y_i) can be reached. Determine the 32-bits Hash value by the position of the coordinates (y_i) falling into the region of [0, 1), then, finally, juxtaposes these bits from left to right to get a 128-bit hash value.[28]

TABLE 1
Algorithm for generating the hash



Step		Operation	

1	$q \leftarrow 1, j \leftarrow 1\ , i \leftarrow 2q-1$
	go to Step 2
		5	$i_1 \leftarrow 23,\ x'_{j,2ka_{(j)}+2} \leftarrow f$
			if $f_{1,p-1} = 1 i_p \leftarrow (1.1)^p + [i_{p-1}]$,
			elseif $i_p \leftarrow i_{p-1}$, $p+1 \leftarrow p$
			end

2	$k \leftarrow 1/\Pi$ $(x_{j,i+4}, y_{j,i+4}) \leftarrow f(x_{j,i}, y_{j,i})$, $i \leftarrow i+1$	6	*if* $p < 33$ go to Step 5
3	*if* $i \leq 2k+2$ go to Step 2	7	ka_1=101 , $ka_{j+1} \leftarrow ([a_{33}])mod\ 23 + 61$
4	$p \leftarrow 2\ , k+1 \leftarrow k$, *if* $k \leq ka$ go to Step 2	8	$d \leftarrow [1,4]$, $y_{j+1,d} \leftarrow y_{j,2ka_{j+2+d}}$

4 Performance Analysis

The proposed algorithm is used to perform hash simulation under the following five kinds of conditions:

Condition 1: A hash function is a fundamental building block of information security and plays an important role in modern cryptography. It takes a message as input and produces an output referred to as a hash value. Generally, hash functions can be classified into two categories: unkeyed hash functions for data integrity, and keyed hash functions, usually known as message authentication code (MAC).

Condition 2: Change the first character A in the original message to D.

Condition 3: Change the word function in the original message to functions.

Condition 4: Change the full stop at the end of the original message to a comma.

Condition 5: Add a blank space to the end of the original message.

The corresponding hash values in hexadecimal format are given as follows, followed by the equivalent number of different bits compared with the hash value obtained under Condition 1:
Condition 1 : 0484FCD104D3614104D5A185002832E9
Condition 2 : 3D87661B4F87F263FAEF3DAD5E7AAB
Condition 3 : 96A08482053A5537316D50181DFB7640
Condition 4 : D700EAE9701A4A55680C96A133F3B023
Condition 5 : 17F91FD2D4C22388C74C0811F5874687

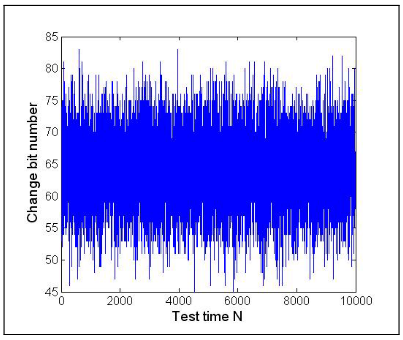


Figure 5 : Distribution of changed bit number B_i

This kind of test is performed N times, and the corresponding distribution of changed bit number is plotted in Figure 5 , where N = 10,000. Obviously, the changed bit number corresponding to 1 bit changed message concentrates around the ideal changed bit number, i.e., 64 bits. It indicates that the algorithm has very strong capability for diffusion and confusion.

4.1 Statistical Analysis of Diffusion and Confusion

Confusion and diffusion are two fundamental design criteria for encryption algorithms, including hash functions. Shannon introduced diffusion and confusion in order to hide message redundancy [18,19]. Hash function, like encryption system, requires the plaintext to diffuse its influence into the whole Hash space. This means that the correlation between the message and the corresponding Hash value should be as small as possible. Diffusion means spreading out the influence of a single plaintext symbol over many ciphertext bits so as hiding the statistical structure of the plaintext. Confusion means the use of transformations to complicate the dependence of ciphertext statistics on plaintext statistics. In the hash value in binary format each bit can be only 0 or 1. Therefore, the ideal diffusion effect should be that any tiny change in the initial condition leads to a 50% changes, probability of each bit. Usually six statistics are defined as follows:

Minimum changed bit number:

$$B_{min} = min\left(\{B_i\}_1^N\right) \tag{7}$$

Maximum changed bit number:

$$B_{max} = max\left(\{B_i\}_1^N\right) \tag{8}$$

Mean changed bit number:

$$\bar{B} = \frac{1}{N}\sum_{1}^{N} B_i \tag{9}$$

Mean changed probability:

$$P = \frac{\bar{B}}{128} \times 100 \tag{10}$$

Standard variance of the changed bit number:

$$\Delta B = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(B_i - \bar{B}\right)^2} \tag{11}$$

Standard variance:

$$\Delta P = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{B_i}{128} - P\right)^2} \times 100 \tag{12}$$

Where N is the total number of tests andB_i is the number of changed bits in the ith test. The following diffusion and confusion test has been performed that: A paragraph of message is randomly chosen and the corresponding hash value is generated. Then a bit in the message is randomly selected and toggled and a new hash value is generated. Finally, the two hash values are compared with each other. This kind of test is performed N times, and the corresponding distribution of changed bit number is plotted in Fig. 8, where N = 10,000. perceptibly the changed bit number corresponding to 1 bit changed message concentrates around the ideal changed bit number, i.e., 64 bits. It signifies that the algorithm has very strong capability for diffusion and confusion.The test results in N = 256, 512, 1024, 2048, 10,000 are listed in Table 1 respectively. Based on the results, the following conclusion was found that, the mean changed bit number $\bar{B}$ and the mean changed probability P are both very close to the ideal value 64 bits and 50% . While ΔB and ΔP are extremely small, which indicate the diffusion and confusion capability are very stable. Therefore a small difference in the plaintext will cause great changes in the hash value, which donates to the high plaintext-sensitivity of our hash function. This property is important in maintaining the secrecy against statistical attacks.

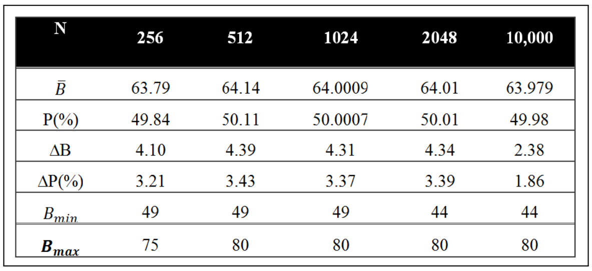
4.2 Collision Analysis

Birthday-attack

Collision resistance and birthday-attack are lying to each others roots. Both are derived from the probability problem that two random input data are found to hash to the same value. The results of the proposed algorithm are shown in TABLE 2:

TABLE 2

Statistical performance of proposed algorithm for 128 bits outputs with $\alpha = 0$ and $\mu = 60.5$



N	256	512	1024	2048	10,000
$\bar{B}$	63.79	64.14	64.0009	64.01	63.979
P(%)	49.84	50.11	50.0007	50.01	49.98
ΔB	4.10	4.39	4.31	4.34	2.38
ΔP(%)	3.21	3.43	3.37	3.39	1.86
B_{min}	49	49	49	44	44
$\boldsymbol{B_{max}}$	75	80	80	80	80

Collision resistance and birthday-attack are lying to each others roots. Both are derived from the probability problem that two random input data are found to hash to the same value. Given a function f the goal of the attack is to find two different inputs $f(x)$ such that $f(x_1) = f(x_2)$. Such a pair x_1, x_2 is called a collision. The method used to find a collision is simply to evaluate the function f for different input values that may be chosen randomly or pseudo randomly until the same result is found more than once. Because of the birthday problem, this method can be rather efficient. Specifically, if a function $f(x)$ yields any of H different outputs with equal probability and H is sufficiently large, then we expect to obtain a pair of different arguments x_1, x_2 with $f(x_1) = f(x_2)$after evaluating the function for about 1.25H different arguments on average. From a set of H values we choose n values uniformly at random there by allowing repetitions. Let $p(n; H)$ be the probability that during this experiment at least one value is chosen more than once. This probability can be approximated as

$$p(n; H) \approx 1 - e^{-\frac{n(n-1)}{2H}} \approx 1 - e^{-\frac{n^2}{2H}} \tag{13}$$

Let $n(p; H)$ be the smallest number of values we have to choose, such that the probability for finding a collision is at least p. By inverting this expression above, we find the following approximation

$$\mathrm{n}\left(\mathrm{p};\mathrm{H}\right) \approx \sqrt{2Hln\frac{1}{1-p}} \tag{14}$$

and assigning a 0.5 probability of collision we arrive at:

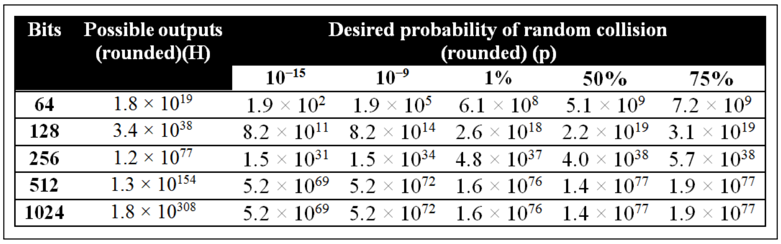
$$\mathrm{n}\left(0.5;\mathrm{H}\right) \approx 1.1774\sqrt{H} \tag{15}$$

Let $Q\left(H\right)$ be the expected number of values, it must be chosen before finding the first collision. This number can be approximated by

$$\mathrm{Q}\left(\mathrm{H}\right) \approx \sqrt{\frac{\pi}{2}H} \tag{16}$$

As an example, if a 64-bit hash is used, there are approximately 1.8×10^{19} different outputs. If these are all equally probable, then it would take only approximately 5.1×10^{9} attempts to generate a collision using brute force. This value is called birthday bound [4] and for n-bit codes it could be computed as 2^{n-1} [3]. Table shows number of hashes $n\left(p\right)$ needed to achieve the given probability of success, assuming all hashes are equally likely. The results of the proposed algorithm are shown in TABLE 3:

TABLE 3
probability of random collision of proposed algorithm for various bits outputs



Bits	**Possible outputs (rounded)(H)**	**Desired probability of random collision (rounded) (p)**				
		10^{-15}	**10^{-9}**	**1%**	**50%**	**75%**
64	1.8×10^{19}	1.9×10^{2}	1.9×10^{5}	6.1×10^{8}	5.1×10^{9}	7.2×10^{9}
128	3.4×10^{38}	8.2×10^{11}	8.2×10^{14}	2.6×10^{18}	2.2×10^{19}	3.1×10^{19}
256	1.2×10^{77}	1.5×10^{31}	1.5×10^{34}	4.8×10^{37}	4.0×10^{38}	5.7×10^{38}
512	1.3×10^{154}	5.2×10^{69}	5.2×10^{72}	1.6×10^{76}	1.4×10^{77}	1.9×10^{77}
1024	1.8×10^{308}	5.2×10^{69}	5.2×10^{72}	1.6×10^{76}	1.4×10^{77}	1.9×10^{77}

The state of proposed is related to each message bit. By iterations, significant changes are obtained at the final state even if there is only a one-bit change in the message. According to the above analysis, the proposed algorithm is secured against statistical attacks. For birthday-attack, the security of the cryptosystem is determined by the length of the hash value, which is 128-bit in this proposed function. According to the definition of birthday-attack [4, 22, 23] , the attack difficulty is 2^{64}

Meet-in-the-middle Attack

Meet-in-the-middle attack [10,11] means to find a contradiction through looking for a suitable substitution of the last plaintext block. In this paper we focus on random routing solutions because their consciousness is simple. The proposed algorithm is principally the same analyzed in [1,9].To clarify its operations, let ns represent the current attack represents the sections within the attacker belongs of M_i, and n_i represents the attacker in $\Phi\left\{n_s\right\}$which is in the straight path between ns and attacked n0. The basic idea of the random routing algorithm considered in this paper is very simple as it defines the following two phases:

1. Basic parameter of the principal phase is the probability p^*Actually, the current M_i immediately selects n_s, which represents the M_i in the shortest block near the attacked n_0, as the next M_i with probability p^*. If this occurs, then the algorithm stops here, otherwise, it performs the second phase.

2. In the second phase, the current M_i randomly selects any of the attackers in $\Phi\left\{n_s\right\}$as next one. Note that according to such algorithm M_i selects n_s, that is the attack in the shortest path with probability

$$P_S = p^* + \left(1-p^*\right) / \left|\Phi\left\{n_s\right\}\right| \tag{17}$$

where $\Phi\left\{n_s\right\}$ is the number of next M_i of n_s ; whereas any given one belonging to $\Phi\left\{n_s\right\}-\left\{n_0\right\}$ is selected with probability$\left(1-p^*\right)/\left|\Phi\left\{n_{CR}\right\}\right|$. Note that, the higher$p^*$, the higher the probability that packets will follow the shortest path, and, vice versa. Accordingly, on the one hand, if p^* decreases it can be expected longer end-to-end routes between packets source and the attacked n_0 and therefore larger energy consumption. However, on the other hand when p^* becomes smaller, it is more difficult for the attacker to intercept messages sent by the actual reader of target, and therefore, the level of security increases. In order to evaluate the impact of the countermeasures to the meen-in-the-middle attack, the output of the reading process is compared in normal conditions and when the meen-in-the-middle attack is performed. To evaluate the attack impact of the disturbing bits on the performance of the proposed scheme, in Figure 6, the bit error rate is shown versus the normalized amplitude of the disturbing bits. Note that when the normalized bits reach the unitary value, bit error probability is 0.5, which means that the meen-in-the-middle attack is not effective. Observe, however, that the increase in energy expenditure goes as the square of the disturbing bits.

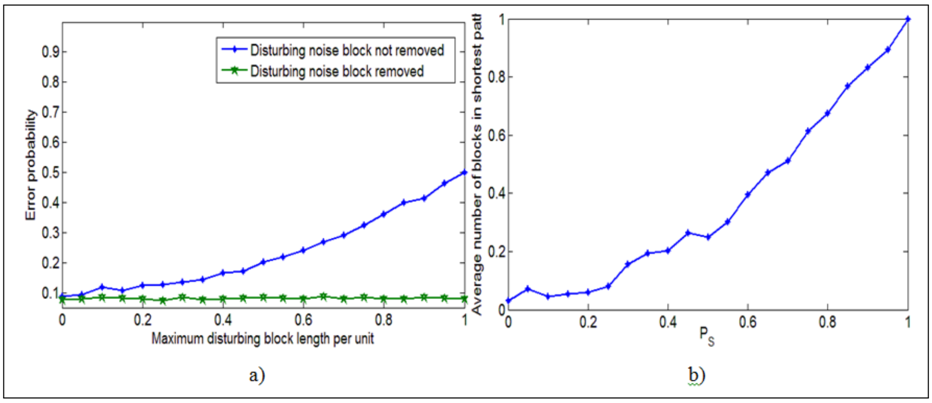


Figure 6 : a) Bit error probability versus the normalized length of the disturbing block, b) Block percentage of the shortest block included in the end-to-end path versus the probability p*

Accordingly, appropriate tradeoff must be identified to reduce energy consumption. In order to evaluate the effects of the adoption of random routing in Figure 6, it is shown the average number of blocks that are included in both the shortest block and in the block obtained by using random routing versus the value of the probabilityp^*. Obviously, this value increases as p^* increases and, therefore, the proposed scheme becomes less effective. However, it can be expected that as p^* increases, the energy consumption decreases. This is indeed demonstrated in Figure 9 where shows the average block length obtained by using random routing versus the value of the probabilityp^*. Note that as p^*increases, the average path length decreases. Note that Figure 7 have been obtained by randomly selecting the positions of the main and the attacked.

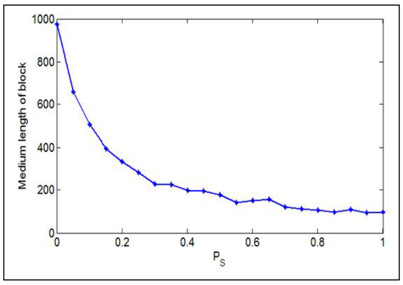


Figure 7 : Average Block length obtained by using random routing versus the value of the probability p*

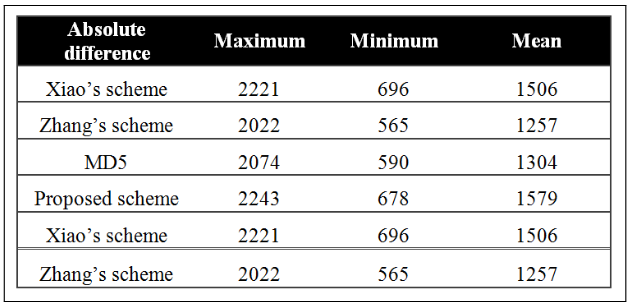
Collision Test

The following test has been performed for the quantitative analysis on collision resistance [24]: first, the hash value for a paragraph of randomly chosen message is generated and stored in ASCII format. Then a bit in the message is selected randomly and toggled. A new hash value is then generated. The two hash values are compared and the number of ASCII characters with the same value at the same location is counted. Moreover, the absolute difference between the two hash values is calculated by the following formula:

$$d = \sum_{i=1}^{N} t\left(e_i\right) - t\left(e\prime_i\right) \tag{18}$$

Where e_i and $e\prime_i$ is the$i_t h$ASCII character of the original and the new hash value, respectively. The function t( ) converts the entries to their equivalent decimal values. This kind of collision test is performed 10,000 times. The maximum, mean, and minimum values of d are listed in Table 4 . The distribution of the number of ASCII characters with the same value at the same location in the hash value is given in Table 6. Notice that the maximum number of equal characters is only 2 and the collision probability is very low.

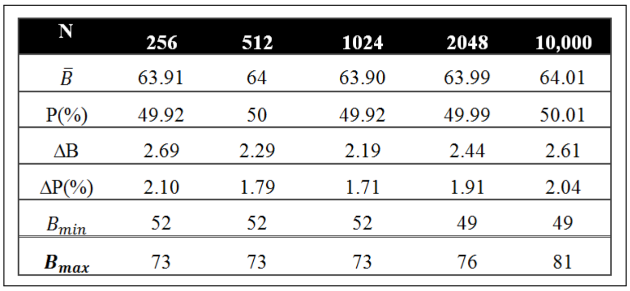
TABLE 4
Absolute difference of two hash value



Absolute difference	Maximum	Minimum	Mean
Xiao’s scheme	2221	696	1506
Zhang’s scheme	2022	565	1257
MD5	2074	590	1304
Proposed scheme	2243	678	1579
Xiao’s scheme	2221	696	1506
Zhang’s scheme	2022	565	1257

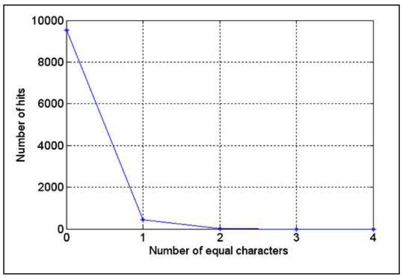
Due to the progresses in technology, NIST plans to replace SHA-1 to longer and stronger hash functions (SHA-224, SHA-256, SHA-384 and SHA-512) by the year 2010 [25], as long hash values are needed in the future. To produce a long hash value, two simple modifications are introduced in the proposed algorithm: increase the size of proposed or extract more bits. For example, 256,512,1024-bit hash values are obtained by using a proposed model with extracting 8 hexadecimal numbers, by extracting 16 hexadecimal numbers and extracting 32 hexadecimal numbers from each lattice value in the origin proposed respectively which is shown in Table 5.

TABLE 5
Statistical performance of proposed algorithm for different output length 512 bits



N	256	512	1024	2048	10,000
$\bar{B}$	63.91	64	63.90	63.99	64.01
P(%)	49.92	50	49.92	49.99	50.01
ΔB	2.69	2.29	2.19	2.44	2.61
ΔP(%)	2.10	1.79	1.71	1.91	2.04
B_{min}	52	52	52	49	49
$\boldsymbol{B_{max}}$	73	73	73	76	81

Figure 8 : Distribution of the number of ASCII characters with the same value at the same location in the hash value.



Hereinto Statistic Analysis of Diffusion and Confusion for Variable Parameter is considered,is the controlled variable. The certain critical value of μ is 0.7 . The followed is some conclusion on chaotic nonlinear map. By varying this parameter as can be seen in the Figure 9, the Statistic analysis of diffusion and confusion are near the same for changing this parameter and it means this algorithm has a stable manner in all variable parameters.

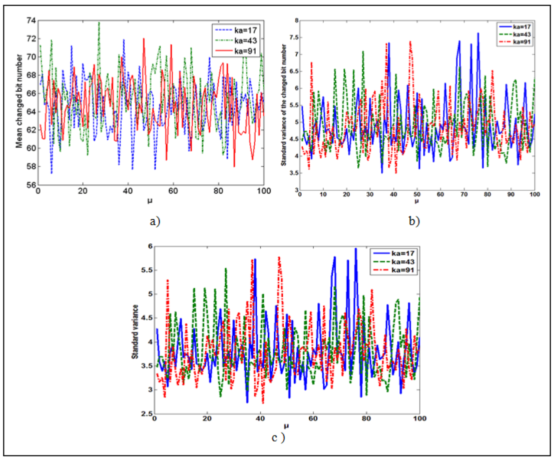
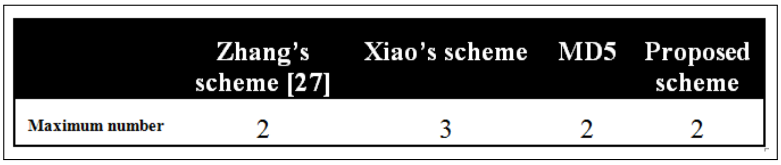


Figure 9 : a) Mean changed bit number for changing μ, b) Standard variance of the changed bit number for changing μ ,c) Standard variance for changing μ.

TABLE 6
Maximum number of equal characters at the same location in the hash value



	Zhang's scheme [27]	Xiao's scheme	MD5	Proposed scheme
Maximum number	2	3	2	2

Based on the result in Tables 1 and 6, the statistical performance of the proposed algorithm is as good as that of MD5. Moreover ΔB and ΔP of the proposed algorithm are smaller than those of MD5, which indicate that the result is more stable. Based on the data in Tables 4 and 6, the proposed algorithm has a bigger exact difference between two hash values than MD5. Meanwhile, the maximum numbers of equal character of MD5 and the proposed algorithm are the same, i.e., only two, which indicates the collision chance is very low.

Speed analysis

Since the proposed algorithm is based on a complex chaotic system, complicated computations are needed in producing a hash value and so this algorithm is slower than MD5 algorithms in [27]. Nevertheless, it still owns a sufficiently high hashing speed for practical use. The PROPOSED model is a kind of spatiotemporal chaos. Compared with the algorithm in [26], which is also based on spatiotemporal chaos, the proposed algorithm has a much higher efficiency. These two algorithms are applied to use Matlab which are running on a personal computer with a Pentium IV 2.66 GHz processor and 4 GB RAM. As monitored from these figures, the execution time of the proposed algorithm is much shorter, especially when the message is long.

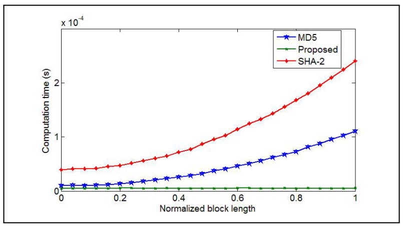


Figure 9: Computation time comparison between the proposed, MD5 and SHA-2 hash function for various normalized block length

5 Conclusion

Based on the Chebyshev Halley methods with variable parameters, a parallel Hash algorithm structure is proposed and analyzed. The algorithm alters the expanded message blocks into the equivalent ASCII code values. The two initial inputs and steps of iterations are generated by last round of iteration, which iterates the chaotic nonlinear map and wholly increases the rise influence of Hash function, and makes the final Hash value has high sensitivity to the initial values, increase the security of Hash function. The analysis designates that the algorithm can meet all the requisites of the Hash function efficiently. And the algorithm is easy to realize a swift and practical program to Hash function structure. The length of the final Hash value generated by this algorithm is 128 bits. The two primary inputs and steps of iterations are generated by last round of iteration, which iterates the chaotic nonlinear map and wholly increases the rise power of Hash function, and makes the final Hash value has high sensitivity to the initial values, increase the security of Hash function. Theoretical analysis and computer simulation show that the proposed algorithm presents numerous interesting features, such as high message, good statistical properties, collision resistance, and secure against meet-in-the-middle attacks that can assure the performance requirements of Hash function. Furthermore the proposed algorithm can present some extra advantages for having convenient controller by the variable parameters.

Bibliography

[1] Boris S. Verkhovsky; Information Assurance Protocols: Efficiency Analysis and Implementation for Secure Communication, *Journal of Information Assurance and Security*, 3(4): 263-269, 2008.

[2] B. Surekha G.N. Swamy, K. SrinivasaRao, A. Ravi Kumar; A Watermarking Technique based on Visual Cryptography Information Assurance Protocols, *Journal of Information Assurance and Security*, 470-473, 2009.

[3] W. Luo, D.C. et al, Hashing via finite field, *Information Sciences*, 176: 2553-2566, 2006

[4] A. Menezes, P. van Oorschot , S. Vanstone, *Handbook of applied cryptography*; CRC Press , 1996.

[5] X. Wang, D. Feng, X. Lai, H. Yu ; Collisions for hash functions MD4 , MD5 ; HAVAL-128 and RIPEMD, Rump Session of Crypto04 E-print, 2004.

[6] X. Wang, H. Yu ; How to break MD5 and other hash functions, *Proceedings of Eurocrypt05*, Aarhus ; Denmark, 19-35, 2005.

[7] S. Wang, W. Liu, H. Lu ; et al., Periodicity of chaotic trajectories in realizations of finite computer precisions and its implication in chaos communications, *International Journal of Modern Physics B*, 18: 2617-2622, 2005.

[8] P. Li, Z. Li, W.A. Halang; et al. G. Chen, A multiple pseudorandom-bit generator based on a spatiotemporal chaotic map, *Physics Letters A*, 349: 467-473, 2006.

[9] SPengFei, QiuShui-Sheng , One-way hash functions based on iterated chaotic systems , *IEEE conference proceedings: communications, circuits and systems, 2007. ICCCAS 2007. International conference on*, 11-13 July; p. 1070-74 , 2007.

[10] Schmitz R., Use of chaotic dynamical systems in cryptography", *Journal of the Franklin Institute*, 38(9):429-441, 2002.

[11] Deng S, Liao X F, Xiao D, A Parallel Hash Function Based on Chaos, *Computer Science*, 35(6): 217- 219, 2008.

[12] Wang X M, Zhang J S and Zhang W F, One way Hash function construction based on the extended chaotic map s switch , *Chin. Phys. Sin*, 52(11): 2737-2742, 2003.

[13] Gao J S, Sun B Y, Han W , Construction of the control orbit function based on the chaos theory, *Electric machines and control*, 2: 150-155, 2002.

[14] Parliz U, Junge L, Kocarev L , Synchronization-based parameter estimation from time series , *PhsRevE*, 4(6): 6253-6259, 1996.

[15] A.M. Ostrowski, *Solution of Equations in Eucilidean and Banach Space*, third ed., Academic Press, New York, 1973.

[16] J.M. GutieÂ´rrez, M.A. Hernandez, A family of Chebyshev Halley type methods in Banach spaces, *Bull. Austr.Math.Soc.*, 55: 113-130, 1997.

[17] J.F. Traub, *Iterative Methods for Solution of Equations*, Prentice-Hall, Englewood Cliffs, NJ,1964.

[18] J.M. GutieÂ´rrez, M.A. Hernandez, An acceleration of Newtons method: super-Halley method, *Appl. Math. Comput.*, 117: 223-239 , 2011.

[19] S. Amat, S. Busquier, J.M. Gutierrez,Geometric constructions of iterative functions to solve nonlinear equations, *J. Comput. Appl. Math.*, 157: 197-205, 2003.

[20] M. Grau, J.L. Daz-Barrero, An improvement of the Euler-Chebyshev iterative method, *J. Math. Anal. Appl.*, 315: 1-7 , 2006.

[21] Jisheng Kou, Yitian Li, Xiuhua Wang, A family of fifth-order iterations composed of Newton and third-order methods, *Appl. Math. Comput.*, in press, doi:10.1016/j.amc..07.150 ,2006.

[22] Secure Hash Standard, Federal Information Processing Standards Publication (FIPS PUB) 180-2, 2002.

[23] Security Requirements for Cryptographic Modules, Federal Information Processing Standards Publication (FIPS PUB) 140-1, 2002.

[24] K. Wong, A combined chaotic cryptographic and hashing scheme, *Physics Letters A*, 307:292-298 , 2003.

[25] NIST Brief Comments on Recent Cryptanalytic Attacks on Secure Hashing Functions and the Continued Security Provided by SHA-1, 2004,
$http://csrc.nist.gov/hash_standards_comments.pdf$

[26] H. Zhang, X. Wang, Z. Li, D. Liu, One way Hash function construction based on spatiotemporal chaos, *Act a PhysicaSinica*, 54(9): 4006-4011 (in Chinese) , 2005.

[27] J. Zhang, X. Wang, W. Zhang, Chaotic keyed hash function based on feed forward feedback nonlinear digital filter, *Physics Letters A*, 362: 439-448 , 2007.

[28] D. Goldberg, D. Priest, What every computer scientist should know about floating-point arithmetic, *ACM Computing Surveys*, 23(1): 548 , 1991.

[29] M.Nouri, S.Abazari Aghdam, P.Pourmahdi and M.Safarinia, Analysis of a Novel Hash Function Based upon Chaotic Nonlinear Map with Variable Parameter, *Journal of Computer Science and Information Security (IJCSIS)*, 221-228, 2011.

[30] M.Nouri, A.Khezeli, A.Ramezani and A.Ebrahimi , Dynamic Chaotic Hash Function Based upon Circle Chord Methods , *Sixth International Symposium on Telecommunications (IST)*, 1044 - 1049, 2012.

[31] Nouri, M.; Farhangian, N.; Zeinolabedini, Z.; Safarinia, M., Conceptual authentication speech hashing base upon hypotrochoid graph, *Sixth International Symposium on Telecommunications (IST)*, 1136 - 1141, 2012.

[32] Nouri, M. ; Zeinolabedini, Z.; Farhangian, N. ; Fekri, N.; Analysis of a novel audio hash function based upon stationary wavelet transform, *2012 6th International Conference on Application of Information and Communication Technologies (AICT)*, 1 - 6, 2012.

INT J COMPUT COMMUN, ISSN 1841-9836
9(1):37-47, February, 2014.

Development of an Agent Based Specialized Multi-Lingual Web Browser for Visually Handicapped

R. Ponnusamy, M.S.S. Babu, T. Chitralekha

R. Ponnusamy*, M.S. Satish Babu
Dept of Computer Science & Engineering,
Madha Engineering College, Chennai - 600 069, India,
*Corresponding author: rponnusamy@acm.org

T. Chitralekha
Dept of Computer Science,
Pondicherry University, Pondicherry - 605 014, India

Abstract:
In the modern age everyone needs access to Internet; Visually handicapped are not an exception for that. SPECS (SPEcialized Computer System) is a system developed to give access to the visually handicapped. It has a Braille shell. The user can enter his spoken language. After the selection of the language the user can present his request to the browser through chosen language in Braille. The output generated by the browser is given out as voice message. The effectiveness of this system is measured based on number of requests processed, access speed and precision of the system.

Keywords: Visually handicapped, Braille shell, Internet, Multi-Lingual Web Browser.

1 Introduction

The impact of visual loss has profound implications for the person affected. The majority of blind people live in developing countries. The number is huge, also due to the sheer size of the population in developing countries. Especially in south and East Asia itself there exist 27% (1,590.80 Millions, 2002 Estimate) [1] [2] [3] of blindness. The first global estimate on the magnitude and causes of visual impairment was based on the 1990 world population data (38 million blind). This estimate was later extrapolated to 1996 world population as 45 million blind, and projected to 2020 world population as 76 million blind, indicating a twofold increase in the magnitude of visual impairment in the world by 2020 [4]. Further, the survey estimated that 3.9% comes under child blindness and incurable categories.

In the past decade, the Internet revolution throughout the computing world, catalyzed largely by the World Wide Web (WWW) [5], has enabled the widespread dissemination of information worldwide. However, much of this information is in English or in languages of Western origin. Presently, the Internet is positioned to be an international mechanism for communications and information exchange, the precursor of a global information superhighway. For this vision to be realized, one important requirement is to enable all languages to be technically transmissible via the Internet, so that when a particular society is ready to absorb Internet technology, the language capability comes prepackaged. The term "Multilingual Computing" refers to the use of computer applications in Indian languages. Traditionally, computer applications were based on English as the medium of interaction with the system. In India, when one attempts to use computers for education and literacy, one faces the problem of language where majority of the population that should get the benefit of Information Technology, does not speak English. This is a non-trivial multilingual information-processing problem.

There is an urgent need to recognize that the true burden of blindness has changed with the rapid pace of industrialization and technology, and must adopt these people for development.

Most legally blind people (70% of them across all ages, according to the Seattle Lighthouse for the Blind) do not use computers. Only small fractions of this population, when compared to the sighted community, have Internet access. There is urgent need to develop Information Technology Tools which can be used by blind people knowing only their own languages. Especially the regional visually handicapped people need special methods and tools for accessing the web. It is the responsibility of the technocrats to develop such technology. Today's research challenge is to give an optimal access to the computers and internet for the visually handicapped people in different languages.

In this paper an attempt has been made to design and develop a special system for visually handicapped people. A specialized browser using role based agents for regional visually handicapped users. Section 2 explains the working nature of Specialized Browser. Section 3 explains the components and architecture of the Specialized Browser. Section 4 explains the design and functionality of Tamil and English Braille keyboard. Section 5 explains the SPECS Machine Learning System. Section 6 explains the simulation experiment. Section 7 gives experiment results and discussion and Section 8 Concludes the paper.

2 Literature Survey

Stuart Goose, 2000 [18] presenting a new idea of creation of a hypertext document in both the visual and auditory realms. In this approach an intelligent agent that is able to convert HTML contents to VXML contents to provide voice services for text disabilities via web. Prior to interpreting HTML documents and separating contents, the contents for the conversion must be selected, however, there are no good solutions for selecting the desired group contents. If an audio document is designed straight from the author's intentions, it may correspond to the author making an explicit recording the user study presenting voice based html structure in audio: user satisfaction with audio hypertext of the document or pieces of the document. Patrick Roth [19] and his group project aims at providing sight handicapped people with alternative access modalities to pictorial documents. More precisely, our goal is to develop an augmented Internet browser to facilitate blind users access to the World Wide Web. The main distinguishing characteristics of this browser are 1.Generation of a virtual sound space into which the screen information is mapped; 2.Transcription into sounds not only of text, but also of images; 3. Active user interaction, both for the macro-analysis and micro-analysis of screen objects of interest; 4.Use of a touch-sensitive screen to facilitate user interaction. Several prototypes have been implemented, and are being evaluated by blind users.

Guan neng huang,2007 [20], designed a special web browser called eguidedog is designed for the visually impaired people. this web browser can extract the structure and the content of an html document and represent it in the form of audio. it helps the blind finding out information they concern more quickly. IBM Accessibility browser [21] provide an additional facility to access audio while enjoying a streaming video, visually impaired people can now select the play button by simply pressing a predefined shortcut key instead of searching in the content for buttons that control the video. Users can also adjust the volume of an individual source in order to identify and listen to different sound sources without losing track of the screen-reading software because of the sound of a video. The main problems with these work is accessing the multilingual content. Tim Morris [22] and his team have reported on a prototype screen reader that is intended to vocalize the information displayed on the LCD or LED screen of home or office equipment. Tadayoshi Fujiki, 2006 [23], Developed a new tool easy bar to cater the needs of the visually handicapped users. The functions of the Easy Bar are to change the size of web texts and images, to adjust the color, and to clear cached data that is automatically saved by the web browser.

3 About SPECS

The SPECS [6] [7] is the specialized browser for Visually Handicapped Users (VHU). This system allows the VHU to browse the restricted set of regional language and English WebPages. It receives the inputs through Braille/Normal Keyboard and gives the voice output in respective Language of choice. In order to browse the webpage the user must know the specific website in advance. Also, the trainer must train system by giving English website address and its equivalent regional language website through the Braille keyboard attached with the system. The system is able to read only the static websites. Further the user is not able to travel into the complete set of hyperlinks provided in the site and the present browser functionality is restricted to access the text messages alone. The trainer is not Visually Handicapped.

The Braille keyboard is designed to work in the two modes. One is the normal mode and another is command mode. The Visually Handicapped Users (VHU) can travel in the website through the command mode. The VHU can type the website address in normal mode of operation and in case of command mode the user can move from the hyperlink to anther hyperlink. This is the main difference between the previous work and the present work. Another main restriction in the present work is that the VHU cannot access the .gif, .bmp, .jpg fixed text contents. This is great challenge to the user. Also the user cannot feel the pictures that appear on the webpage. At present the SPECS is designed to accommodate only two languages. In future the complete design of SPECS is expected to incorporate all Indian languages.

4 Architecture and Component Functionality of SPECS

The overall architecture of SPECS is shown in the following figure 1. This architecture of SPECS consists of three layers. These are 1. SPECS Browser Layer 2. Multi-functional Agent Layer 3. Knowledge Base Layer. Further, the SPECS System Interface (IOCS) is built under Windows platform.

The Windows API, informally WinAPI, is Microsoft's core set of application programming interfaces (APIs) available in the Microsoft Windows operating systems. Developer support is available in the form of the Microsoft Windows SDK, providing facilities and tools necessary to build software based upon the Windows API and associated Windows technologies. Various wrappers were developed by Microsoft that took over some of the more low level functions of the Windows API, and allowed applications to interact with the API in a more abstract manner. Microsoft Foundation Class Library (MFC) wrapped Windows API functionality in C++ classes, and thus allows a more object oriented way of interacting with the API. The Active Template Library (ATL) is a template oriented wrapper for COM. The Windows Template Library (WTL) was developed as an extension to ATL, and intended as a lightweight alternative to MFC. Over and above such facilities the SPECS IOCS has been developed to cater the needs. The layered representation of this specialized browser is shown in the following Figure 2.

4.1 Brower Layer

The SPECS Browser is the general browser capable of browsing in regional language and in English language. Font availability is the big problem of this layer. This problem is solved through the Machine Learning System. The SPECS browser is designed using the Internet Explorer Active X component [13]. ActiveX is a framework for defining reusable software components in a programming language independently. Internet Explorer 7 was used to design this system. Software applications can then be composed from one or more of these components in order to provide their functionality. Active X controls (small program building blocks), can

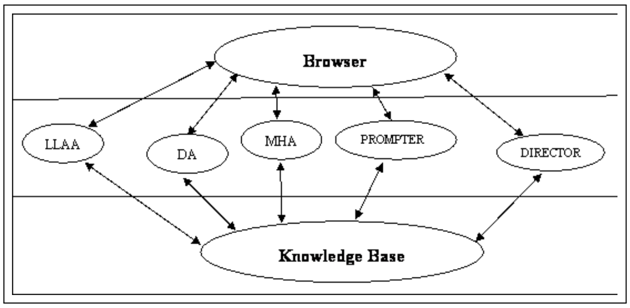


Figure 1: SPECS Architecture, LLAA – Language Learning Adaptation Agent, DA – Dialogue Agent, MHA – Message Handling Agent

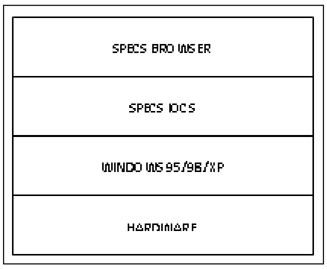


Figure 2: Layered Representation of Specialized Browser

serve to create distributed applications working over the Internet through web browsers. ActiveX controls can then be embedded into other applications. Internet Explorer also allows embedding ActiveX controls onto web pages.

Windows Internet Explorer [14] (formerly Microsoft Internet Explorer; commonly abbreviated to IE), is a series of graphical web browsers developed by Microsoft and included as part of the Microsoft Windows line of operating systems. Since its first release, Microsoft has added features and technologies such as basic table display (in version 1.5); XMLHttp Request (in version 5), which aids creation of dynamic web pages; and Internationalized Domain Names (in version 7), which allow Web sites to have native-language addresses with non-Latin characters.

Internet Explorer has introduced an array of proprietary extensions to many of the standards, including HTML, CSS, and the DOM. This has resulted in a number of web pages that appear broken in standards-compliant web browsers and has introduced the need for "quirks mode" rendering improper elements meant for Internet Explorer in such browsers. Considering these advantages the IE 7 has been chosen as the suitable component for designing the SPECS browser.

4.2 Multi-Function Agents Layer

This layer performs different functions such as language learning, VHU interaction, error messaging, voicing and VHU direction. The study of multi-agent systems (MAS) [16] focuses on systems in which many intelligent agents interact with each other. The agents are considered to be autonomous entities, such as software programs or robots. Their interactions can be either cooperative or selfish. That is, the agents can share a common goal (e.g. an ant colony), or they can pursue their own interests (as in the free market economy). The characteristics [16] of MASs are that (1) each agent has incomplete information or capabilities for solving the problem and, thus, has a limited viewpoint; (2) there is no system global control; (3) data are decentralized; and (4) computation is asynchronous.

In the present design the system functions are asynchronous and hence the MAS architecture is chosen as the best fit architecture. This layer has different agents to perform all these functionalities, such as Language Learning Adaptation Agent, Dialogue Agent, Message Handling Agent, Prompter Agent and Director Agent. These agents are operating independently performing the operations.

Language Learning Adaptation Agent is a simple component. It scans the user language selection. Normally, it permits two types of users as of now. One is the normal user and second user is the VHU users. It understands the user and displays different screen for different users. In the Braille key board the system is designed to operate in a command mode. This command mode first supports the language selection. The system is able to understand the language selection and change the system setup accordingly. Message Handling Agent displays error and other messages from the knowledge base in regional language or prompts the display in English for other user. An error message alerts users of a problem that has already occurred. Error messages can be presented using modal dialog boxes, in-place messages, and notifications.

Dialogue Agent gets the input from the visually handicapped user in particular language Braille through special input keyboard attached with the SPECS system. On the other hand it also gets the normal input form the trainer. The design of the Braille keyboard [11] [12] [15] and its components are explained in the Section 4.

Prompter Agent gives the voice output after filtering the output from the browser. This prompter gets the sequence of string from the browser in the form of HTML/DHTML and checks the FONTFACE tag, if it is trained language tag, then it stores the sequence of text in file until it encounters FONT tag, otherwise it simply truncates those tags and HTML input. Then the other tags are given to the sound component. The sound component is able to read the given regional language/English words. A system is designed to read the words in regional language and in English. A girl voicer recorded regional language/English alphabets with different rhythm based on their occurrence in the word at different places. The occurrence may be at the beginning or at the middle or at the end. Then the sound component is designed to pronounce the word with different voice synthesizing.

Director Agent directs the browser to browse the regional language/English web sites according to the selection. Even if the user typed words with small mistakes it is able to direct to the correct website.

4.3 Knowledge Base Layer

The third layer consists of Knowledge base which stores different types of fonts, Regional Language Voice Database, Various Regional Language web site information, etc. The main mechanism of this knowledge base development is knowledge representation, acquisition, learning and reasoning. The section 5 explains how this knowledge is acquired, represented, learning and reasoning.

4.4 Mutli-Lingual Font Repository

This repository stores all the fonts available in each web site and the common regional language fonts of different designers. As soon as the new website is visited it will fetch the new font from that website.

4.5 Regional Language Voice Database

A Prerecorded UNICODE character set is stored in the database. The UNICODE character set is available in the http://unicode.org/. This database provides the equivalent voice file for the particular character as soon as it is requested. These voices are recorded by the voicer in a recording room with three different types of appearance of letters in different places. That is first type of tag is at the beginning of letters, second one is for the middle letters and third for the end letters. The voice chord in Regional Language/English will differ when it appear in different places.

4.6 Multi-Lingual Web Site Database

This database contains the multi-lingual web name and the equivalent English web site name. These websites are taught by the normal person during the training phase of the system. Even through the system is designed for the people; a non-blind person can also use the system in a normal way. His / her duty is to teach the equivalent multi-lingual web site name for every known multi-lingual portal. and third for the end letters. The voice chord in Regional Language/English will differ when it appear in different places.

5 Design and Functionality of Multi-Lingual Braille Keyboard

Braille is a touch and feel system for the Visually Handicapped person, which uses an arrangement of 6 dots called a cell. The cell is three dots in height and two dots wide. Each Braille character is formed by placing one or more dots in specific positions.

A printed sheet of Braille normally contains upwards of twenty five rows of text with forty cells in each row. The physical dimensions of a standard Braille sheet are approximately 11 inches by 11 inches. The dimensions of the Braille cell are also standardized but these may vary slightly depending on the country. The dimension of a Braille cell, as printed on an embosser is shown below.

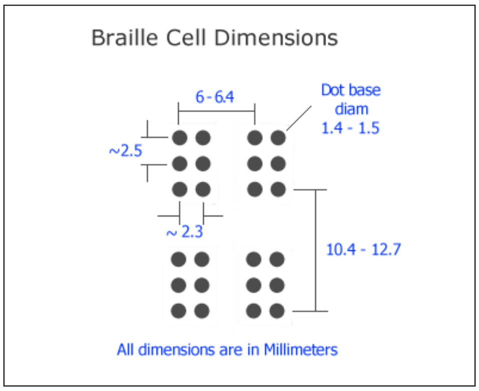


Figure 3: Braille Cell Dimensions

A sheet of Braille may thus appear to hold information amounting to about a thousand characters (letters of the alphabet). Later we will see that the designers of the Braille system had foreseen the need to present information in compact form so that a set of cells could convey much more information in the string of letters forming the cells. Try reading the following sentence shown in Grade-1 Braille.

To aid in describing these characters the positions in the Braille cell are numbered 1, 2, 3 downward and on the left, and 4, 5, 6 downward on the right. It is shown in the following Figure 3.

The Braille Keyboard has six keys, a line spacer, a back spacer and a space bar. The six keys correspond to his six dots of the Braille cell. The keys are struck one or more at a time so that one Braille cell is written with each stroke. There are three keys each side of the space bar. The

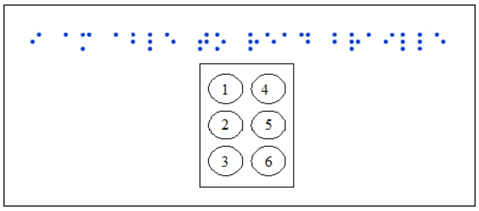


Figure 4: Braille Key Board Design

left index finger uses the key to the left of the space bar, which strikes dots1; the middle finger, dot 2; and the left ring finger, dot 3. The right finger, middle finger and the ring index finger strike the keys for dots 4, 5 and 6 respectively. Thumb strike on the space bar leaves a blank cell. The Bharathi Braille Tamil fonts and Grade-1 Braille are used to key in the fonts in to the SPECS system. The Bharathi Braille fonts and Grade-1 Braille are shown in the following Figures 5 and 6 respectively.

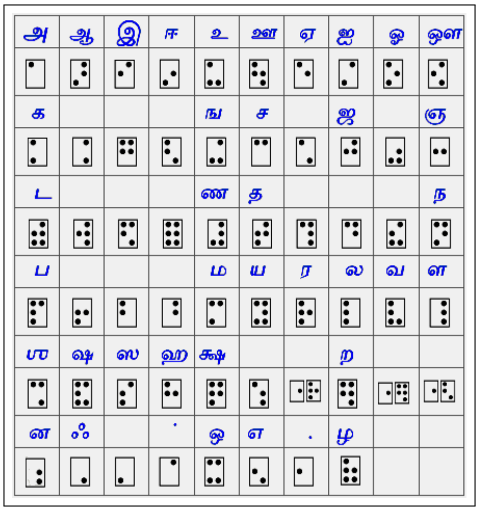

அ
ஆ
இ
ஈ
உ
ஊ
ஏ
ஐ
ஓ
ஔ
க
ங
ச
ஜ
ஞ
ட
ண
த
ந
ப
ம
ய
ர
ல
வ
ள
ஶ
ஷ
ஸ
ஹ
க்ஷ
ற
ன
ஃ
்
ஒ
எ
.
ழ

Figure 5: Bharathi Braille fonts

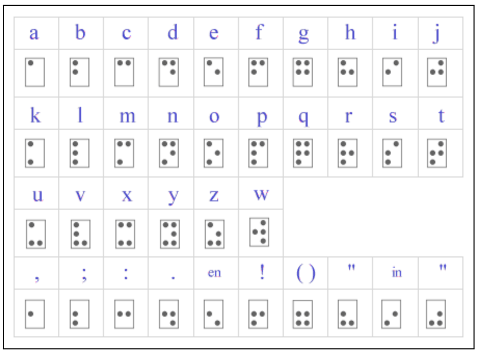

a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
q
r
s
t
u
v
x
y
z
w
,
;
:
.
en
!
( )
"
in
"

Figure 6: Bharathi Braille fonts

6 SPECS Machine Learning System

The SPECS system uses three different learning systems for usage. First learning mechanism applied is the simple rote learning mechanism to understand the English URLs and their equal Multi-lingual Braille names. The trainer must explicitly train the system to understand the equivalent URLs. The idea is that one will be able to quickly recall the meaning of the URL by repeating it often. The second learning is the reinforcement learning mechanism that must understand the occurrences of an alphabet in different words in different places and sequence the sound files according to their requirement. It can be used in cases where there is a sequence of inputs and the desired output is only known after the specific sequence occurs. Reinforcement learning is concerned with how an agent ought to take actions in an environment so as to maximize some notion of cumulative reward. In machine learning [17], the environment is typically formulated as a Markov decision process (MDP), and many reinforcement learning algorithms for this context are highly related to dynamic programming techniques. The main difference to these classical techniques is that reinforcement learning algorithms do not need the knowledge of the MDP and they target large MDPs where exact methods become infeasible. This process of identifying the relationship between a series of input values and a later output is temporal difference learning. This algorithm is adapted to understand the word and then present the voice to the VHU users. The third learning is the font usage for the different system and again it is a rote leaning mechanism. As soon as the agent finds new fonts it downloads those fonts and stores them in the local database.If you wish to include color illustrations in the electronic version in place of or in addition to any black and white illustrations in the printed version, please provide the managing editor and the editorial assistant with the appropriate files.

7 Simulation Experiment

This system is developed using Microsoft Visual C++ under Microsoft Windows 9x platform. The special browser developed can be facilitating the normal user to give the inputs and train the system. All the given components and agents are developed and integrated and installed with the Braille keyboard and a speaker. The system is installed with the normal Hardware; 800 MHz Intel Pentium 3, 256 MB RAM machine and tested. After the installation the normal users trained the system by visiting different Tamil web sites. The fonts from these websites also downloaded and put into the repository. The Language Learning Adaptation Agent takes care of installation of fonts in the respective system font's directory as soon as it is downloaded. The SPECS browser developed and sample screen is shown in the following Figure 7.

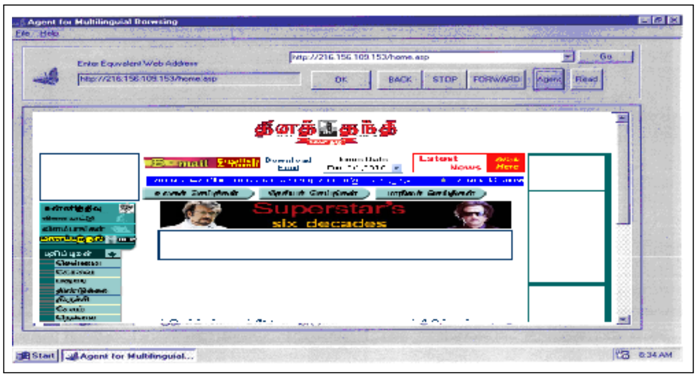

Agent for Multilingual Browsing
File Help
Enter Equivalent Web Address
http://216.156.109.153/home.asp
http://216.156.109.153/home.asp
Go
OK
BACK
STOP
FORWARD
Agent
Read
Superstar's
six decades
Start
Agent for Multilingual...

Figure 7: A view of SPECS browser

User type	**VHU**	**Non-VHU**
Mean-Time to Access Time (s)	800	200

Table 1: Comparison of Navigation Time of VHU Vs Non-VHU

8 Experiment Result and Discussion

In these experiments, the system is trained with 600 web sites by the normal non-blind person. Then four blind persons were brought from the blind school and they were asked to browse in the system using the specialized browser. After getting their opinion the system performance is evaluated in two different ways. In order to evaluate the effectiveness of this browsing [24], system is measured with the well known precision measure used for different query in both normal browser and special browser and is shown by

Precision = Total Number of Documents Retrieved /Total Number of Documents Trained

The Precision is the probability that a (randomly selected) retrieved document is relevant. Based on the measurement pertaining to precision is compared for both the blind user and with the non-blind users. It is found that the graph with precision taken in both experiments justifying equality in retrieval effectiveness.

To evaluate the efficiency of access system, it is essential to find the Mean Time to Access [8] [9] [10] the browser by both the VHU and Non-VHU. This result is compared and is presented in the Table 1. It is found that the VHU takes four times of the access time compared with the Non-VHU-Figure 8.

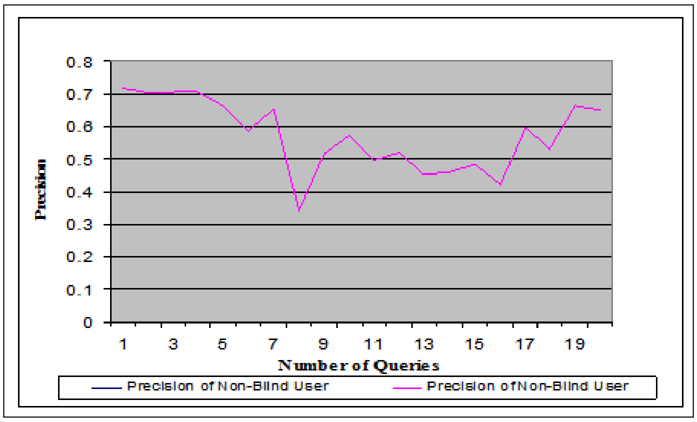


Figure 8: Comparison of Precision for Both VHU and Non-VHU

9 Conclusion

In the present work a special browser has been designed and developed for visually handicapped persons. The mean-time to access is taken into account for the VHU as well as for Non-VHU (Internet Explorer 7.0 for Non-blind person) and the comparative results are presented. In addition to this the precision has been measured under different situations. This browser is very much helpful visually handicapped persons of regional language to access the Internet without any difficulties. The system is able to read only the static websites. The user is not able to travel into complete set of hyperlinks provided in the site and the present browser functionality is restricted to access the text messages alone. Further work is necessary to take care of these problems of this system.

Bibliography

[1] Serge Resnikoff et al, Global magnitude of visual impairment caused by uncorrected refractive errors in 2004, *Bulletin of the World Health Organization*, 86(3):63-70, Jan 2008.

[2] Murthy GV, Gupta SK, Bachani D, Jose R, John N.; Current estimates of blindness in India, *Br J Ophthalmol* 89:257-60, PMID: 15722298 doi:10.1136/bjo.2004.056937, 2005.

[3] Serge Resnikoff et al., Policy and Practice, Global data on visual impairment in the year 2002, *Bulletin of the World Health Organization*, 82(11):844-852, Nov 2004.

[4] Magnitude and Causes of Visual Impairment, Fact Sheet May 2009, World Health Organization, http://www.who.int/mediacentre /factsheets/ fs282/en/

[5] Leong Kok Yong et al., Multiple Language Support over the World Wide Web, *World Wide Web Journal*, 2: 165-176, 1997.

[6] S. Kuppswami et al, SPECS: Friendly Computer System for the Visually Handicapped – A Proposal, *Proc. of the National Conference on Creating Convenient and Friendly Environment for Education and Training of the handicapped in Technical Institutions*, December 1999, Roorkee University.

[7] S.Kuppuswami, V.Prasanna Venkatesan, T.Chithralekha, Role Based Agents for Internet Access Through SPECS , *CSI Conference*, Chennai, Sept 2000.

[8] Browser Speed Comparisons, http://www.howtocreate.co.uk/ browserSpeed .html

[9] Tenni Theurer, Performance Research, Part 2: Browser Cache Usage – Exposed: A web Blog, http://yuiblog.com /blog /2007/01/04/performance-research-part-2/, 2007.

[10] Michael Czeiszperger, Evaluating Apple's Browser Performance Claims in The Real World, http://www.webperformanceinc.com/library/reports/Safari % 20 Benchmarks/

[11] Bharathi Braille Fonts, http;//acharya.iitm.ac.in/

[12] Standard Grade-1 Braille Fonts, http;//acharya.iitm.ac.in/

[13] R. Ponnusamy; T. Chithralekha; Prasanna Venkatesan; S. Kuppuswami, Development of an Agent Based Specialised Web Browser for Visually Handicapped Tamils, *Lecture Notes in Computer Science*, Springer-Verlag LNCS 5616, ISSN 0302-9743, Universal Access in HCI, Part III, HCII2009, 778-786, 2009.

[14] Windows XP Technical Manual, Microsoft Corporation.

[15] http://acharya.iitm.ac.in/sdi.html

[16] http://www.aaai.org/AITopics/pmwiki/pmwiki.php /AITopics /MultiAgentSystems

[17] Sio-long Ao et. al., Machine Learning and Systems Engineering, *Lecture Notes in Electrical Engineering*, Springer, 1st Edition, 2010.

[18] Stuart Goose , Mike Newman , Claus Schmidt , Laurent Hue, Enhancing Web accessibility via the Vox Portal and a Web-Hosted Dynamic HTML-VoxML Converter, *Proc. of the 9th int. World Wide Web conference on Computer networks : the int. J. of Computer and Telecommunications Networking*, 583-592, June 2000.

[19] Patrick Roth et al., AB – Web : Active audio browser for visually impaired and blind users, *Int. Conference on Auditory Display*, University of Glasgow, UK, November 1-4, 1998.

[20] Jing Xiao, GuanNeng Huang,Yong Tang, An Open Source Web Browser for Visually Impaired, *Third Int. Conference on Intelligent Computing, ICIC 2007*, LNCS, Vol. 4681: 90-101, 2007.

[21] IBM Accessibility Internet Browser for Multimedia, A tool that enables multimedia content on the Internet to be enjoyed by people with visual impairments, September 27, 2007, http://www.alphaworks.ibm.com /tech/aibrowser

[22] Tim Morris et. al., Clearspeech: A Display Reader for the Visually Handicapped, *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(4):492-500 , Dec 2006.

[23] Tadayoshi Fujiki et. al., A Tool for Improving the Web Accessibility of Visually Handicapped Persons, *J Med Syst*, 30: 83–89, 2006.

[24] Barbara Leporini and Fabio Patern, Applying Web Usability Criteria for Vision-Impaired Users: Does ItReally Improve Task Performance?, Intl. Journal of Human–Computer Interaction, 24:17–47, 2008.

INT J COMPUT COMMUN, ISSN 1841-9836
9(1):48-55, February, 2014.

Modeling and Simulation of Genetic Fuzzy Controller for L-type ZCS Quasi-Resonant Converter

M. Ranjani, P. Murugesan

Mani Ranjani*
Department of Electrical and Electronics Engineering
Sathyabama University, Chennai, 600119 Tamilnadu, India
*Corresponding author: meranjanisasi@gmail.com

Palzha Murugesan
Department of Electrical and Electronics Engineering
S.A.Engineering College, Chennai, 600077 Tamilnadu, India
murugu1942@yahoo.co.uk

Abstract: A new method of speed control of DC drives using series Quasi-Resonant (QR) Zero current switching (ZCS) DC to DC converters is proposed. It employs a Fuzzy logic controller (FLC) in feedback loop conventionally but due to the advent of new intelligent techniques the FLC are optimized by Genetic algorithm (GA). In this paper the GA optimization technique is applied for speed control of series ZCS-QRC fed drive. The main objective of this work is to obtain reduced transient response, reduced switching stresses and switching losses which in turn enhances the efficiency and commutation capability of motor.

Keywords: Fuzzy logic controller (FLC), Zero Current Switching Quasi-Resonant Converter (ZCS-QRC), Genetic Algorithm (GA), Integral Absolute Error (IAE), Direct Current (DC)

1 Introduction

The dc motor drives have the advantage of high controllability and are used in many applications such as robotic manipulators, position control, steel mining, and paper and textile industries. In some industrial applications, the dynamic response of drives is bounded by certain limitations such as transient time and steady state error. In addition when they are fed from PWM converter [1], they suffer from high switching losses, reduced reliability, electromagnetic inference and acoustic noise. To overcome the above difficulties quasi-resonant converters [2]- [3] are used which can be either zero current (ZC) or zero voltage switching (ZV) .In order to improve the speed response and regulation of the converter fed drives it is necessary to have closed loop control. Conventionally closed loop response employs PI controller but the performance of the drive is sensitive to load disturbances and parameter variations. The advent of FLC has been suggested as an alternative approach to conventional control techniques for complex control system like non linear system. The design of FLC does not require the exact mathematical model of the system and can compensate the parameter variation due to load disturbances ([4]- [5]). Unfortunately a good performance cannot be obtained for incorrect membership functions, fuzzy rules and scaling factors. This necessitates the optimization technique of FLC by Genetic algorithm [6]- [8] to achieve optimal solutions of membership function, fuzzy rules and scaling factors. In this paper speed control of series ZCS-QRC fed DC drive is composed of two steps. The first step is conventional control of DC drive by FLC. The second step is optimization of FLC by GA.

2 Analysis of Series FM-ZCS-QRC Fed DC Drive

The QRC with ZCS topology is considered for the present work.Fig.1. shows the circuit diagram of the QRC fed motor.The waveform for the model is shown in fig.2.

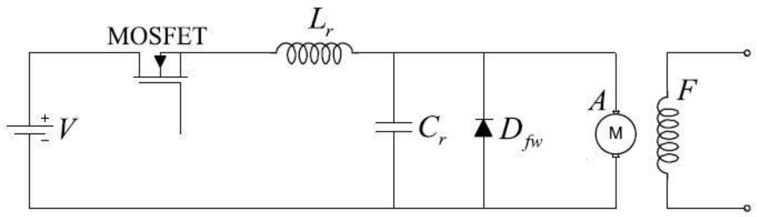


Figure 1: ZCS-Half-wave series quasi-resonant Converter

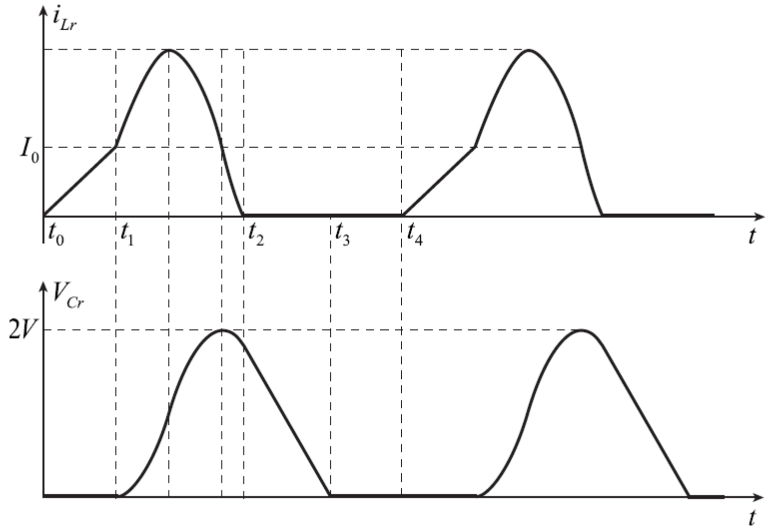


Figure 2: Waveforms for Half wave series ZCS-QRC

To analyze its behavior, the following assumptions are made:

- Armature inductance is much larger than resonant inductance.
- The DC motor is treated as a constant current sink.
- Semiconductor switches are ideal.
- Reactive elements of the tank circuit are ideal.

A switching cycle can be divided into four stages.. Suppose that before MOSFET turns on, diode carries the steady-state output current Ia and capacitor voltage Vcr is clamped at zero. At time t0, MOSFET turns on, starting a switching cycle. *MODE 1 : Linear Stage [t0 , t1]* Input current iLr rises linearly and its waveform is governed by the state equation:

$$Lr(diLr/dt) = E \tag{1}$$

the duration of this stage td1 (= t1 - t0) can be solved with boundary conditions of ILr (0) = 0 and ILr (td1)= Ia , thus

$$td1 = (LrIa)/E \tag{2}$$

MODE 2: Resonant Stage [t1 , t2] At time t1, the input current rises to the level of Ia , freewheeling diode is commutation off, and the difference between the input current and the output current iLr(t) - Ia flows into Cr, Voltage Vcr rises in a sinusoidal fashion.The state equations are

$$Cr(dVcr/dt) = ILr(t)Ia \tag{3}$$

$$Lr(diLr/dt) = EVcr(t) \tag{4}$$

with initial conditions Vcr(0) =0 , iLr(0)= Ia and therefore,

$$iLr(t) = Ia + (E/Z0)sint \tag{5}$$

$$Vcr(t) = E(1 - cost) \tag{6}$$

MODE 3 : Recovering Stage [t2 , t3] Since MOSFET is off at time t2, capacitor begins to discharge through the output loop and Vcr drops linearly to zero at time t3 . The state equation during this interval is

$$Cr(dVcr/dt) = -Ia \tag{7}$$

The duration of this stage td3 (= t3 t2) can be solved with the initial condition Vcr (0) = Vcr

$$td3 = CrVcr/Ia \tag{8}$$

$$td3 = CrE(1 - cos)/Ia \tag{9}$$

MODE 4: Freewheeling Stage [t3 , t4] After t3, output current flows through diode. The duration of this stage is td4 (t4 t3),

$$td4 = Ts - td1 - td2 - td3 \tag{10}$$

where Ts is the period of the switching cycle. After an interval of Toff, during which It is zero and Vcr = 0, the gate drive to the MOSFET is again applied at T4 to turn it on, and the operation during the next cycle is similar to that of the preceding cycle. By controlling the dead time (T4- T3), the average value of the armature voltage and hence the speed of the dc motor can be controlled.

- characteristic impedance

$$Zn = (L_1/C_1) \tag{11}$$

- resonant angular frequency

$$\omega = 1/(L1C1) \tag{12}$$

- resonant frequency

$$fn = \omega/2\pi \tag{13}$$

3 Fuzzy Logic Controller

Fuzzy Logic Control is derived from fuzzy set theory introduced by Zadeh in 1965. In fuzzy set theory, the transition between membership and non-membership can be gradual. Therefore, boundaries of fuzzy sets can be vague and ambiguous, making it useful for approximate systems. The Fuzzy logic controller employed to control the speed of ZCS-QRC fed DC drive is as shown in Fig.3.The FLC is an attractive choice when precise mathematical formulations are not possible. Other advantages of FLC are it can work with less precise inputs, it does not need fast processors,

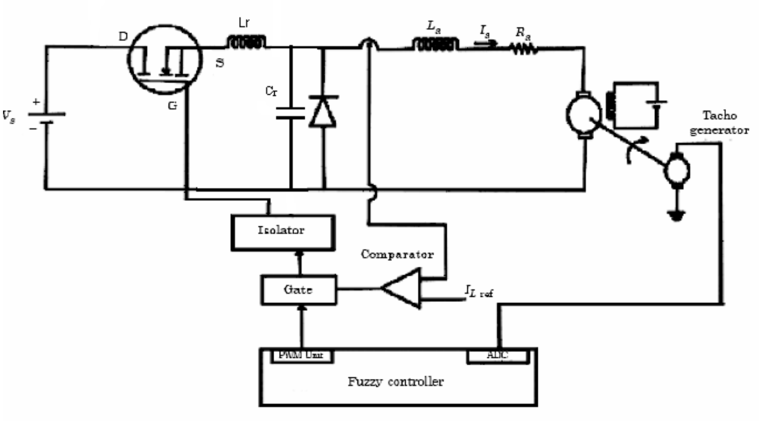


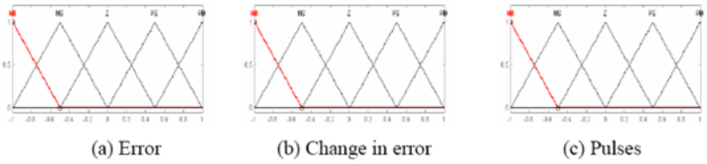
Figure 3: Fuzzy controlled Series ZCS QRC fed DC drive

Table 1: Rule table with 25 rulres

u(t)		e(t)				
		NB	NS	Z	PS	PB
de(t)	NB	NB	NB	NS	NS	Z
	NS	NB	NS	NS	Z	PS
	Z	NS	NS	Z	PS	PS
	PS	NS	Z	PS	PS	PB
	PB	Z	PS	PS	PB	PB

it needs less data storage in the form of membership functions and rules than conventional look up table for non-linear controllers; and it is more robust than other non-linear controllers, parallel with the z-network output terminals.

The simplest form of membership function is triangular membership function and it is used here as the reference. As Sugeno type of implication is considered, the singleton membership function is used for the output variable namely the change in duty cycle. The spread of membership functions for the inputs (error and change in error) and output (pulses) are shown in Fig.4.respectively.



(a) Error (b) Change in error (c) Pulses

Figure 4: Triangular membership functions for FLC

The rules for the designed fuzzy controller are given in the Tables 1 .Table 1 uses five linguistic variables for error and change in error with 25 rules. The five sets used for fuzzy variables ’error’ and change in error are negative big (NB), negative small (NS), zero (Z), positive big (PB), and positive small (PS). From the rule table, the rules are manipulated as If error is NB and change in error is NB, then output is NB.

4 GA Optimized FLC

Genetic algorithms [9]- [10], which are adopted from the principle of biological evolution, are efficient search techniques that manipulate the coding representing a parameter set to reach a near optimal solution. Hence by strengthening fuzzy logic controllers with genetic algorithms the searching and attainment of optimal fuzzy logic rules and high-performance membership functions will be easier and faster. GAs is used regularly to solve difficult search, optimization and machine-learning problems that have previously resisted automated solutions. They can be used to solve difficult problems quickly and reliably. These algorithms are easy to interface with existing simulations and models, and they are easy to hybridize. GAs includes three major operators: selection, crossover, and mutation, in addition to four control parameters: population size, selection crossover and mutation rate. This paper is concerned primarily with the selection and mutation operators. There are three main stages of a genetic algorithm; these are known as reproduction, crossover and mutation. The flow chart of a genetic algorithm is shown in figure 5.

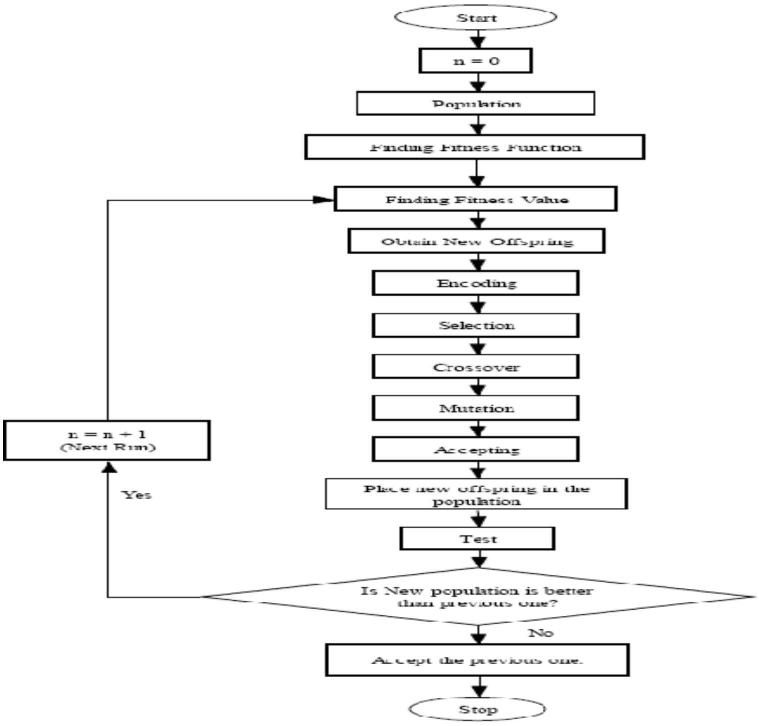


Figure 5: Flowchart for Genetic Algorithm

The steps involved in genetic algorithm are described below.

- Start: Generate random population of n chromosomes (suitable solutions for the problem).
- Fitness: Evaluate the fitness f(x) of each chromosome x in the population.
- New population: Create a new population by repeating following steps until the new population is complete.
  - Selection: Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected).
  - Crossover: With a crossover probability, cross over the parents to form new offspring (children). If no crossover was performed, offspring is the exact copy of parents.
  - Mutation: With a mutation probability, mutate new offspring at each locus (position in chromosome).

  - Accepting: Place new offspring in the new population.

- Replace: Use new generated population for a further run of the algorithm

- Test: If the end condition is satisfied, stop, and return the best solution in current population.

- Loop: Go to step 2.

The objective functions IAE (Integral Absolute Error). The main objective of controller is to minimize the error signal or in other words we can say that minimization of performance indices.

$$IAE = \int_0^t |e(t)dt| \tag{14}$$

The fitness value of the chromosome is the inverse of the performance indices. The fitness value is used to select the best solution in the population to the parent and to the offspring that will comprise the next generations. The fitter the parent greater is the probability of selection. This emulates the Evolutionary process of survival of the fittest. Parents are selected using roulette wheel selection method. Fitness function is reciprocal of performance indices. In this paper we have taken the discrete form of ITAE. ITAE is treated as performance indices and fitness function denoted by J can be described as

$$J = 1/(100 + \sum_{K=1}^{N} |\omega ref - \omega m|) \tag{15}$$

The membership functions obtained for error, change in error and pulses for GA optimized FLC are shown in fig.6. The simulation structure of optimized fuzzy speed controller with the IAE is shown in fig.7 & fig.8 respectively.

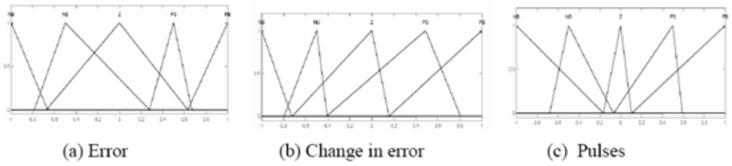

(a) Error (b) Change in error (c) Pulses

Figure 6: GA Optimized membership functions

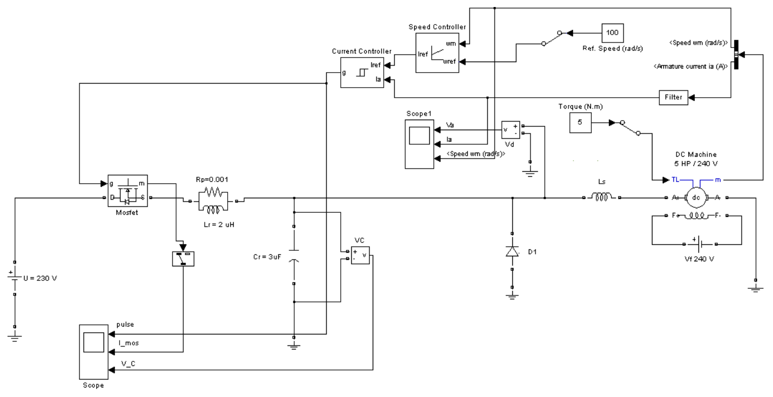


Figure 7: Simulated structure of GA based FLC of series ZCS-QRC fed DC drive

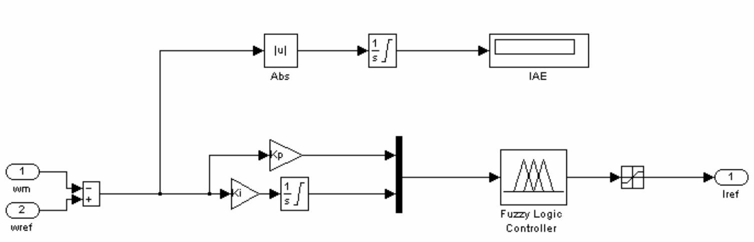
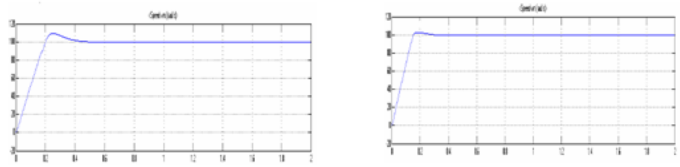


Figure 8: Simulated structure of Optimized Fuzzy speed controller

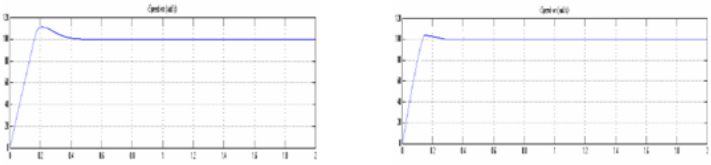
5 Simulation Results

The closed loop operations of series FM-ZCS-QRC fed DC drive has been simulated. Controlling the freewheeling period using the controller regulates the speed of the drive. The speed variations for sudden load disturbances are shown for increased load torque and decreased load torque in fig.9 and fig.10 respectively. It can be said that the speed remains constant even for various disturbances of load torque.



(a) with Fuzzy speed controller (b) with GA optimized Fuzzy speed controller

Figure 9: Speed response for increase in load torque



(a) with Fuzzy speed controller (b) with GA optimized Fuzzy speed controller

Figure 10: Speed response for decrease in load torque

6 Conclusion

We design the speed controller of the Series FM-ZCS-QRC in two steps. In step one the conventional FLC is designed and simulated and second step optimization of membership function ,fuzzy rules and scaling factors of FLC by Genetic algorithm is considered. The simulation results show that the optimal fuzzy logic controller is functioning better than a conventional FLC in terms of the rise and settling time. Hence it can be concluded that the GA optimized FLC implemented in ZCS-QRC fed drive enhances the drive robustness by reducing the transient time and steady state error and superior than the conventional fuzzy controller.

Bibliography

[1] Ned Mohan et.al. (1995); *Power electronics conveiters, Application and Design*,John Wiley and Sons.

[2] Liu.K.H.; Lee.F.C. F.C. Lee.; Zero- voltage switching technique in DC/DC converters; *IEEE Trans. Power Electron*, 5(3):293-304, 1990.

[3] Rama Reddy et al, A step down frequency modulated Zero current switching QRC fed DC drive,*ETEP Joumal*, Ref.No.ET-1088, 1995.

[4] Abraham Kandel; Gideon Langholz; *Fuzzy Control Systems*, CRC Press, 1994.

[5] Kung.Y.S.; Liaw.C.M.; A Fuzzy Controller improving a Linear Model Following Controller for Motor Drives; *IEEE Transactions on Fuzzy Systems*, 2(3):194-201, 1994.

[6] Reznik.L; Evolution of Fuzzy Controller Design; *IEEE Transactions on Fuzzy Systems*, 503-508, 1997.

[7] Hu.B.G.; Gosine.R.G.; Theoretic and Genetic Design of a Three-Rule Fuzzy PI Controller; *IEEE Transactions on Fuzzy Systems*, 489-496, 1997.

[8] Tan. G.V.; Hu.X.; More on Designing Fuzzy Controllers using Genetic Algorithms: Guided Constrained Optimization;*IEEE Transactions on Fuzzy Systems*, 497-502, 1997.

[9] Mohammadian.M.; Stonier. R.J.; Generating fuzzy rules by genetic algorithms; *Proceedings of 3rd IEEE International Workshop on Robot and Human Communication*, 362-367, 1994.

[10] Arulselvi.S.; Uma Govindarajan.; Real time implementation of modified Fuzzy logic controller for a non linear quasi resonant DC-DC converte ; *IETE Journal of research*, 53(5):401-416, 2007.

INT J COMPUT COMMUN, ISSN 1841-9836
9(1):56-61, February, 2014.

Improving the Efficiency of Image Clustering using Modified Non Euclidean Distance Measures in Data Mining

P. Santhi, V. Murali Bhaskaran

P.Santhi*
Computer Science Department
Paavai Engineering College, Pachal,Namakkal,India
*Corresponding author: santhipalanisamypec@paavai.edu.in

V.Murali Bhaskaran
Paavai College of Engineering
Pachal, Namakkal, India
murali66@gmail.com

Abstract: The Image is very important for the real world to transfer the messages from any source to destination. So, these images are converted in to useful information using data mining techniques. In existing all the research papers using kmeans and fuzzy k means with euclidean distance for image clustering. Here, each cluster needs its own centric for cluster calculation and the euclidean distance calculate the distance between the points. In clustering this process of distance calculation did not give efficient result. For make this in to efficient, this research paper proposes the non Euclidean distance measures for distance calculation. Here, the logical points are used to find the cluster. The result shows that image clustering based on the modified non Euclidean distance and the performance shows the efficiency of non euclidean distance measures.

Keywords: Data Mining, Image Mining, Kmeans, Fuzzy Kmeans, Euclidean Distance

1 Introduction

In real world, the image plays an important role in the entire field. Normally, all the images having some information related to any application [1] [2]. So, these images are converted in to some useful information using data mining techniques. This process of mining the image is called as image mining. Here, the clustering technique is applied to these images. The process of grouping the similar data or pixels is called as clustering [5].. The data mining having two types of learning is supervised learning and unsupervised learning. In supervised learning, the training dataset was provided to train the system and got the output based on these training data. This type of system is called as classifiers. In classification, group the same item based on the physical behavior. In unsupervised learning, the training data set is not provided to the learning system. In this paper proposes the clustering for the Image. The clustering is one of the techniques in unsupervised learning. The clustering is having many algorithms are partitioned based clustering, Hierarchical based clustering, dense based clustering and distribution based clustering. In partitioned based clustering the cluster based on the centre point [6][4]. Here, the initialization of centre is very difficult and this process takes more time to execute. In hierarchical clustering the clusters are formed by using distance connectivity [3]. In density based clustering the cluster formed using the dense value.

The distance between the points is calculated by using the distance measures. Data mining having the two types of measures called Euclidean distance and non Euclidean distance measure [7]. In euclidean distance measure, the distance is calculated based on the physical points of the cluster. But this measure is not efficient for the clustering to produce the efficient result.

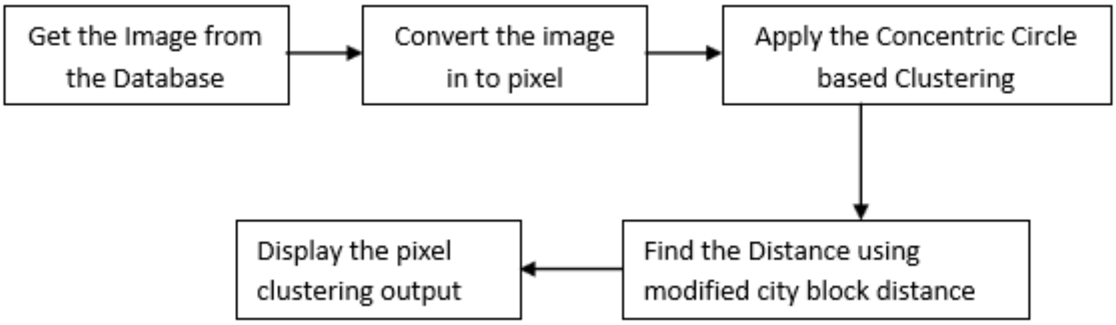

Get the Image from the Database
Convert the image in to pixel
Apply the Concentric Circle based Clustering
Find the Distance using modified city block distance
Display the pixel clustering output

Figure 1: Steps for the proposed system [8]

In non euclidean distance, the distance is calculated using feature vectors between the clusters [4]. In the literature survey most of the researches are done by using kmeans and fuzzy kmeans algorithm for clustering and the city block distance measures for measure the distance between the clusters. In Existing formula for city block distance is [9]

$$D = \sum_{j=1}^{n} [a_j - b_j] \tag{1}$$

In this paper proposes the modified city block distance measures for calculating the distance between the clusters and using concentric circle based clustering algorithm.

2 Research Methodology

Non euclidean distance Measure is very important for the clustering to measure the distance between the points. In this research using the modified non euclidean distance measure of city block distance or manhatten distance. In Existing formula is

$$D = \sum_{j=1}^{n} [a_j - b_j]$$

In this formula is modified in to

$$D = \sum_{j=1}^{n} [a_j - b_k] \tag{2}$$

The equation 2 is called as modified city block or manhatten distance. The steps for this system will be discussed in figure 1.

1. Get the Image.
2. Convert this image into pixel.
3. The concentric circle based clustering algorithm is applied to pixel.
4. Find the distance between the clusters using modified city block distance.
5. Step 4 will be repeated until all the pixels come under any one of the clustering.

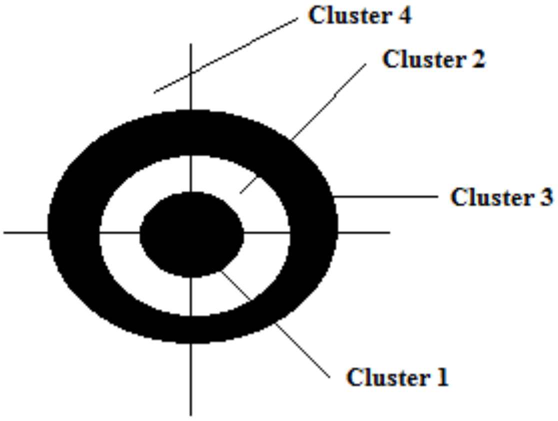


Figure 2: Sample for Concentric Circle Based Clustering

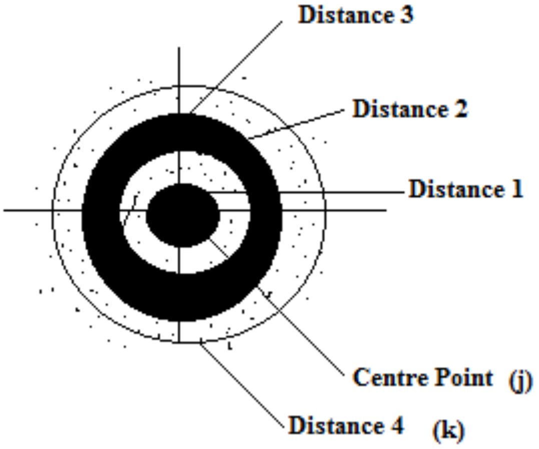


Figure 3: Sample for Pixel Based Clustering

6. Display the output of pixel clustering

In concentric circle based clustering; only one centre is used for all the clusters. This process will reduce the time for image clustering. In existing research, the reduction in the running time is already discussed. Now, this paper discusses the improvement in the efficiency of image clustering. The figure 2 shows the sample of concentric circle based clustering. Here it contains the four groups or clusters using only one centre. The figure shows the pixel clustering using modified city block distance or manhatten distance. Consider the centre assumption is 0.

The formula is

$$D = \sum_{j=1}^{n} [a_j - b_k]$$

- D is the Modified city block or Modified Manhatten distance Measure,
- J is the dimension from 1 to k dimension
- a_j is the centre pixel assumption point

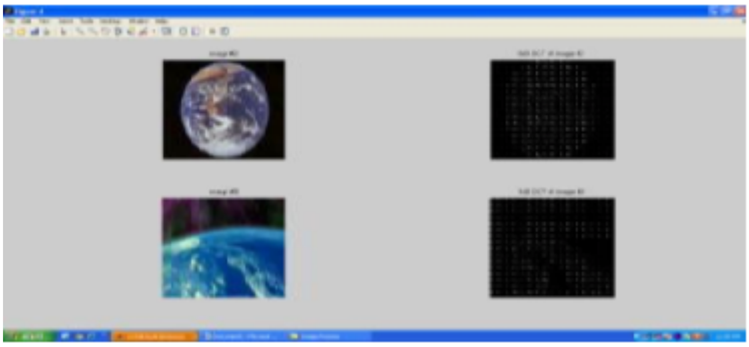


Figure 4: Input Images for Pixel Clustering

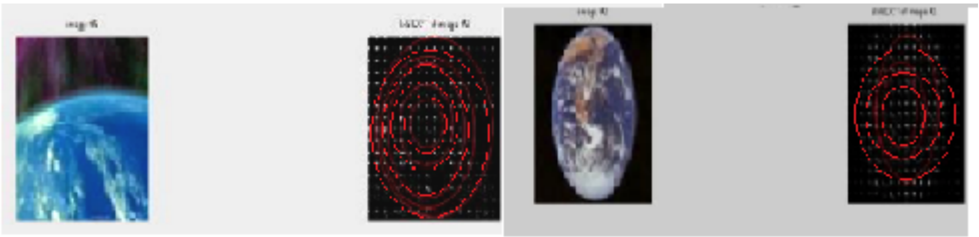


Figure 5: Output images with Pixel Clustering

- b_k is the Top most pixel assumption point

Based on this formula, a_j=0 and the value of b_k=1.Here, the j is the centre value of 0.This value is same for all the clusters. The k value is min distance from that centre. The same manner all the k values are calculated. For the cluster 1, D=0-1=[-1]=1 The cluster 1, the min distance from the centre is 1.So, the pixels within the 1 from X and Y axis are considered as the first cluster. For the cluster 2, D=0-2=[-2]=2, Here a_j=0 and b_k= 2. The cluster 2, the next to the min distance from the centre is 2.So, the pixels within the 2 from X and Y axis are considered as the second cluster. For the cluster 3, D=0-3=[-3]=3, Here a_j=0 and b_k= 3. The cluster 3, the next to the min distance from the centre is 3.So, the pixels within the 3 from X and Y axis are considered as the second cluster. The same way all the clusters are formed.

3 Results and Discussion

The result shows the original image and the equivalent pixel to that image.The figure 5 shows the clustering of pixel using concentric circle based clustering. The image 1 having the 3 clusters and the image 2 shows the 5 clusters.

In this result all the clusters in the image 1 and the image 2 using the single centre point for all the clusters and the distance between the centers is calculated using modified city block distance or manhatten distance.

4 Performance Analysis

The graph shows the performance of clustering algorithm using modified non euclidean distance. In the performance analysis the concentric circle based clustering with modified city block

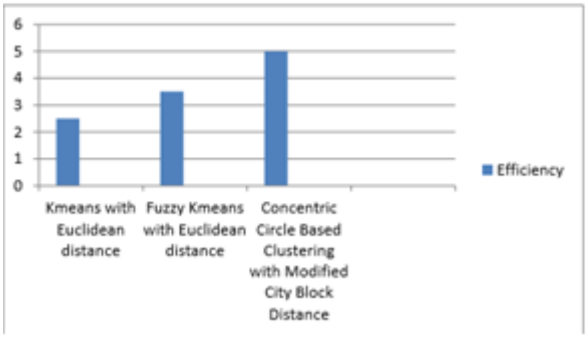


Figure 6: Performance Graph

distance algorithm having more efficiency comparing to existing system.

5 Conclusion and Future Enhancement

The images are very important to the real world in entire field. The images are converted in to useful knowledge using data mining techniques. In existing the researches are did by using k means and Fuzzy k means for clustering with euclidean distance. But it is having more difficulties in assigning centre for each cluster and the efficiency is not high. So, improving the efficiency this research paper gives the image clustering using concentric circle based clustering with non euclidean distance measure of modified city block distance with single centre point. This Result gives the better performance in the accuracy comparing to existing system. In Future, the concentric circle based clustering is combined with soft computing techniques for the image clustering.

Bibliography

[1] Nor Ashidi Mat Isa; Samy Salamah, A, Umi Kalthum Ngah; Adaptive Fuzzy Moving K-means Clustering Algorithm for Image Segmentation, *IEEE Trans on Knowledge and Data Engineering*, 2145-2153, 2009.

[2] Siti Noraini Sulaiman; Nor Ashidi Mat Isa; Adaptive Fuzzy-K-means Clustering Algorithm for Image Segmentation, *IEEE Trans on Knowledge and Data Engineering*, 2661-2668, 2010.

[3] Maduri, A.Tayal;Raghuwanesh,M.M.; Review on various Clustering Methods for Image Data, *J. of Emerging Trends in Computing and Information Sciences*, 34-38, 2010.

[4] Keh-Shih Chuang; Hong-Long Tzeng; Sharon Chen; Jay wu; Tzong-Jer Chen; Fuzzy c-Means Clustering with spatial information for image segmentation,*Elseiver*, 9-15, 2006.

[5] Sneha Silvia, A.; Vamsidhar, Y.; Sudhakar,G.; Color Image Clustering using K-Means,*IJCST*, 11-13, 2011.

[6] Vasuda, P.; Satheesh, S.; Improved Fuzzy C-Means Algorithm for MR Brain Image Segmentation, *IJCSE*, 1713-1715, 2010.

[7] Fahim, A.M.; Saake, G.; Salem, A.M.; Torkey, F.A.; K-Means for Spherical Clusters with Large Variance in Sizes, *World Academy of Science, Engineering and Technology*, 177-182, 2008.

[8] John Peter, S.; Minimum Spanning Tree-based Structural Similarity Clustering for Image Mining with Local Region Outliers,*Int. J. of Computer Applications*, 33-40, 2010.

[9] Chawan, P.M.; Saurabh Bhonde,R.; Shirish Patil; Concentric Circle-Based Image Signature for Near-Duplicate Detection in Large Databases,*Electronics and telecommunication Research Institute*, 2010.

INT J COMPUT COMMUN, ISSN 1841-9836
9(1):62-70, February, 2014.

Using the Breeder GA to Optimize a Multiple Regression Analysis Model used in Prediction of the Mesiodistal Width of Unerupted Teeth

F. Stoica, C.G. Boitor

Florin Stoica
Department of Mathematics and Informatics, Faculty of Sciences,
"Lucian Blaga" University of Sibiu,
Romania, 550012, Dr. Ion Ratiu 5-7, Sibiu
florin.stoica@ulbsibiu.ro

Cornel Gheorghe Boitor
Department of Preventive Dentistry, Faculty of Medicine "V. Papilian",
"Lucian Blaga" University of Sibiu,
Romania, B-dul. Victoriei, No. 10, Sibiu
boitor.cornel@ulbsibiu.ro

Abstract: For the prediction of the unerupted canine and premolars mesiodistal size, have been proposed different variants of multiple linear regression equations (MLRE). These are based on the amount of the upper and lower permanent incisors with a tooth of the lateral support. The aim of the present study was to develop a method for optimization of MLRE, using a genetic algorithm for determining a set of coefficients that minimizes the prediction error for the sum of permanent premolars and canines dimensions from a group of young people in an area Romania's central city represented by Sibiu. To test the proposed method, we used a multiple linear regression equation derived from the estimation method proposed by Mojers to which we adjusted regression coefficients using the Breeder genetic algorithm proposed by Muhlenbein and Schlierkamp. A total of 92 children were selected with complete permanent teeth which had not clinically visible dental caries, proximal restorations or orthodontic treatment that requires the decrease of the mesiodistal size of teeth. For each of these models was made a hard dental stone which was then measured with a digital calliper, the instrument having an accuracy of 0.01 mm. To improve prediction equations, we divided data into training and validation sets. The Breeder algorithm, using the training set, will provide new values for regression coefficients and error term. The validation set was used to test the accuracy of the new proposed equations.

Keywords: Genetic Algorithms (GA), mixed dentition analysis, multiple regression equations, mesiodistal teeth size.

1 Introduction

The estimation of the mesiodistal size of permanent canine and of the two premolars before their eruption is important for early evaluation of the need for space in this area and consequently to mandible and maxillary. This represents an important part of diagnosis and orthodontic treatment strategy.

Methods of estimation, performed during mixed dentition, can be grouped into three categories: those using multiple linear regression equations, those using radiographs and those using a combination of the two methods [1], [3]- [7].

Among these methods recently reported in literature, those based on MLRE have the highest predictive capacity of the mesiodistal dimension (MDD) for unerupted canines and premolars.

Prediction capacity of these methods can vary depending on the characteristics of different types constitutional areas, sometimes may vary even in the same countries [5]- [8], [9]- [17].

Our aim was to verify if optimizing through an original method the MLRE used recently in the literature [3], [14], can be predicted with sufficient accuracy the sizes of unerupted teeth from the support area in a group of children in Sibiu located in a central area of Romania.

The first objective of the study was to verify the accuracy of recently used MLRE based on known variables, namely the mesiodistal dimensions of teeth 42, 21 and 46, used in prediction of the sizes of unerupted teeth from the support area. The second objective of the study was to use an evolutive calculation method based on the genetic algorithm Breeder, to optimize the regression coefficients used in the MLRE. Thus the accuracy of the predictions can be improved [1], [9], [19], [20].

2 Subjects, materials and prediction methods

A representative public school with a population of 321 children ages 12-15 years from Sibiu (Romania) was selected for this study. From these subjects, a random simple technique was used to select 92 students (47 females and 45 males) fulfilling the selection criteria:

- To have the parents' written consent to participate in the study;
- To present the dental arches fully erupted permanent teeth (molars 3 was not considered);
- The erupted teeth show no abnormalities of shape, size or structure;
- The teeth must not have missing of substance in the mesiodistal size, due to decay, trauma or orthodontic treatments have provided striping.

Dental impressions were taken with alginate impression material and immediately poured with hard dental stone to avoid any distortion. The measure tooth size models we used a digital calliper manufactured by Mega (Germany) with an accuracy of 0.01 mm.

Measurements were performed after the procedure proposed by Seipel [12]. All models were measured 2 times by the same author and the result used was the average of two values. It was calculated the Pearson correlation coefficient between measurements and method error (ME) was calculated using the formula of Dahlberg: $ME = \sqrt{d^2/2n}$ where d is the difference between the two measurements and n is the number of patterns measured a second time.

For estimation the size of the unerupted canines and premolars we chose a recently proposed equation [3] based on known variables 21, 42 and 46. The form of this equation is: $Y = X_1*A_1+X_2*A_2+X_3*A_3+A$, where Y is the outcome expected, X_1, X_2, X_3 are independent variables determined by the size of teeth 42, 46 and 21, A_1, A_2 and A_3 are regression coefficients for used teeth and A is a specific constant. The values of constant A and regression coefficients of the equation are presented in Table 1:

Canines premolars group	Constant A	A_1 (42)	A_2 (46)	A_3 (21)
Maxillary	6,563	0,822	0,595	0,411
Mandible	3,350	0,872	0,710	0,538

Table 1: Parameters of multiple linear regression equation used [3]

In the following is presented our approach in optimization of the regression coefficients presented above, to provide a more accurate method for prediction of the mesiodistal width of unerupted permanent canines and premolars.

Because parameters of the multiple linear regression equation are real values, we are using a Breeder genetic algorithm, in order to avoid a weak point of classical GAs, represented by their discrete representation of solutions, which implies a limitation of the power of the optimization process [20].

The Breeder genetic algorithm, proposed by Mühlenbein and Schlierkamp-Voosen [2] represents solutions (chromosomes) as vectors of real numbers, much closer to the reality than normal GA's.

The selection is achieved randomly from the $T\%$ best elements of current population, where T is a constant of the algorithm (usually, $T = 40$ provide best results). Thus, within each generation, from the $T\%$ best chromosomes are selected two elements, and the crossover operator is applied over them. On the new child obtained from the mate of the parents is applied the mutation operator. The process is repeated until are obtained $N-1$ new individuals, where N represents the size of the initial population. The best chromosome (evaluated through the fitness function) is inserted in the new population (1-elitism). Thus, the new population will have also N elements.

2.1 The Breeder genetic operators

Let be $x = \{x_1, x_2, ..., x_n\}$ and $y = \{y_1, y_2, ..., y_n\}$ two chromosomes, where $x_i \in \mathbb{R}$ and $y_i \in \mathbb{R}$, $i = \overline{1,n}$. The crossover operator has a result a new chromosome, whose genes are represented by values $z_i = x_i + \alpha_i(y_i - x_i)$, $i = \overline{1,n}$, where α_i is a random variable uniformly distributed between $[-\delta, 1+\delta]$, and δ depends on the problem to be solved, typically in the interval $[0, 0.5]$.

The probability of mutation is typically chosen as $1/n$. The mutation scheme is given by $x_i = x_i + s_i \cdot r_i \cdot a_i$, $i = \overline{1,n}$ where: $s_i \in \{-1, +1\}$ uniform at random, r_i is the range of variation for x_i, defined as $r_i = r \cdot domain_{x_i}$, where r is a value in the range between 0.1 and 0.5 (typically 0.1) and $domain_{x_i}$ is the domain of the variable x_i and $a_i = 2^{-k \cdot \alpha}$ where $\alpha \in [0,1]$ uniform at random and k is the number of bytes used to represent a number in the machine within is executed the Breeder algorithm (mutation precision).

2.2 The Breeder genetic algorithm

The skeleton of the Breeder genetic algorithm may be defined as follows [19]:

```
Procedure Breeder
**begin**
$t = 0$
```

```
Randomly generate an initial population $P(t)$ of $N$ individuals
Evaluate $P(t)$ using the fitness function
while (termination criterion not fulfilled) do
    for $i = 1$ to $N - 1$ do
        Randomly choose two elements from the $T\%$ best elements of $P(t)$
        Apply the crossover operator
        Apply the mutation operator on the child
        Insert the result in the new population P'(t)
    end for
    Choose the best element from $P(t)$ and insert it into $P\prime(t)$
    $P(t+1) = P\prime(t)$
    $t = t + 1$
end while
end
```

2.3 The optimization process

The aim of the Breeder genetic algorithm is to find new values for the parameters of multiple linear regression equation presented in table 1, in order to reach a better prediction.

Each chromosome contains four genes, representing the real values $A_i, i = \overline{1,3}$ and A. The fitness function for chromosomes evaluation is represented by the number of cases from the training set having an approximation error obtained with the new equation (in absolute value) bigger than prediction error provided by original equation. In our tests, parameters of Breeder algorithm are assigned with following values: $\delta = 0$, $r = 0.1$ and $k = 8$. The initial population has 1500 chromosomes and algorithm is stopped after 30000 generations.

Data provided by our study models was randomly divided in two sets: the training set, containing 50 cases and the validation set, composed by 42 study models.

Implementation of our new optimization method was accomplished in Java language, using Net Beans 7.01.

3 Results

The MLRE method is using two equations, one for mandible and other for maxillary.

Because there are differences in measurements of teeth between left and right quadrants for mandible and respectively for maxillary, in order to improve the prediction, we are using four equations, one for each quadrant.

Using the data from training set, the Breeder algorithm finds new values for the parameters of the initial multiple linear regression equation (Table 2). The accuracy of prediction made by optimized equations was verified using the validation set.

The order of reliability of both compared prediction methods is the same. As we can see from the table 3, the correlation coefficient r calculated for the all four linear regression equations is almost the same for the original MLRE equations as for the Breeder optimized equations.

In the following figures are evaluated the original and respectively the optimized multiple linear regression equations, using as criteria the number of cases better evaluated.

A comparison of prediction error in estimating the mesiodistal widths of the canines and premolars in the mandible and maxilla using multiple linear regression equations in original form and respectively in optimized form is presented in figures 3-6:

Quadrant	A	A_1	A_2	A_3
1	5.1917	0.7571	0.85332	0.28341
2	5.16292	0.90463	0.68192	0.41011
3	3.31241	0.89357	0.72022	0.51352
4	3.28732	0.70242	0.84793	0.47736

Table 2: Optimal values of parameters for multiple linear regression equations provided by the Breeder genetic algorithm

Quadrant	Linear regression equations	
	Original MLRE	Optimized with Breeder
1	0.546	0.572
2	0.509	0.510
3	0.671	0.671
4	0.625	0.664

Table 3: The correlation coefficients r for multiple linear regression equations

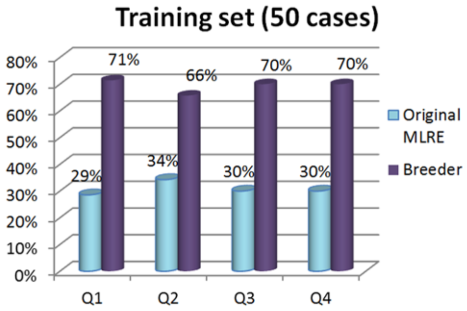


Figure 1: Predictions on the training set

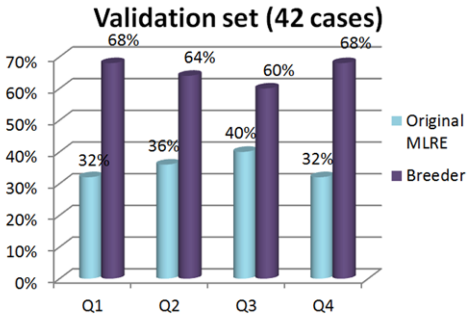


Figure 2: Predictions on the validation set

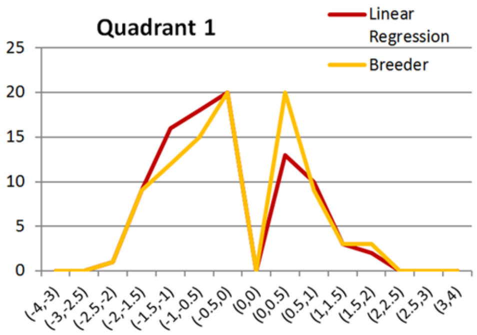


Figure 3: The comparison of prediction error in quadrant 1

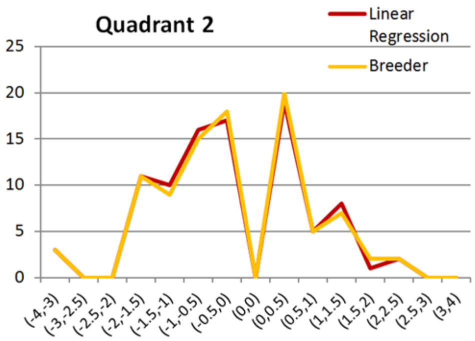


Figure 4: The comparison of prediction error in quadrant 2

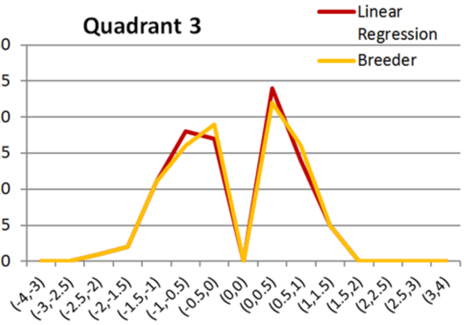


Figure 5: The comparison of prediction error in quadrant 3

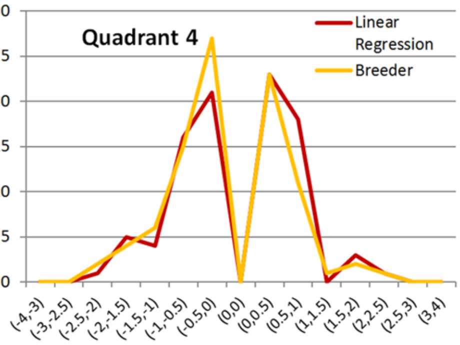


Figure 6: The comparison of prediction error in quadrant 4

4 Comparative analysis

The optimization using the Breeder genetic algorithm was made on all four quadrants, providing the following equations:

$$
\begin{aligned}
Y_{Q1} &= 5.1917 + 0.7571 * X_{42} + 0.85332 * X_{46} + 0.28341 * X_{21} \\
Y_{Q2} &= 5.16292 + 0.90463 * X_{42} + 0.68192 * X_{46} + 0.41011 * X_{21} \\
Y_{Q3} &= 3.31241 + 0.89357 * X_{42} + 0.72022 * X_{46} + 0.51352 * X_{21} \\
Y_{Q4} &= 3.28732 + 0.70242 * X_{42} + 0.84793 * X_{46} + 0.47736 * X_{21}
\end{aligned}
$$

Figure 7: The optimized equations using genetic algorithm

where Y_{Qi} denote the outcome expected for the quadrant $i \in \{1, 2, 3, 4\}$ and X_{mn} represents the mesiodistal width of the tooth specified by index mn.

In our study, if the difference in millimetres between the measured and predicted value of the sum of the mesiodistal sizes of unerupted canines and premolars is situated in interval $[-0.75, 0.75]$, the prediction is considered as a correct estimation, if the difference is < -0.75 mm we have an overestimation, and a prediction error > 0.75 mm is considered an underestimation.

A comparison of correct estimations, overestimations and underestimations, provided by the original MLRE equations [3] and respectively by the optimized equations from figure 7, is presented in the table 4:

Canines premolars group	Method	over-estimations %	correct estimations %	under-estimations %
Maxillary	Original MLRE	35	51	14
	Breeder	32	54	14
Mandible	Original MLRE	24	63	13
	Breeder	21	66	13

Table 4: Correct estimations, overestimations and underestimations in percents

Maximum errors in predicting the sizes of the canines and premolars in the mandible and maxilla are presented in Table 5:

Quadrant	Original MLRE		Breeder	
	over-estimating	under-estimating	over-estimating	under estimating
1	-2.32	1.50	-2.11	1.23
2	-3.97	2.21	-3.56	1.84
3	-2.13	1.14	-1.86	1.01
4	-2.15	2.25	-2.01	1.93

Table 5: Maximum errors in estimating the sum of the mesiodistal sizes of unerupted canines and premolars

Comparing predictions provided by the new and respectively old method, we can conclude that Breeder genetic algorithm is capable to provide the best values for parameters of multiple linear regression equations, and thus our equations are optimized for best performance. The results obtained by the new multiple linear regression equations are significant better than those provided by some classical statistical approaches [3], [15], [17].

The proposed technique is an adaptive tool for predicting the sizes of unerupted canines and premolars with greater accuracy than standard linear regression analyses, the fitness function ensuring optimization of predictions for data collected from different groups selected from different countries.

5 Conclusions

Using a Breeder genetic algorithm, we can automatically find the optimal values for the parameters of multiple linear regression equations used in prediction of the mesiodistal width of unerupted permanent canines and premolars.

After evaluation, we found that our new parameters, used in the regression equations, are providing a better prediction than original MLRE method.

Thus, the prediction error rates of the optimized equations using the Breeder genetic algorithm are smaller than those provided by the multiple linear regression equations proposed in [3].

Using a fitness function related to the prediction error provided by original linear regression equations, the evolution process guided by our implementation of Breeder genetic algorithm was capable to find new MLRE equations which outperform the original equations in terms of qualitative results of the prediction process.

Bibliography

[1] Moyers, R. E.; Handbook of orthodontics Chicago: Year Book Medical Publishers, 1988.

[2] Mühlenbein H., Schlierkamp-Voosen; D-The science of breeding and its application to the breeder genetic algorithm, *Evolutionary Computation*, 1:335-360, 1994.

[3] Boboc, A.; Dibbets, J.; Prediction of the mesiodistal width of unerupted canines and premolars: a statistical approach, *American J. of Orthodontics and Dentofacial Ortopedics*, 137(4):503-507, 2010.

[4] Pancherez, H.; Schaffer, C.; Individual-based prediction of the supporting zones in the permanent dentition. A comparison of the Moyers method with a unitary prediction value *J. Orofacial Orthopedic*, 60(4):227-2355, 1999.

[5] Legovic, M.; Novosel, A.; Legovic, A.; Regression equation for determining mesiodistal crown diameters of canines and premolars, *Angle Orthodontic*, 73(3):314-318, 2003.

[6] Legovic, M.; Novosel, A.; Skrinjaric, T.; Legovic, A.; Madi, B.; Ivancic, N.; A comparison of methods for predicting the size of unerupted permanent canines and premolars, *European Journal Orthodontic*, 28(5):485-490, 2006.

[7] Memon, S.; Fida, M.; Development of a prediction equation for the estimation of mandibilar canine and premolar widths from mandibular first permanent molar and incisor widths, *European Journal Orthodontic*, May 9 [Epub ahead to print], 2011.

[8] Alhaija Abu, E.S.; Qudeimat, M.A.; Mixed dentition space analysis in a Jordanian population: comparison of two methods, *International Journal Paediatric Dentistry*, 16(2):104-110, 2006.

[9] Aquino Melgaco, C.; Araujo de Sousa, M.T.; Roules de Oliveira, A.C. (2007); Mandibular permanent first molar and incisor width as predictor of mandibular canine and premolar width, *American J. of Orthodontics and Dentofacial Orthopedics*, 132(3):340-345, 2007.

[10] Moghimi, S.; Talebi, M.; Parisay, I.; Desing and implementation of a hybrid genetic algorithm and artificial neutral network system for predicting the sizes of unerupted canines and premolars, *European Journal Orthodontic*, 34(4):480-486, 2011.

[11] Van der Merwe, W.S.; Rossouw, P.; Van Wyk Kotze, T.J.; Trutero, H.; An adaptation of the Moyers mixed dentition space analysis for a Western Cape Caucasian population, *The Journal of Dental Association South Africa*, 46(9):475-479, 1999.

[12] Melgaco, C.A.; Araujo, M.T.; Ruellas, A.C.O.; Applicability of three tooth size prediction methods for white Brazilians, *Angle Orthodontic*, 76(4), 644-649, 2006.

[13] Barnabe, E.; Flores-Mir, C.; Apparaising number and clinical significance of regression equations to predict unerupted canines and premolars, *American Journal of Orthopedics and Dentofacial Orthopedics*, 126(2):228-230, 2004.

[14] Barnabe, E.; Flores-Mir, C.; Are the lower incisors the best predictors for the unerupted canine and premolars sums? *An analisis of a Peruvian sample, Angle Orthodontist*, 75(2):202-207, 2005.

[15] Martinelli, F.L.; De Lima, E.M.; Rocha, R.; De Sousa Araujo, M.T.; Prediction of lower permanent canine and premolars width by correlation methods, *Angle Orthodontist*, 75(3):805-808, 2005.

[16] Nourallah, A.W.; Gesch, D.; Khordaji, M.N.; Splieth, C. (2002); New ecuations for predicting the size of unerupted canines and premolars in contemporary population, *Angle Orthodontist*, 72(3):216-221, 2002.

[17] Bonetti, G.A.; Verganti, S.; Zamarini, M.; Bonetti, S.; Mixed dentition space analisis for a northern Italian population: new regression equations for unerupted teeth, *Progress in Ortodontics*, 12(2):94-96, 2011.

[18] Philip, N.I.; Prabhakar, M.; Arora, D.; Chopa, S.; Applicability of the Mojers mixed dentition probability tables and new prediction aids for a contemporary population in India, *American J. of Orthopedics and Dentofacial Orthopedics*, 138(3):339-345, 2011.

[19] Stoica, F.; Simian D.; Optimizing a New Nonlinear Reinforcement Scheme with Breeder genetic algorithm, *Proc. of the 11'th International Conference on EVOLUTIONARY COMPUTING (EC'10)*, Iaşi, Romania, ISSN: 1790-2769, ISBN: 978-960-474-194-6, 273-278, 2010.

[20] Stoica, F.; Cacovean, L.F.; Using genetic algorithms and simulation as decision support in marketing strategies and long-term production planning, *Proceedings of the 9'th International Conference on SIMULATION, MODELLING AND OPTIMIZATION (SMO'09)*, Budapest Tech, Hungary, ISSN: 1790-2769 ISBN: 978-960-474-113-7, 435-439, 2009.

INT J COMPUT COMMUN, ISSN 1841-9836
9(1):71-78, February, 2014.

Energy Efficient Key Management Scheme for Wireless Sensor Networks

N. Suganthi, V. Sumathy

N.Suganthi*
Dept of Information Technology
Kumaraguru College of Technology, Coimbatore-49
*Corresponding author suganthiduraisamy@yahoo.co.in

V.Sumathy
ECE Department
Government College of Technology, Coimbatore-13
sumi_gct2001@yahoo.co.in

Abstract:
Designing an efficient key establishment scheme is of great importance to the data security in Wireless Sensor Networks. The traditional cryptographic techniques are impractical in Wireless Sensor Networks because of associated high energy and computational overheads. This algorithm supports the establishment of three types of keys for each sensor node, an individual key shared with the base station, a pair wise key shared with neighbor sensor node, and a group key that is shared by all the nodes in the network. The algorithm used for establishing and updating these keys are energy efficient and minimizes the involvement of the base station. Polynomial function is used in the study to calculate the keys during initialization, membership change and key compromise. Periodically the key will be updated. To overcome the problem of energy insufficiency and memory storage and to provide adequate security, the energy efficient scheme is proposed. It works well in undefined deployment environment. Unauthorized nodes should not be allowed to establish communication with network nodes. This scheme when compared with other existing schemes has a very low overhead in computation, communication and storage.

Keywords: key management, sensor nodes, polynomial function

1 Introduction

These tiny sensor nodes, which consist of sensing, data processing and communicating components, leverage the idea of sensor networks based on the collaborative effort of a large number of nodes. Sensor nodes are deployed in hostile environments or over large geographical area. The nodes could either have a fixed location or could be randomly deployed to monitor the environment. The nodes then sense environmental changes and report them to other nodes over flexible network architecture. They have thus found application domains in battlefield communication, homeland security, pollution sensing and traffic monitoring. The limited factors of using sensor nodes are that they have limited battery power and less memory capacity. To control information access in a sensor environment only authorized node must know the key to disseminate the information that is unknown to the compromised nodes. The communication keys may be pair wise [7],Chan,Du or group wise [1], these keys to be updated to maintain security and resilience to attacks. Some of the proposed work was based on static schemes [7]Liu and some are on dynamic schemes [1]Eltoweissy Though many protocols have been designed for the purpose of security in sensor environment, unfortunately, node compromising is rarely or not enough investigated and most of these protocols have a weak resilience to attack [13].

In this paper, we propose a key management scheme for WSNs in which the pair wise keys and the group wise key are set up through the broadcast information during the network initialization phase and no further message exchange is needed afterwards. Consequently, the communication overhead is very low. Therefore, the compromise of some sensor nodes will not affect any other non-compromised pair wise keys. For the establishment of keys for new nodes, we propose a composite mechanism based on an algorithm in which resource consumption can also be kept very low and also the transmission of information. Here only the polynomial identifier needs to be communicated to the nodes for establishing group key and pair wise key.

The rest of this paper is organized as follows. In Section 2, some related work and their drawbacks are discussed. In Section 3, energy-efficient key management scheme, in Sections 4 and 5, the security and performance of energy efficient key management scheme are analyzed. Finally, Section 6 deals with the conclusion.

2 Related Works

Many pair wise key distribution schemes [5] [7] [8] [9] have been developed for peer to peer wireless sensor networks and heterogeneous network [6] [12] [14].

In one of the hierarchical schemes [1], the base node calculates the group key using partial keys in bottom up fashion. The partial key of the child node is generated using random number and which is passed to its parent to calculate its partial key which further goes in bottom up fashion finally to calculate the group key. The partial keys are calculated by using a function. The function is expressed as

$$f(k_1, k_2) = \alpha^{k_1+k_2} mod\ p \tag{1}$$

p is the prime number, k1, k2 are the partial keys

The decision for choosing a number of partial keys is based on the key size for the security requirements and the corresponding energy consumption. To guarantee that all the nodes in a group received the information, they send the reply (REP) message. If the cluster head does not get the (REP) from all the node, it re-broadcasts.

When a new node joins the group, the group key is recalculated and again the cluster head broadcasts the newly created group key to all the nodes in the group. The same is repeated when a node leaves the group. This makes the old node, which is deleted, not to know the new key that is created. Also communication takes places between two cluster heads. Due to poor memory capacity and low power of sensor nodes it will be difficult to store all the partial keys and the communication becomes costly as it needs to broadcast the group key once it is created and changed. In our energy efficient key management protocol scheme the group key need not be broadcasted each time.

A tree based key management protocol [2] in which each sensor node is pre-deployed with three keys. One of the keys is used for initial communication i.e for key exchange and tree spanning. After the tree is spanned, this key is deleted from the memory of the sensor node. Then for further communication, the remaining two keys are used. One of these keys is symmetric and is used to encrypt (or decrypt) information sent from the child to the parent. The third key is also symmetric and is used to encrypt (or decrypt) information sent from the parent to the child. The two keys are used to make the task of cryptanalysis attacker difficult. The disadvantage of this scheme is that an attacker gaining access (physical) to the sensor node can obtain the information. Xing Zhang et al [11] proposed an energy efficient distributed deterministic key management protocol (EDDK). Though this scheme provides higher security than the above two schemes namely hierarchical and tree-based protocol, it requires large memory to store data. Sencun Zhu et al [10] proposed a LEAP: efficent security mechanisms for large scale distributed sensor networks and Du et al [11]proposed a scheme using depolyment knowledge.

3 Proposed Scheme: Energy-Efficient Key Management Scheme

To fix the flaws present in the existing key management schemes, we propose an energy efficient key management scheme for WSNs. This scheme mainly focuses on the establishment and maintenance of the pair wise keys as well as the group keys. Unlike hierarchical and tree-based key management schemes, this scheme does not require additional memory for storing the keys after deployment. In this scheme, no key is broadcasted. Each and every node generates the group key and pair wise key using one of the polynomial functions. The polynomial function is identified with the help of its ID. To enhance message security in the network initialization phase, each sensor node makes use of its own individual key. The base station computes the individual key of all the nodes using the unique keys and the IDs that it has stored. This method also enhances security in data transmission with periodic key update. To avoid the replay attack, sequence number is used.

3.1 Overall System Description

The diagram (Figure 1) shows the overall system description. Key will be established after node initialization. If any node joins in the system or a node is compromised, key update will be performed. After detecting the compromised node the keys will be removed from the node memory. The overall system is designed to reduce the computation overhead and it requires less communication between nodes.

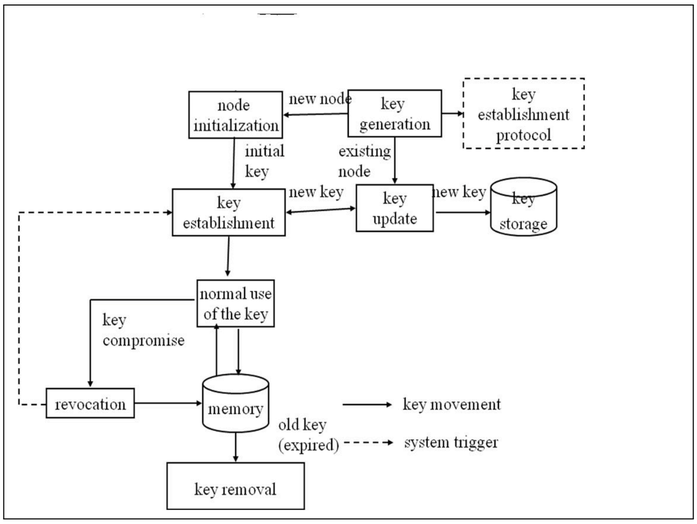


Figure 1: system model

3.2 Spanning of tree

Before deployment, every sensor node is pre-distributed with a network wide shared pseudo-random function (Rf) and an initial key. It is assumed that each node is tamper proof, and it will not be affected by capture attack. The pseudorandom function and the initial key is used by every node in the network to compute its own individual key. This individual key is used for initial communication with the base station. For example, node A s individual key can be computed as per equation (2).

$$K_a = R_f(K_I, Id_a). \tag{2}$$

KI - Initial key, Ida - Individual node identifier

Ka - Individual key shared with base station, Rf -shared pseudorandom function

The hello message which is used to span the tree is also encrypted using this individual key. The hello message contains the ID of the sender and the HELLO keyword. The base station broadcasts hello message. The nodes that reply to hello message become the children of the base station. Here acknowledgment of the child node is essential to accept the node as the child. These nodes then broadcast hello message to other nodes. The nodes that reply to hello message become the children of these nodes. The spanning of the tree is stopped when the nodes do not get a reply to the hello message.

3.3 Key Establishment Phase

Base station will transmit function identifier and random number, encrypted with individual key to individual nodes as shown in equation (3).

$$E_{ka}(Pf_{id}, R_n) \tag{3}$$

Pfid function identifier , Rn random number

Pair Wise key

After sensor nodes are placed in the sensor field, each sensor node communicates with its neighbor via the pair wise key. This key is used to make sure that the message to the intended neighbor node is not known by other neighbor nodes. Nodes A and B are used to show the calculation of the pair wise key using equation (4). Let us consider that node A wants to communicate with node B.

$$K_{ab} = Pf_{id}(R_n, Id_a) \tag{4}$$

Kab Pairwise key, Rn Random number, Ida Identifier of node a

The polynomial function takes the random number and the ID of the node which initiates the communication as input to calculate the pair wise key. The ID of the polynomial function is used to identify one of the functions from the set of polynomial functions. The base node communicates the random number and the ID of the polynomial function to all the nodes. After calculating the pair wise key, it will transmit the message encrypted with this key to the node B along with its ID in plain. Then node B will calculate the pair wise key as it has all the information needed for calculation. Using the key it will decrypt the message transmitted by node A. Here each node, thus, calculates the pair wise key by knowing initiator ID. As no transmission of ID or key information takes place, communication overhead is avoided here.

Group key

A group key is a key shared by all the nodes in the network, and it is needed when the base station is distributing a secure message, (e.g. a query on some event of interest or a confidential instruction) to all the sensor nodes in the network. In a conventional method the parent encrypts M with its cluster key and then broadcasts the message. Each neighbor receiving the message decrypts it to get the message and re-encrypts with its own cluster key, and then transmits the M. This process is repeated until all the nodes receive the message. However, this method has a drawback. In this method, every node has to encrypt and decrypt the message, thus consuming a large amount of energy on computation. So encryption using group key is the most desirable one from the performance point of view. The simple way to store the group key for a node is to preload every node with it. An important problem that arises immediately is the secured updation of this key when a compromised node is detected. In our proposed scheme, to enable base station-individual nodes communication, group key is used. The group key generator that is present in all the nodes is used for generating the group key using the equation(5). The random number that is transmitted to each node is also involved in key calculation. The group key is calculated as follows.

$$K_g = Pf_{id}(R_n, G_k) \tag{5}$$

Kg - group key , Pfid - polynomial function , Rn - random number, Gk - group key generator

The timer is set for the reestablishment of the key. When the timer reaches the threshold value that is assigned, the re-keying is done. Here re-keying is done by changing the coefficient of the polynomial function. The previously calculated keys are deleted periodically. Therefore, even if an adversary could compromise some legitimate nodes, it still could not compute the pair wise keys and the group key. Note that each sensor node only needs to broadcast one communication message during the key establishment phase with no further message exchange required for key calculation. Thus, the communication overhead can be very low. During the data transfer phase the sequence number is used to indicate the message transfer between the nodes. Once the sequence number reaches the threshold value that is already set, the sequence number is reset to 1 again and it can prevent the replay attacks. This sequence number helps to know the number of the messages sent and received. It is also used for receiving the acknowledgement.

3.4 Key Update Phase

Pair wise key and group key should be updated to avoid cryptanalysis and to prevent attacks from adversaries after one or more sensor nodes are compromised. And also after the threshold time, nodes need to update the keys using the same formula as mentioned in equation (4) and (5). The coefficient of the polynomial function is changed. It is done by adding the constant with the previous values and the modulus is taken to be used as new coefficients. Thus it makes the key update easily and avoids communication overhead.

4 System Analysis

For system analysis, we have implemented the key management algorithm in matlab. Compared with the EDDK, Tree based protocol and hierarchical scheme.

4.1 Computation Costs

Computation costs are measured in terms of number of encryptions required to change the keys in the event of node compromise and node addition. When a node is added, only the new random number and new id of the polynomial function will be transmitted by base station. Individual nodes will receive it and compute the group key and pair wise key using the formulas. Calculation using the polynomial function will consume less energy. But in other schemes lot of encryption and decryption is involved to get the key update. It consumes lot of battery energy. The comparison is made between the existing schemes and the proposed scheme in terms of time. This graph (Figure 2) shows the computation time differences between various schemes and proposed scheme depending upon the number of nodes during initial key calculation.

4.2 Memory Requirement for Key Storage

Let x represent the number of neighboring nodes around a sensor and n be the number of polynomial functions. Each sensor node has x storage units for the pair wise keys, nt + n

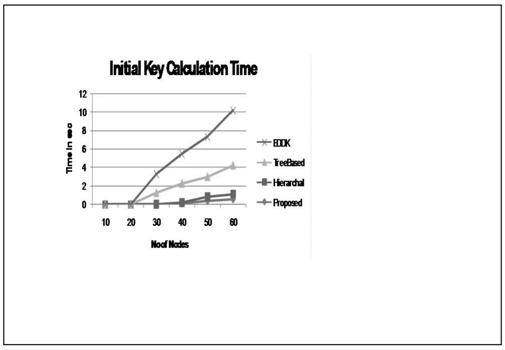


Figure 2: Time required to calculate initial key

storage units for the t-degree polynomial functions, two storage units for random number and pseudo random function and single storage unit to store the group key. In terms of memory requirement to store keys for each scheme, the proposed scheme needs less memory, hence it provides scalability. Even when the number of nodes increases, the memory required to store the keys remains the same. On the contrary the tree based protocol requires more memory as the number of nodes increases. The following graph (Figure 3) shows the comparison of memory requirement for all schemes. The proposed scheme need less memory even when number of nodes increases.

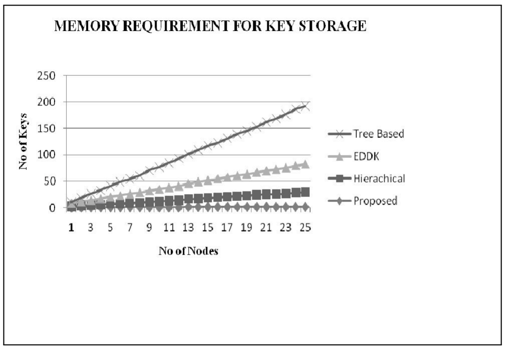


Figure 3: Memory requirement for key storage

4.3 Communication Overhead For Key Exchange

Communication cost is measured in terms of number of messages needed to be exchanged in order to update the existing keys as a result of events like; addition of new node, node compromise, and key refreshing at regular intervals. The communication overhead for the existing scheme is more when compared to the proposed scheme. This is because they need to exchange the keys to enable communication. The proposed scheme requires transmitting the IDs of the polynomial function and random numbers only. So overhead for key exchange is minimal.

As shown in Figure 4, the number of nodes increases, the communication overhead increases for tree based protocol; whereas EDDK and proposed scheme require less communication overhead.

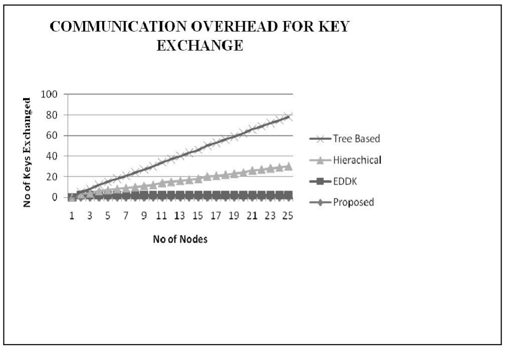


Figure 4: Communication overhead for key exchange

5 Security Analysis

Adversaries with a single compromised node and adversaries with n compromised nodes are chosen selectively. In all cases we study their impact on the desired network properties assuming that the adversary acts maliciously at different layers of the communication protocols. Insiders are adversaries that can compromise nodes or otherwise have a valid identity in a network with appropriate key material. Insiders therefore have the same capabilities as outsiders plus the ability to participate in the network protocols and deviate from the normal behavior of the protocols. Stronger security considerations have to be taken into account for insiders. A minimum level of fault tolerance has to be designed into the network inside attackers. But in our proposed algorithm every node is loaded with the set of polynomials, and every time one function will be used to calculate the key. Adversary nodes cannot generate this polynomial functions. And also it doesn't know which function will be used to calculate the key at that time. The random number will be communicated to the individual nodes by the base station after encrypting with secret key shared by the base station and the individual node.

6 Conclusion

The key exchange problem for sensor networks has been introduced and it is believed that information can be secured by not exchanging the keys directly. A mechanism that makes use of pre-deployed functions has been proposed to fulfill our idea. By using this mechanism, the impacts of many attacks in wireless sensor networks can be limited. This scheme incorporates mechanism that allows for the scalability of memory in the sensor nodes. By comparing the proposed scheme with the existing scheme, it becomes clear that memory required to store the key information is less. Similarly the communication overhead to exchange the keys is also low. Because of the increase in the complexity of the algorithm, the immunity of the sensor networks towards various attacks has been greatly increased. Thus the proposed naming mechanism shields the network from various attacks.

Bibliography

[1] Biswajit Panja; Sanjay Madria; Bharat Bhargava; Energy-Efficient Group Key Management Protocols for Hierarchical Sensor Networks,*Int. J. of Distributed Sensor Networks Taylor Francis Group*, 201-223, DOI:10.1080/15501320701205225, 2007.

[2] Messai,L.; Aliouat,M.; Seba,H.; Tree Based Protocol for Key Management in Wireless Sensor Networks, *EURASIP J.on Wireless Communications and Networking*, Article ID 910695, DOI:10.1155/2010/910695, 2010.

[3] Xing Zhang; Jingsha He; QianWei; EDDK: Energy-Efficient Distributed Deterministic Key Management for Wireless Sensor Networks,*EURASIP J. on Wireless Communications and Networking*, Article ID 765143, DOI:10.1155/2011/765143, 2011.

[4] Eltoweissy,M.; Moharrum,M.; Mukkamala,R.; Dynamic key management in sensor networks, *IEEE Communications Magazine*, 44(4):122- 130, 2006.

[5] Du,W.; Deng,J.; Han,Y.S.; Varshney,P.K.; Katz,J.; Khalili,A.; A pairwise key predistribution scheme for wireless sensor networks, *ACM Trans. on Information and System Security*, 8(2):228-258, 2005.

[6] Jen Yan Huang; I-En Liao; Hao-Wen Tang; A Forward Authentication Key Management Scheme for Heterogeneous Sensor Networks, *EURASIP J. on Wireless Communications and Networking*, Article ID 296704, DOI:10.1155/2011/296704, 2011.

[7] Eschenauer,L.; Gligor,V.D.(2002); A key-management scheme for distributed sensor networks, *Proc. of the 9th ACM Conference on Computer and Communications Security*, Washington, DC, USA, 41-47, 2002.

[8] Chan,H.; Perrig,A.; Song, D.; Randomkey predistribution schemes for sensor networks, *Proc. of IEEE Symposium on Security And Privacy*, 197-213, 2003.

[9] Liu,D.; Ning, P.; Establishing pairwise keys in distributed sensor networks,*Proc. of the 10th ACM Conference on Computer and Communications Security (CCS 03)*, Washington, DC, USA, 52-61, 2003.

[10] Sencun Zhu; Sanjeev Setia; Sushil Jajodia (2003); LEAP: Efficient Security Mechanisms for Large Scale Distributed Sensor Networks, *Proc. of the 10th ACM Conference on Computer and Communications Security*, pp.62-72.

[11] Du,W.; Deng,J.; Han,Y.S.; Chen,S.; Varshney,P.; A Key Management Scheme for Wireless Sensor Networks Using Deployment Knowledge, *Proc. IEEE INFOCOM04*, 586-597, 2004.

[12] Kausar,F.; Hussain,S.; Yang,L.T.; Masood,A.; Scalable and efficient key management for heterogeneous sensor networks, *J. of Supercomputing*, 45(1):44-65, 2008.

[13] Xiao,Y.; Rayi,V.K.; Sun,B.; Du,X.; Hu,F.; Galloway,M.; A survey of key management schemes in wireless sensor networks,*J. of Computers Communications*, 30(11-12):2314-2341, 2007.

[14] Du,X.; Xiao, Y.; Guizani, M.; Chen,H.H.; An effective key management scheme for heterogeneous sensor networks, *J. of Ad Hoc Networks*, 5(1):24-34, 2007.

INT J COMPUT COMMUN, ISSN 1841-9836
9(1):79-84, February, 2014.

Efficiency of a Combined Protection Method against Correlation Power Analysis Side-Attacks on Microsystems

H.-N. Teodorescu, E.-F. Iftene

Horia-Nicolai Teodorescu*, Emanuel-Florin Iftene
1. "Gheorghe Asachi" Technical University of Iasi Romania, Iasi, 8 Bd. Carol I, and
2. Institute of Computer Science of the Romanian Academy, Romania, Iasi
*Corresponding author:hteodor@etti.tuiasi.ro

Abstract: We analyze the efficiency of the masking of instruction patterns using a chaotic driven clock and power supply, in front of a side attack intruding the power supply of a microsystem. The differential analysis is supposedly conducted by correlation power analysis. We demonstrate that the use of a chaotically-driven masking based on relatively simple circuits may be a significant candidate for the protection of embedded systems.

Keywords: physical security, protection, hardware, side attack, chaos, control signal, security evaluation.

1 Introduction

With a field less than 20 year old (the first paper, by Paul C. Kocher [1], was published in 1998, with the first significant expansion published in 2000, [2]), the protection against hardware-level attacks of the information in microsystems, including embedded systems is fast developing, due to the huge interest of the banks, security companies, card manufacturers, and military, moreover due to the interest in power minimization [3]. Citing [4], 'Side Channel Analysis is a [$\cdots$] form of attack [$\cdots$] that uses information that leaks, unintentionally, from the real-world implementations of cryptographic hardware.' Side-channel attacks (SCA) extract and decode the executed instructions and the manipulated data in microsystems, bypassing the cryptographic protections [1], [2]. The basic methods of attack were named simple power analysis (SPA), respectively differential power analysis (DPA), depending on the details of the attack. For various approaches of DPA, see [5].

While the literature includes numerous papers on the attacks and on the sibling topic of power analysis for software optimization [3], understandably fewer papers present hardware methods of mitigating these attacks. Several manufacturers include various solutions against side attacks. For example, Newell and Juliano [4] cite FreeScale Inc., who uses 'patented DPA functions, licensed from Cryptography Research.' Other manufacturers, as MAXIM Inc. and INFINEON also use various protection means, but details on them are not public. For example, the 32 RISC 'DeepCover Secure Microcontroller' MAX32590 released in 2013 by MAXIM includes on the chip, according to the manufacturer data- sheet, a 'tamper detection controller' that 'monitors voltage, frequency, temperature, die shield, and external sensors', erasing essential information when any type of suspicious external activity is detected.

In [6], [7], and [8] we introduced the masking of the instructions using a chaotically-driven clock and power supply. However, a detailed analysis of the masking efficiency under DPA has not been performed for that method. In this paper, we provide results of the analysis of differential power attacks, when the chaotic masking as above is used to protect the system. The protection method proposed in [6], [7], and [8] and further analyzed here is, at the hardware level, more of a proof of concept of the capabilities of the method, not a blueprint solution ready to put into silicon.

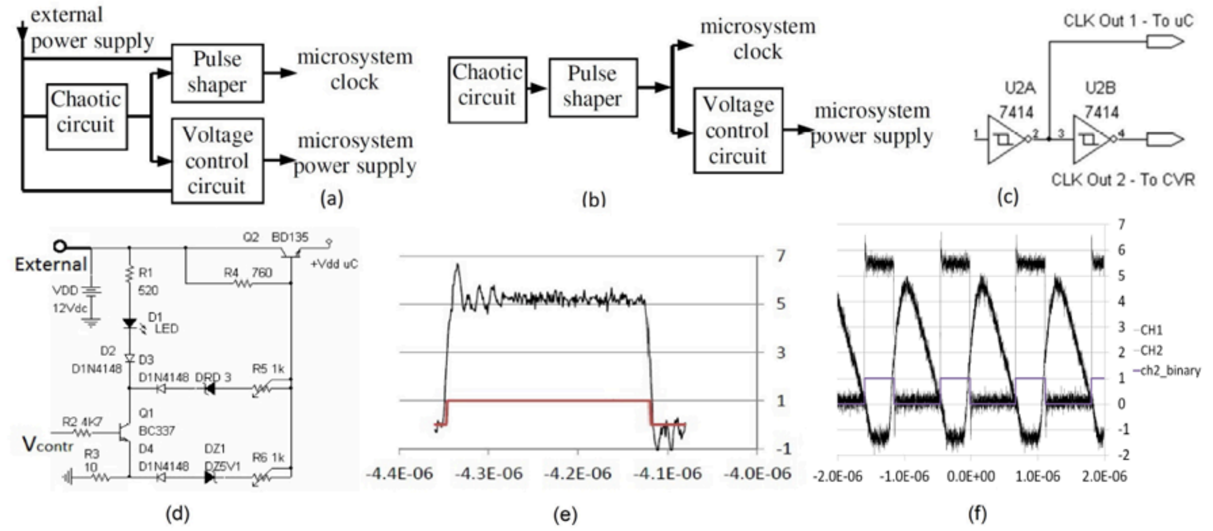


Figure 1: Block diagram for the hardware for protection against SCA. (a) Diagram with analog control of the voltage. (b) Both the clock and the voltage control are digital signals. (c) Pulse forming circuit. (d) Controlled voltage regulator. (e) Example of noisy control pulse. (f) Waveforms from the chaotic circuit and the corresponding pulses generated for control

2 The proposed protection method

Beyond software protection methods, such as specific algorithms, hardware protection methods play an essential role today in data and code security, as proved by several specific chips produced by companies as the ones quoted in Section 1.

We present in this paper an operation principle demonstrator of the protection method. We assume that only the external power supply is available to the intruder for monitoring with a series resistor. While the proposed method is similar to the typical injection of (pseudo-)random pulses on the power supply line, in this type of protection the random signal produced by the chaotic circuit is used to drive a controlled voltage regulator (CVR), which modifies the voltage that powers the microcontroller, moreover is used to generate the clock signal of the system. The protection circuits include a chaotic signal generator, a pulse shaper, and the CVR, as shown in Fig. 1 (a-c). The CVR designed and used in this research is shown in Fig. 1 (d) and an example of control pulse in Fig. 1 (e). The circuits where described in [9] and [8]. More than one level of voltage jump can be produced with such a scheme, provided that multiple loops with different Zener diodes are used in parallel on the lower branch of the circuit in Fig. 1 (d).

The random character of the pulses produced by the pulse shaper refers to the variation of their duration, especially on the long run, due to the change of operation condition of the chaotic circuit (changes in the ambient temperature, fluctuations of the power supply of the chaotic circuit.) For recordings spaced in time by about 10 minutes, under apparently unchanged laboratory conditions, variations of the number of samples per pulse were of more than 20 % for measurements performed during the same day. On the other hand, changes from one pulse to the other were less than 0.5 %. The slow change of the pulse duration due to the change of the chaotic regime produced by ambient factors is beneficial for the protection because it makes difficult the learning of the patterns of the instruction, as they continuously and unpredictably change. Notice in Fig. 1 (f) that the control pulses remain noisy with an amplitude of the noise of about 0.5 V (peak amplitude more than 1 V). This high frequency noise makes the masking process more effective, therefore we have not tried to reduce the noise. The evaluation of the randomicity of the clock signal, as produced by the pulse shaper of the chaotic clock generator proposed in [8] was performed by determining the fluctuations of the width of the pulses. For this purpose, the time between two successive up-down impulse edges was determined for all pulses during a long period of time. The analysis showed that the width of the pulses varies by about 0.1 to 2% over short periods (less than 1 ms), but with almost 40% over longer periods (minutes to hours). The presented circuits serve only to illustrate the operation principle and the feasibility. These circuits are not designs for on-chip implementations. On the other hand, the hardware-level protection discussed in this paper assumes that the protection circuits are built on the same chip as the protected microsystem (SoC - system on chip technology), or at least that they are on a chip included in the same package (multi-chip technology).

3 Analysis of the strength of the protection method

Various attack methods and related countermeasures were presented in the literature, see [5], [10], [11], [12]. Attacks based on correlation analysis are among the most common. The inter-correlation function for two sequences, $\{x(n)\}$ and $\{y(n)\}$, is defined as $C_{x,y}(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} x[n]y$ $[n+\tau]$, where τ is the delay. Let $V_k = (s_N, s_{N+1}, , s_{N+T})$ be the expected (average) vector standing for the pattern of an unperturbed (unmasked) instruction k. Let $X = (x_M, x_{M+1}, , x_{M+T})$ be the vector of an instantiation of a masked, unknown instruction. The duration (number of samples) is taken the same as for V_k, when the clock can be determined independently and the number of clock periods for an instruction is known. The purpose of the attack is to identify the instruction from its signature, X. Several approaches for the attack are possible, among others the determination of the distance between X and all the patterns of the instructions, V_k, $k = 1 \cdots n$, the computation of the inter-correlations between all V_k and X, or determining the distances between the Fourier transforms of the unknown, masked sequence X, $F(X)$, and the Fourier transform of the sequence of the instructions, $F(V_k)$. Some authors, e.g. [5], consider the correlation power analysis (CPA) a distinct, more advanced method than DPA.

When using correlation functions, attackers may try to determine the instruction in various ways, depending on the information they can acquire about the microsystem. When the attackers are able to determine the patterns of the unmasked instructions, they could proceed as follows. The attackers may compute in the first place the intercorrelation functions between the 'clean' patterns of instructions and segments of the waveform that correspond to one machine cycle (m.c.), assuming the instructions take one m.c. The attackers may reason that the true instruction is there where the intercorrelation is the greatest. Denote the 'clean' pattern of the instruction $\#k$ by X_k^0. We denote an instance of the masked instruction $\#j$ by X_j^m. The correlation between them is denoted by $C_{X_k^0,X_j^m}(t)$. In the simplest (ideal) case, $max_t C_{X_k^0,X_k^m}(t) \ggg max_t C_{X_k^0,X_j^m}(t), j \neq k$. Then, the instructions are easily identifiable. If, instead, there is some index j such that $max_t C_{X_k^0,X_k^m}(t) < max_t C_{X_k^0,X_j^m}(t), j \neq k$, confusion appears between the instructions $\#k$ and $\#j$.

In the next Section, we demonstrate that a key sub-set of the instruction set of the microcontrollers in the 16FXXX series is securely masked by the method we proposed in [7], [8] against CPA analysis. For this purpose, we compute the correlation functions between the waveforms produced by various instructions when they are masked, respectively unmasked.

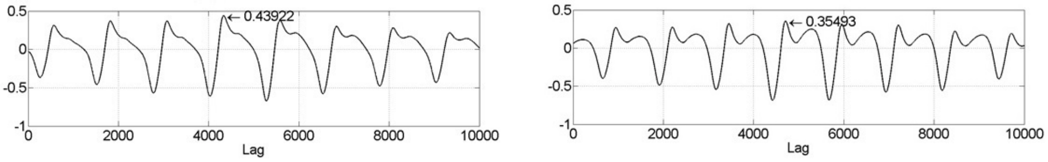


Figure 2: Examples of intercorrelations with masked instructions

4 Results and assessment of the robustness against CPA-SCA

The main results refer to the efficiency of the masking as determined by the lack of correlation between the unmasked pattern of the instruction and the masked ones. The results are summarized in Tables I and II. Table I shows that the maximal values of the self-correlations of unmasked instruction waveforms is (almost) 1 for all the instructions, as expected, while the maximal values of the intercorrelations between unmasked instructions is less than 0.8. This allows the easy discovery of the unknown instruction by performing the correlation of its waveform with the waveforms of the other instructions. In our case, as Table II shows, seven out of eight instructions have maximal values of intercorelations with other instructions than with themselves. For example, the instruction **movlw**, when masked, better intercorrelates with **addwf**, **andwf**, **movf**, **rrf**, and **btfss** than with itself (larger correlation values, see Table II).

TABLE I. Maximal values of the inter-correlation functions for eight instructions; unmasked operation, at 4 MHz clock

	addwf#2	andwf#7	movf#4	rrf#3	btfsc#4	btfss#3	andlw#2	movlw#7
addwf#2	1	0.73801	0.73584	0.60686	0.39528	0.75491	0.66487	0.78685
andwf#7	0.73801	1	0.85923	0.60065	0.4209	0.77955	0.73043	0.8248
movf#4	0.73584	0.85923	1	0.5919	0.39232	0.76756	0.67754	0.80839
rrf#3	0.60686	0.60065	0.5919	1	0.81175	0.57169	0.5507	0.57
btfsc#4	0.39528	0.4209	0.39232	0.81175	1	0.40628	0.46722	0.38796
btfss#3	0.75491	0.77955	0.76756	0.57169	0.40628	1	0.85267	0.73405
andlw#2	0.66487	0.73043	0.67754	0.5507	0.46722	0.85267	1	0.68391
movlw#7	0.78685	0.8248	0.80839	0.57	0.38796	0.73405	0.68391	1

TABLE II. Maximal values of the inter-correlation functions between eight instructions, when one instruction is unmasked (first column in the table) and the other one is masked (first row).

	addwf#3	andwf#7	movf#3	rrf#6	btfsc#6	btfss#5	andlw#4	movlw#4
addwf#2	0.4508	0.30583	0.33292	0.27566	0.37544	0.37642	0.30343	0.44512
andwf#7		0.34256	0.36752	0.28884	0.41852	0.39059	0.30863	0.48923
movf#4			0.36011	0.29453	0.38931	0.40252	0.32866	0.45862
rrf#3				0.23007	0.47391	0.29884	0.27176	0.51101
btfsc#4					0.39219	0.22641	0.20155	0.40093
btfss#3						0.3588	0.28241	0.45588
andlw#2							0.21974	0.33482
movlw#7								0.44189

Notice that the first table is symmetrical with respect to the main diagonal (Hermitian). Tables I and II should be considered from the point of view of the identification of the instruction based on correlation functions. Each element of the tables (matrices) is the maximal value of the correlation function for specified execution instances of a first and second instructions. The instruction is identified when the correlation is 1, in Table I. The attacker is supposed here to have access to the true waveform of the instruction and to be able to directly or indirectly determine the clock frequency of the attacked system.

Assume that the attackers have acquired the waveforms of the non-masked instructions. The attackers can determine the true clock frequency of a system by running in a loop the correlation between an interpolated, respectively extrapolated version of the clean waveforms with the unknown, masked waveforms. For some value(s) of the interpolation, the correlation function exhibits the maximal highest value and a strong periodicity, due to the machine cycles in the waveform. In that case of interpolation, the time alignment between the known 'clean' waveform and the given waveform with unknown clock is the best and, therefore, the unknown clock period is found. We assume that the attackers have performed this determination. Next, the attackers can perform with a known interpolation factor all the intercorrelations, to extract the information on the instructions in the attacked program.

Figure 2 shows examples of self-correlations of unmasked and masked instructions and correlations between masked and unmasked instructions. As expected, in all cases the correlations exhibit the periods of the clock and of the machine cycles, but not the instruction patterns. Notice in Table II that values of selfcorrelation for a specified instruction that are lower than values of the correlation of the same instruction with others means that the criterion of maximal value of correlation will not work for the discovery of the instruction, based on correlations.

5 Discussion and conclusions

This paper synthesized partial and preliminary results reported in [6], [7], [8] and presented a thorough analysis of the masking efficiency under CPA attacks against a microsystem protected with the masking method proposed. The method is based on the randomization of both the clock and the supply voltage. The randomization uses an approach based on a simple chaotic system and the related circuitry.

The proposed protection can be effective only when the attacker has no access to the chaotic circuit, or to the controlled voltage regulator. These circuits should be included in the same package as the microsystem. Moreover, the electromagnetic radiation (EMR) from the CVR should not be easy measured, because it reveals to the attacker the control of the voltage (that is, the chaotic circuit output). With the chaotic signal known, the attacker would be able to demodulate the masked signal and the masking one. In addition to limiting the direct and indirect (EMR mediated) access to the chaotic signal, the protection must insure that the modulating signal and the protected one have similar characteristics, for example, similar amplitudes and heavily overlapping spectra. Only with all these conditions satisfied, could the protection be effective. We reported only on an idea demonstration, not on an effective circuit. Therefore, neither the condition on the amplitude of the swings of the VCR, nor the overlapping spectra condition is satisfied.

Concluding, we presented a method for instruction masking against CPA and showed that the method proves highly effective even with simple circuits for protection. The core of the method is the use of a chaotic circuit to alter at the same time the clock frequency and the supply voltage of the protected microsystem. The method is appropriate for integration either on the chip of the microsystem or in a multi-chip package.

Acknowledgment. This research was not supported by any institution; most of the conceptual work was benevolently performed by HNT. While there is a partial overlapping of this research with the project 'Cognitive Systems' of the Romanian Academy, there was no support from that project.

Authors' contributions. HNT proposed the protection method in Fig. 1, the schemes of the circuits for chaotic signal generation and power supply control, determined the tests and participated in the tests and experiments, performed part of the data processing, derived conclusions and wrote the paper. EFI built all the circuits based on the design and schemes provided by HNT, wrote the test programs based indicated by HNT, and made most of the experiments. Both authors discussed the paper and agreed with its final form. **Conflicts of interest.** The authors declare no conflict of interest.

Bibliography

[1] P. Kocher, J. Jaffe, B. Jun, (1998), Introduction to Differential Power Analysis and Related Attacks, Cryptography Research Inc, www.cryptography.com/public/pdf/DPATechInfo.pdf. Accessed Jan. 2012.

[2] P. Kocher, J. Jaffe, B. Jun, (2000), Differential Power Analysis, Cryptography Research Inc, www.cryptography.com/public/pdf/DPA.pdf. Accessed Jan. 2012.

[3] V. Tiwari, S. Malik, A. Wolfe, M. T.-C.Lee, Instruction Level Power Analysis and Optimization of Software, *J. VLSI Signal Processing*, 13(2-3):223-238, Aug 1996.

[4] R. Newell, F. Juliano, Protecting Sensitive Networked Embedded Systems from Aggressive Intrusion. EDN, Electronic Design News Magazine, May 5, 2013. www.edn.com/Pdf/ViewPdf?contentItemId=4413418

[5] T.-H. Le, M. Berthier, Mutual Information Analysis under the View of Higher-Order Statistics. In: Echizen, I., Kunihiro, N., Sasaki, R., (Eds.), Advances in Information and Computer Security, *LNCS*, Springer, Berlin Heidelberg, 6434: 285-300. 2010.

[6] H.-N. L. Teodorescu, E.-F. Iftene, Analysis of the Code Masking Efficiency of Chaotic Clocks in Microcontroller Applications, *3rd Int. Symposium on Electrical and Electronics Engineering (ISEEE2010), Sep 16-18, Galati*, 261-266, 2010.

[7] E.-F. Iftene, H.-N. L. Teodorescu, Masking the Instructions of a Microcontroller using a 'Chaotic' Power Supply, Bull. Polytechnic Inst. Iasi, E&E, LIX (LXIII), 1:21-28, 2013.

[8] E.-F. Iftene, H.-N. L. Teodorescu, Protecting the Code against Side Attacks using Chaotically Controlled Clock and Supply, *Proc. ECAI 2013 - 5th Int. Conf. Electronics, Computers and A.I., IEEE Conf. #20924, 27-29 June 2013, Pitesti, Romania*, 79-82, 2013.

[9] H.-N.L. Teodorescu, V. P. Cojocaru, Complex Signal Generators based on Capacitors and on Piezoelectric Loads. In: C. H. Skiadas, I. Dimotikalis and C. Skiadas (Eds), Chaos Theory: Modeling, Simulation and Applications. World Scientific Publishing Co., 423-430, 2011.

[10] E. Brier, C. Clavier, F. Olivier, Correlation Power Analysis with a Leakage Model. In M. Joye and J.J. Quisquater (Eds.), Cryptographic Hardware Embedded System, CHES 2004, Vol. 3156, LNCS, pp. 16-29, Springer-Verlag, 2004.

[11] Y. Zhang, A. Juels, M.K. Reiter, T. Ristenpart, Cross-VM Side Channels and Their Use to Extract Private Keys. ACM, 2012. Available at http://dx.doi.org/10.1145/2382196.2382230, 2012.

[12] R.E. Atani, S. Mirzakuchaki, S.E. Atani, W. Meier, On DPA-Resistive Implementation of FSR-based Stream Ciphers using SABL Logic Styles, *Int J Comput Commun*, ISSN 1841-9836, 3 (4):324-335, 2008.

INT J COMPUT COMMUN, ISSN 1841-9836
9(1):85-92, February, 2014.

Spectrum Migration Approach Based on Pre-decision Aid and Interval Mamdani Fuzzy Inference in Cognitive Radio Networks

Z. Wang, H. Wang, H. Lv, G. Feng

Zhendong Wang*, Huiqiang Wang,
Hongwu Lv, Guangsheng Feng
College of Computer Science and Technology
Harbin Engineering University, Harbin, China
*Corresponding author: wangzhendong@hrbeu.edu.cn

Abstract: This study intends to improve the QoS of SUs and CRNs performance. A novel spectrum migration approach based on pre-decision aid and interval Mamdani fuzzy inference is presented. we first define spectrum migration factors as spectrum characteristic metrics for spectrum migration decision. In addition, we use pre-decision aid to reduce system complexity and improve spectrum migration efficiency. To shorten spectrum migration decision time and seek the optimal spectrum holes, interval Mamdani fuzzy inference is put forward. Finally, simulation results show the proposed approach can inhibit the upward trend of service retransmission probability and average migration times effectively, and improve the effective utilization of CRNs spectrum resource significantly.

Keywords: cognitive radio networks, spectrum migration, pre-decision aid, interval Mamdani fuzzy inference

1 Introduction

Opportunistic spectrum access (OSA) technology due to help SUs utilize the idle spectrum, effectively improve spectrum usage and system throughput for CRNs, and becomes a hotspot that academia concerned [1]- [5]. However, existing OSA technologies mainly focus on improving CRNs system throughput and spectrum resource utilization, the QoS requirement of SUs and the effective utilization of CRNs spectrum resource are little considered, that maybe cause system throughput and utilization of spectrum resource increase to a higher level, but the effective utilization of spectrum resource is still maintained at a lower level [6]- [7].

In order to improve the QoS of SUs and the CRNs performance, this paper introduces the concept of spectrum migration. Spectrum migration means SUs change their using spectrum dynamically for improving the success rate of SUs connections, it describes the whole process of SUs service transmission. The occurrence of spectrum migration includes two cases: (1) PU arrives at the spectrum that SU is using, (2) the quality of spectrum SUs using drops below the minimum value that can maintain normal data transmission [8]- [10]. Under normal circumstance, the probability of spectrum environment deterioration is lower, so we only consider SUs spectrum migration when PU arrives. Actually, frequency spectrum migration operations can decrease system performance and SUs QoS because of the operations is time-consuming. Therefore, the objective of this paper we pursue are the longest occupation time to single spectrum hole, the least spectrum migration times and the shortest spectrum migration decision time for SUs. In this paper, we define spectrum migration factors as spectrum characteristic metrics for spectrum migration decision, and use pre-decision aid to reduce system complexity and improved spectrum migration efficiency. At last, we propose interval Mamdani fuzzy inference (IMFI) method based on Mamdani fuzzy inference to shorten spectrum migration decision time and search for suitable spectrum holes.

2 Spectrum Migration Factors

2.1 Spectrum Occupation Probability

Spectrum occupation probability indicates the spectrum occupation degree for PUs and SUs. For a SU, when the spectrum is being occupied by PUs or other SUs, it can not migrate to the spectrum. Therefore, spectrum occupation includes spectrum occupation include PUs and SUs. From this point, spectrum occupation probability is the occupation probability of sub-spectrum divided, i.e., channel occupation probability. A higher spectrum occupation probability can lead to higher migration blocking probability for SUs. Until now, there is no accurate spectrum occupation model proposed, so we use statistical method to obtain spectrum occupation probability by calculating spectrum occupation data in the past time.

Considering the ON-OFF average time of spectrum η are t^α and t^β over a period of time $t\prime$ ($t\prime = t^\alpha + t^\beta$) respectively, then spectrum occupation probability of spectrum η can be calculated as $SOP = \frac{t^\alpha}{t^\alpha + t^\beta}$. When spectrum η undergoes k times state changes, the spectrum occupation probability can be updated as

$$SOP = \frac{\sum\limits_{i=1}^{k} t^{\alpha,i}}{\sum\limits_{i=1}^{k} (t^{\alpha,i} + t^{\beta,i})} \tag{1}$$

Where $t^{\alpha,i}$ and $t^{\beta,i}$ indicate the ON-OFF time in the i^{th} period on spectrum η respectively.

2.2 Link Maintenance Probability

Link maintenance probability mainly reflects the link support degree for SUs data transmission. It indicates the capacity of continuous data transmission for SUs on a specific licensed spectrum. From Sect. 1 we know only PUs arrival can force SUs vacate the using spectrums and makes link maintenance fail. From this point, link maintenance probability is the same for licensed spectrum and its channels. When a PU arrives, there are three consequences for SUs: (1) SU needs to vacate its using spectrum and migrate to other spectrum to continue its data transmission. (2) SU vacates its using spectrum and waits for PU to leave, then SU continues its data transmission through the original spectrum. (3) SU vacates its using spectrum and link maintenance is failed.

Let P_v denote the probability that an SU vacates its using spectrum, we have link maintenance probability as follows

$$LMR = P_v[(1 - r_b) + r_b P_r(\psi^t < \tau_t)] \tag{2}$$

Where r_b denotes PU call blocking probability, and it is given by

$$r_b = (\rho^M / M!) \Big/ (\sum_{i=1}^{M} \rho^i / i!) \tag{3}$$

Where ρ denotes the PU traffic intensity.

2.3 Spectrum Migration Degree

Spectrum migration degree reflects the pros and cons of the spectrum holes on each licensed spectrum for SUs directly. Before defuzzification operations, it is denoted as a specific level, and after defuzzification operations, it denotes as a specific value. We consider that the spectrum with higher spectrum migration degree, the more suitable for SUs spectrum migration.

Spectrum migration degree can be obtained by SOP and LMP using fuzzy inference, it is expressed as

$$SMD = Inf(SOP, LMP) \tag{4}$$

3 Spectrum Migration Approach Based on Pre-decision Aid and IMFI

3.1 Spectrum Migration Pre-decision Aid

In order to avoid all the SUs enter into fuzzy decision module and reduce CRNs complexity, we propose spectrum migration pre-decision aid method. It can be described as follows

1) When SU arrives at the spectrum for the first time, i.e., $A = A^f$, where A denotes the arrival of SU, and A^f denotes the arrival of SU for the first time. When there are more than one spectrum hole, SU needs to enter into fuzzy decision. If there is only one spectrum hole, SU migrates to it directly, and if all spectrums are occupied, spectrum migration operations will be blocked. The process is shown as Figure 1 (a).

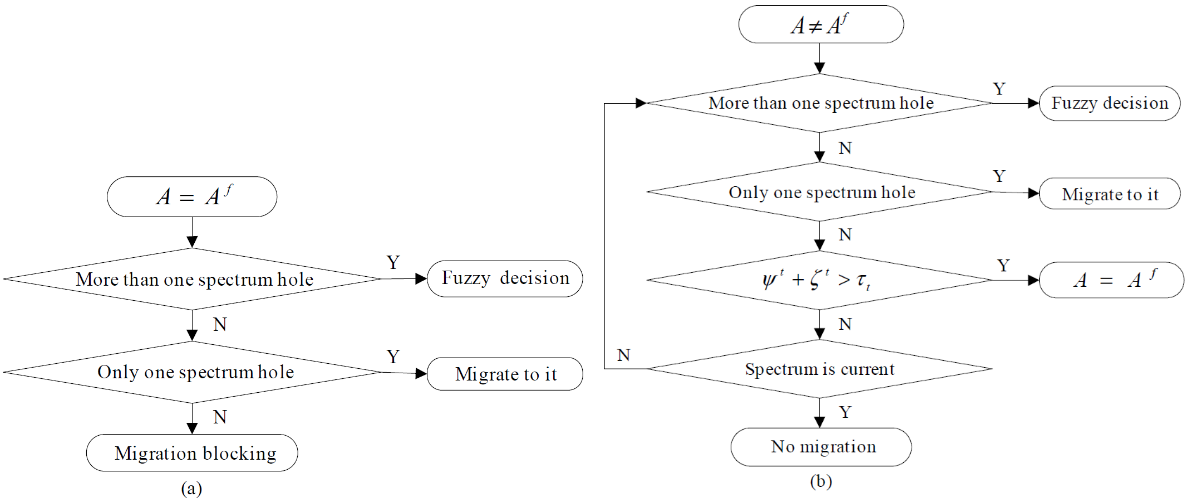


Figure 1: Process of Spectrum Migration Pre-decision Aid

2) When SU arrives and $A \neq A^f$, if there are more than one spectrum hole, SU enters into fuzzy decision module. If there is only one spectrum hole, SU migrates to it directly. If there is no idle spectrum and the occupied duration is more than τ_t , i.e., $\psi^t + \zeta^t > \tau_t$, SU connection will be interrupted, the data has been transmitted will be considered as ineffective. Then, A will be reset to A^f , and the retransmission operation will start for this SU. Otherwise, SU judges whether the idle spectrum is current spectrum, if it is indeed the current spectrum, then SU does not need to migration. Otherwise, SU selects spectrum migration manner according to the number of idle spectrums. This process is described as Figure 1 (b).

3.2 Spectrum Migration Method Based on IMFI

A. Spectrum Migration Factors Fuzzification

In fuzzification process, the more the number of fuzzy sets, the lower the probability of entering into the 2^{nd} decision, then, the spectrum migration decision time can be saved much. However, the number of fuzzy rules is polynomial growth with the growth of the number of fuzzy sets, when SUs enter into the 2^{nd} decision, excessive fuzzy rules will cause longer inference time, which may make SU on the effective use time of spectrum holes shorten, and the system throughput reduced. When $\psi^t + \zeta^t > \tau_t$ happens, it even leads to SU data retransmission. Simultaneously, rare fuzzy sets have fewer fuzzy rules, but they cause the probability of entering into the 2^{nd} decision increase, also extend the fuzzy inference time. In this subsection, we consider *SOP* and *LMP* as input fuzzy parameters, and *SMD* as output fuzzy parameter. Then, we establish the membership functions for *SOP*, *LMP* and *SMD* as Figure 2. In Figure 2, the universes of all the fuzzy variables are set to [0, 1]. For every licensed spectrum, *SOP* and *LMP* have five fuzzy sets, that denote as VL, L, M, H, VH, mean "Very Low", "Low", "Medium", "High" and "Very High" respectively. The fuzzification operations are same for *SOP* and *LMP*, they are shown as Figure 2 (a). Fuzzy variable *SMD* also has five fuzzy sets, and they denote as VS, S, M, B, VB, that mean "Very Small", "Small", "Medium", "Big" and "Very Big" respectively, they are shown as Figure 2 (b).

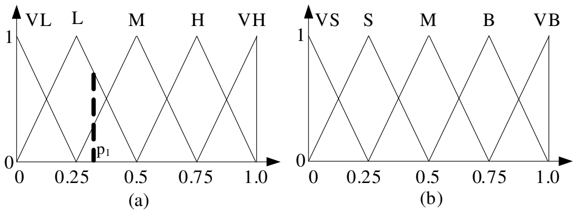


Figure 2: Membership Function of Fuzzy Variables

B. IMFI Method

In fuzzy decision phase, decision time has a big impact on spectrum migration performance. Traditional Mamdani fuzzy inference method calculates each fuzzy rule by max-min mode, makes it compute-intensive, and fuzzy inference time is long. For two-input single-output and 7-divisions fuzzy controller, the inference time accounts for 60% to 80% of the total fuzzy inference time, and the proportion will increase with the increase of the number of rules.

Definition 1. Let the universe of fuzzy variable π is U , its membership function is $F(x)$. If there is an interval $X = [a, b] \subset U$, and its membership function is $f(x)$, $f(x) \in F(x)$, i.e., $f(\pi) = F(\pi)$. Then, we define interval $[a, b]$ as an inference interval for π .

Definition 2. Inference interval $[a, b]$ is an effective inference interval if and only if $f(x) \neq 0$ for any $x \in [a, b]$. Conversely, $[a, b]$ is defined to be ineffective inference interval.

Definition 3. Inference intervals $X_1, X_2, X_3, ..., X_h$ are defined to complete inference interval if and only if $X_1 \cup X_2 \cup X_3 \cup ... \cup X_h = U$.

Theorem 1. *Fuzzy relationship on the universe is equal to fuzzy relationship on complete inference interval.*

Proof: Assuming $A_1, A_2, ..., A_n$ are complete inference intervals on universe U . According to definition 3, we have $A_1 \cup A_2 \cup ... \cup A_n = U$. Assuming the membership functions of $A_1, ..., A_n$ are $\mu_{A_1}(x), ..., \mu_{A_n}(x)$ respectively, and the membership function of U is $\mu(x)$, then, we have $\mu_{A_1}(x) \in \mu(x)$, i.e., $\mu_{A_1}(a_1) = \mu(a_1)$ for any $a_1 \in A_1$ according to definition 1. Similarly, we have $\mu_{A_2}(a_2) = \mu(a_2)$, $\ldots$, $\mu_{A_n}(a_n) = \mu(a_n)$ for $a_2 \in A_2$, $a_n \in A_n$ on the intervals $A_2, ..., A_n$. It completes the proof.

Table 1: Fuzzy Inference Rules

SOP LMP	VL	L	M	H	VH
VL	M	S	S	VS	VS
L	M	M	S	S	VS
M	B	B	M	S	S
H	VB	B	B	M	S
VH	VB	VB	B	M	M

Table 2: Table of Interval Mamdani Fuzzy Inference Decision

SOP LMP	I_1	I_2	I_3	I_4
I_1	(1,2/1,2)	(2,3/1,2)	(3,4/1,2)	(4,5/1,2)
I_2	(1,2/2,3)	(2,3/2,3)	(3,4/2,3)	(4,5/2,3)
I_3	(1,2/3,4)	(2,3/3,4)	(3,4/3,4)	(4,5/3,4)
I_4	(1,2/4,5)	(2,3/4,5)	(3,4/4,5)	(4,5/4,5)

From *Theorem* 1, we can simplify MFI to IMFI.

In order to apply IMFI method to spectrum migration, we formulate spectrum migration fuzzy rules as Table Table 1. Furthermore, according to Figure 2 (a), we make interval Mamdani fuzzy inference decision table as Table 2. For ease of comprehension, we provide detailed information on Table Table 1 and Table Table 2. In Table 3.2, we give an example to explain the expression of fuzzy rules. The three shaded tables with "M", "VL" and "S" stand for the fuzzy rule of "If *SOP* is M and *LMP* is VL, then *SMD* is S". For Table Table 2, I_1 , I_2 , I_3 , I_4 denote effective inference intervals for [0, 0.25], [0.25, 0.5], [0.5, 0.75], [0.75, 1.0] respectively, and 1, 2, 3, 4, 5 stand for VL, L, M, H, VH respectively. It can be seen that $I = I_1 \cup I_2 \cup I_3 \cup I_4 = [0, 1]$, so I is a complete inference interval for the universe. To explain the meaning of Table Table 2, we also use the shaded tables as an example. If the value of *SOP* located at the effective inference interval I_3 , and the value of *LMP* located at the effective inference interval I_2 , I_3 and I_2 denote the effective inference intervals [0.5, 0.75] and [0.25, 0.5], i.e., if $0.5 \le Value_{SOP} \le 0.75$, $0.25 \le Value_{LMP} \le 0.5$, then, we use fuzzy inference rules (3,4/2,3). The number on the left of backslash is fuzzy sets of *SOP*, and number on the right of backslash is fuzzy sets of *LMP*, they correspond to the following four fuzzy rules

If *SOP* is M and *LMP* is L, then *SMD* is S, If *SOP* is M and *LMP* is M, then *SMD* is M

If *SOP* is H and *LMP* is L, then *SMD* is S, If *SOP* is H and *LMP* is M, then *SMD* is S

If we do not use interval Mamdani fuzzy inference, there will be twenty five fuzzy rules used under the same case. Due to the limited space, we will not list the fuzzy rules. For the same inference results, interval Mamdani fuzzy inference can save more than three-quarters of the time compared with Mamdani fuzzy inference under our condition.

C. Spectrum Migration Decision

According to interval Mamdani fuzzy inference, we can get three kinds of inference results: 1) Some of the licensed spectrums have the same *SMD* levels, but there only one spectrum has the highest *SMD* level. 2) Some of the licensed spectrums have the same *SMD* levels, and there are more than one spectrum has the highest *SMD* level. 3) The *SMD* level for each spectrum is different. For 1) and 3), the system selects the spectrum with the highest *SMD* level to migration, we call it the 1^{st} migration decision. For the case of 2), the system should take defuzzification operations and selects the spectrum with the maximum *SMD* value to migration, we call it the 2^{nd} migration decision. Spectrum migration decision can be expressed as

$$Ch^* = \arg\max_{\forall Ch} \Theta_{SMD}(Ch) \tag{5}$$

4 Numerical and Simulation Results

We simulate and evaluate the performance of spectrum migration approach proposed in this paper. Meanwhile, we use the existing approaches that are RANDOM [6], MFI and GREEDY [7] to comparison. The following assumptions are adopted in the simulation. The CRN in which SUs coexist with PUs in a 5km × 5km area. The number of licensed spectrums in the area is 5, and each spectrum is divided to 5 channels. All the spectrums are independent identically distributed. We set the total bandwidth of licensed spectrum is 5 MHz. So, the bandwidth of each secondary channel is 0.2 MHz. This assumption is reasonable since that one voice channel is only 0.2 MHz in GSM cellular network. The signal to interference plus noise ratio (SINR) is set to 3dB. The average arrival rate of SUs is assumed to be 1.0, the retransmission waiting threshold is set to be 1.0s. Simulation data is recorded for 10000 times to avoid the contingency of the results.

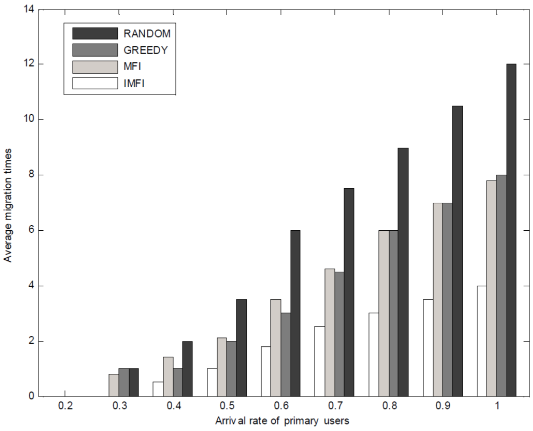


Figure 3: Average Migration Times

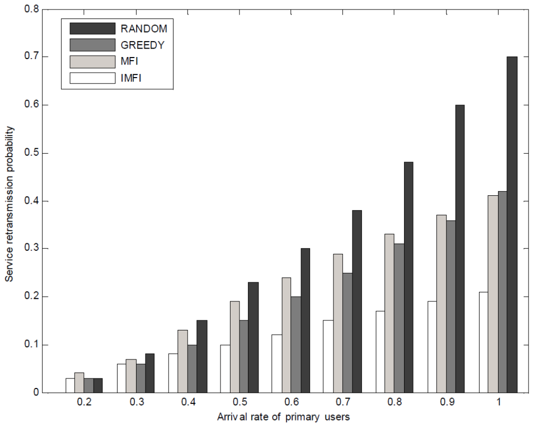


Figure 4: Service Retransmission Probability of SU at τ_t =1.0s

Figure 3 and Figure 4 show the average migration times of SU single service transmission and SU service retransmission probability at τ_t =1.0s respectively. It is obviously that IMFI algorithm has the least average migration times. Compared with RANDOM, the average migration times of IMFI reduce by about 65%, and also reduce nearly half compared with the MFI. Under the premise of service size fixed, average migration times are related to spectrum holes during time, migration waiting threshold and spectrum migration decision time. Compared with RANDOM and GREEDY, the advantage of IMFI reflects the accuracy and timeliness for the optimal spectrum selection. MFI has the same inference results with IMFI, so the advantage of IMFI reflects the timeliness of spectrum migration decision. When τ_t =1.0s, the tendency of service retransmission probability is close to average migration times for the four algorithms, that indicates there exists a positive relationship between SU service retransmission probability and average migration times. When λ =1.0, service retransmission probability only reaches to 20% using IMFI, that is much lower than GREEDY and MFI, which confirms superior performance on the optimal spectrum selection and spectrum migration decision once again.

In Figure 5, we can see that the average throughput of CRNs all decrease with the increase of arrival rate of PUs. This is because the increase of λ makes the spectrum holes duration shorten, and cause the use of spectrum holes tend to be difficult. Compared with RANDOM, the tendencies of GREEDY and MFI decrease obviously, the main reason is the two intelligent algorithms consume an inordinate amount of time for spectrum migration decision. Because of using spectrum migration pre-decision aid, and IMFI method, spectrum migration decision time with IMFI algorithm is shortened greatly. Compared with RANDOM, the average throughput of CRNs increases slightly.

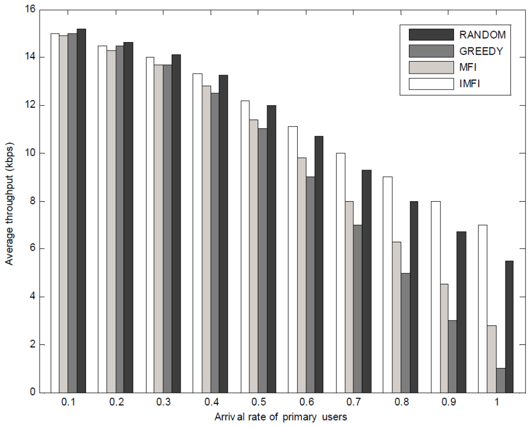


Figure 5: Average Throughput of CRNs

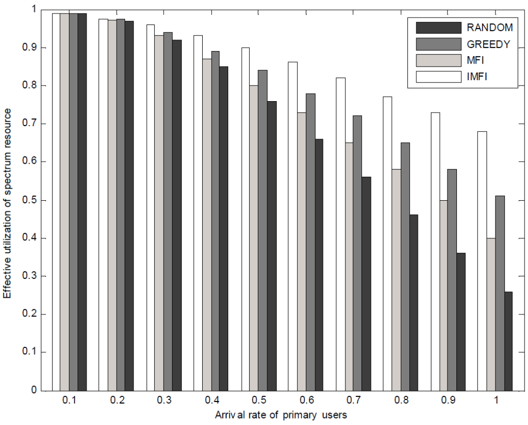


Figure 6: Effective Utilization of CRNs Spectrum Resource of SU

Figure 6 shows the effective utilization of CRNs spectrum resource. When λ is small, the four algorithms are closer to this parameter and the effective utilization for each is higher. With the increase of λ , service retransmission with RANDOM, GREEDY and MFI increases quickly, makes invalid data rapid upward, and cause the effective utilization of CRNs spectrum resource decline rapidly. IMFI can inhibit the service retransmission probability upward effectively, especially in the condition that the arrival rate of PUs is higher, the advantage is more obvious. When λ =1.0, the effective utilization of CRNs spectrum resource using IMFI still reaches to 70%, that makes more efficient use of spectrum sources, and improve system performance effectively.

5 Conclusions

The objective of this paper is to improve the QoS of SUs and CRNs performance. In view of the problems that existing methods exist, we put forward a spectrum migration approach based on pre-decision aid and interval Mamdani fuzzy inference in CRNs. Through the establishment and analysis of the spectrum migration model, we define spectrum migration factors (*SOP*, *LMP* and *SMD*) as spectrum characteristic metrics for spectrum migration decision. Moreover, pre-decision is put forward to reduce system complexity and improve migration efficiency. For shorten spectrum migration decision time and seek the optimal spectrum holes, we propose an interval Mamdani fuzzy inference method based on Mamdani fuzzy inference, which can reduce inference time significantly. At last, simulation results show the effectiveness of our approach compared with other existing algorithms.

Acknowledgments

This work was supported by the Fundamental Research Fund for the Central Universities (HEUCFZ1213), and Doctoral Fund of Ministry of Education of China (20122304130002).

Bibliography

[1] Yuan, G. et al; Performance Analysis of Selective Opportunistic Spectrum Access with Traffic Prediction, *IEEE Trans. Vehicular Technology*, 59(4): 1949-1959, 2010.

[2] Chang, N.B.; Liu, M.(2009); Optimal Channel Probing and Transmission Scheduling for Opportunistic Spectrum Access, *IEEE/ACM Trans. Networking*, 17(6): 1805-1818.

[3] Pawelczak, P. et al; Quality of Service Assessment of Opportunistic Spectrum Access: A Medium Access Control Approach, *IEEE Wireless Communications*, 15(5): 20-29, 2008.

[4] Zhou, X. et al; Probabilistic Resource Allocation for Opportunistic Spectrum Access, *IEEE Trans. Wireless Communications*, 9(9): 2870-2879, 2010.

[5] Huang, S. et al; Optimal Transmission Strategies for Dynamic Spectrum Access in Cognitive Radio Networks, *IEEE Trans. Mobile Computing*, 8(12): 1636-1648, 2009.

[6] Huang, S.; Liu, X.; Ding, Z.; Opportunistic Spectrum Access in Cognitive Radio Networks, *Proc. of INFOCOM*: 1427-1435, 2008.

[7] Wang, L.C.; Wang, C.W.; Chang, C.J.; Modeling and Analysis for Spectrum Handoffs in Cognitive Radio Networks, *IEEE Trans. Mobile Computing*, 11(9): 1499-1513, 2012.

[8] An, C.; Ji, H.; Si, P.; Dynamic Spectrum Access with QoS Provisioning in Cognitive Radio Networks, *Proc. of GLOBECOM*: 1-5, 2010.

[9] Hamdaoui, B.; Adaptive Spectrum Assessment for Opportunistic Access in Cognitive Radio Networks, *IEEE Trans. Wireless Communications,* 8(2): 922-930, 2009.

[10] Stotas, S.; Nallanathan, A.; On the Throughput and Spectrum Sensing Enhancement of Opportunistic Spectrum Access Cognitive Radio Networks, *IEEE Trans. Wireless Communications*, 11(1): 97-107, 2012.

INT J COMPUT COMMUN, ISSN 1841-9836
9(1):93-100, February, 2014.

A Practical Military Ontology Construction for the Intelligent Army Tactical Command Information System

D. Yoo, S. No, M. Ra

Donghee Yoo, Sungchun No, Minyoung Ra*
Department of Electronics Engineering & Information Science, Korea Military Academy
Gongneung-dong, Nowon-gu Seoul, Republic of Korea
donghee.info@gmail.com, is695@kma.ac.kr, myra@kma.ac.kr
*Corresponding author: myra@kma.ac.kr

Abstract: The purpose of this research is to construct a military ontology as the core element for implementing the intelligent Army Tactical Command Information System (ATCIS). Using the military ontology, the system can automatically understand and manage the meaning of military information in the system, and hence it can provide a commander with military knowledge for decision making. To construct the military ontology, we define the core concepts of the ontology based on terms extracted from the ATCIS database and complete the ontology by using the mixed ontology building methodology (MOBM). In addition, we implement intelligent ATCIS as a prototype that provides a military concept navigation service and commanders' decision support service to demonstrate how to use the military ontology in practice.

Keywords: military ontology, ontology building methodology, ATCIS.

1 Introduction

To realize network-centered warfare (NCW), Korea developed the Army Tactical Command Information System (ATCIS) as a ground tactics C4I system [1]. The main service domains of ATCIS are the (1) *'information'* domain for reporting battlefield situations such as the state, location, and movement of the enemy, (2) *'operation'* domain for decision making and operational orders based on the battlefield situation, and (3) *'firepower'* domain for analysis of the target and the order of priority for a strike. It is possible to effectively manage the battlefield because the ATCIS domains are organically linked. However, ATCIS provides only the actual facts of the battlefield, and thus the principal decision making, such as "the possibility of hostile provocation" and "the most effective strike method" depends on the intuition and experience of the commanders and staff officers. Commanders and staff officers can make faster and more accurate decisions in urgent battlefield conditions if they have access to specialized military knowledge for battlefield management.

For this reason, battlefield information must be expressed in machine-understandable language using a standardized format; ultimately, the military ontology, which defines various concepts and the meaning of their relationships (semantics), should be built based on battlefield information. In the case of defining rules for developing military knowledge by using the concepts in the military ontology, commanders and staff officers can obtain refined knowledge to help in the decision-making process.

Therefore, this paper has suggested the construction of a military ontology as the first step in implementing intelligent ATCIS. The mixed ontology building methodology (MOBM) [2] was applied for ontology construction, and the kernel ontology was defined based on terms extracted from the ATCIS database. Then the bottom-up approach and the top-down approach were applied to extend the kernel ontology, and finally the military ontology was completed. Also, this paper has described intelligent ATCIS as a prototype that provides military concept navigation service and commanders' decision support service to demonstrate how to use the military ontology in practice. Here, we suggested a method to minimize the time cost when inferring the military rules for decision making by using the query rewriting model.

The rest of the paper is organized as follows. Section 2 provides a review of previous methods used for ontology construction and introduces the outline of the MOBM. Section 3 presents the military ontology constructed by the MOBM and analyzes additional aspects of applying the MOBM to the ATCIS database information. Section 4 describes the intelligent ATCIS, which provides a military concept navigation service and a commanders' decision support service. Finally, in section 5, we draw conclusions and suggest directions for future research.

2 Related works

2.1 Previous ontology building methodology

A significant amount of research has been conducted on the issue of ontology building methodology. The research has employed essentially two approaches. The first collects terminology and builds the ontology by analyzing concepts, forming a hierarchy for the concepts, and defining the relationships between the concepts and the rules for acquiring domain knowledge. Based on the refinement process assigned to this task, the ontology is then completed. Several methods have been reported for accomplishing this task. The bottom-up method starts with the most specific classes and then groups them into more general concepts [3, 4]. The top-down method starts with the definition of the most general concepts and then divides these into detailed sub-concepts [5]. The middle-out method starts with certain middle-level concepts and then applies the bottom-up method or the top-down method as appropriate [6]. The hybrid method merges ontologies developed from the bottom-up method and top-down method into one ontology [7].

The second approach to ontology building involves developing an ontology from database schemas. This work takes three directions: (1) First, extract the entity-relationship (ER) model from the database schema using reengineering, then from that model extract the ontology [8]; (2) given the database schema and ontology, for semantic web applications, extract the mapping rules between them [9]; and (3) generate the ontology structure itself from the relational database schema [10].

The MOBM, a mixed methodology, was proposed based on these works [2]. The MOBM first generates a kernel ontology, which becomes the core, using database information as much as possible and then completes the ontology by applying the bottom-up method and the top-down method to build additional parts of the ontology.

2.2 Mixed ontology building methodology

As mentioned earlier, the MOBM combines the characteristics of both approaches to more effectively represent organizational knowledge on ontology. In the MOBM, mapping rules are defined to extract the main concepts and relationships of a certain domain ontology from the target database schema. This kind of domain ontology is called kernel ontology. Kernel ontology is enhanced by adding upper-level terms and lower-level terms, which are collected from domain knowledge or instances of the target database, because they may contain new concepts or relationships that did not exist in the target database schema. Based on the top-down method, the upper-level terms are conceptualized into upper concepts. In the same way, the lower-level terms are conceptualized into lower concepts using the bottom-up method. Once the upper and lower concepts are developed, they are linked to the kernel ontology. Thus, the MOBM employs eight steps to build domain ontologies, as follows:

- Step 1: Extracting kernel ontology from database schema.

- Step 2: Developing class hierarchies from upper concepts.
- Step 3: Developing class hierarchies from lower concepts.
- Step 4: Connecting these class hierarchies into kernel ontology.
- Step 5: Enhancing the semantics between inter-terms.
- Step 6: Enhancing any restrictions.
- Step 7: Enhancing additional axioms and rules.
- Step 8: Completing the ontology.

3 Practical military ontology construction

3.1 Building scope of military ontology

We adopted the MOBM as an ontology building methodology. In this section, we describe the process of constructing the military ontology from ATCIS according to the MOBM and provide an analysis of what should be considered when the MOBM is applied to ATCIS domains. A military ontology for the core service domains of ATCIS – *information*, *operation*, and *firepower* – has been constructed among various service domains. Currently, we have collected 10,835 related terms from the ATICS database schema and the defense technology information service (DTiMS) thesaurus and electronic drill book, and 6,515 refined terms have been used for the military ontology construction.

3.2 Extraction of kernel ontology

The heart of the MOBM is the utilization of database schema to construct a practical ontology. First, the kernel ontology was built following the mapping rules [2] of the MOBM after extracting the core terms from the database schema. Because the scope of the ATCIS database schema was overly broad, the kernel ontology was created distinguishing the *information*, *operation*, and *firepower* domains. This paper focused only on the *information* domain, and the middle part of Figure 1 shows the hierarchy structure of the concepts in the domain kernel ontology. The MOBM mentions that the hierarchy structure of the concepts is well presented if the hierarchy of the tables in the relational database is properly defined. However, the hierarchy structure of the concepts is not well represented due to scarcity of hierarchies of the tables in the relational database. Therefore, other upper and lower concepts were added to the kernel ontology.

3.3 Creating class hierarchies from upper concepts

Second, the upper concepts of the kernel ontology were conceptualized as a form of class hierarchies. The terms for the upper concepts were mostly collected from the DTiMS thesaurus and electronic drill book and conceptualized into the upper concepts of the kernel ontology by using the top-down approach. Some of the concepts in the kernel ontology did not connect to the upper concepts because they were not conceptualized in the previous step. This step was intended to alleviate this problem by utilizing the concepts defined by domain experts. The upper part of Figure 1 shows the conceptualized upper concepts, and the underlined concepts (e.g., *Tactical_ Operating_ Spot*) are the ones defined by domain experts.

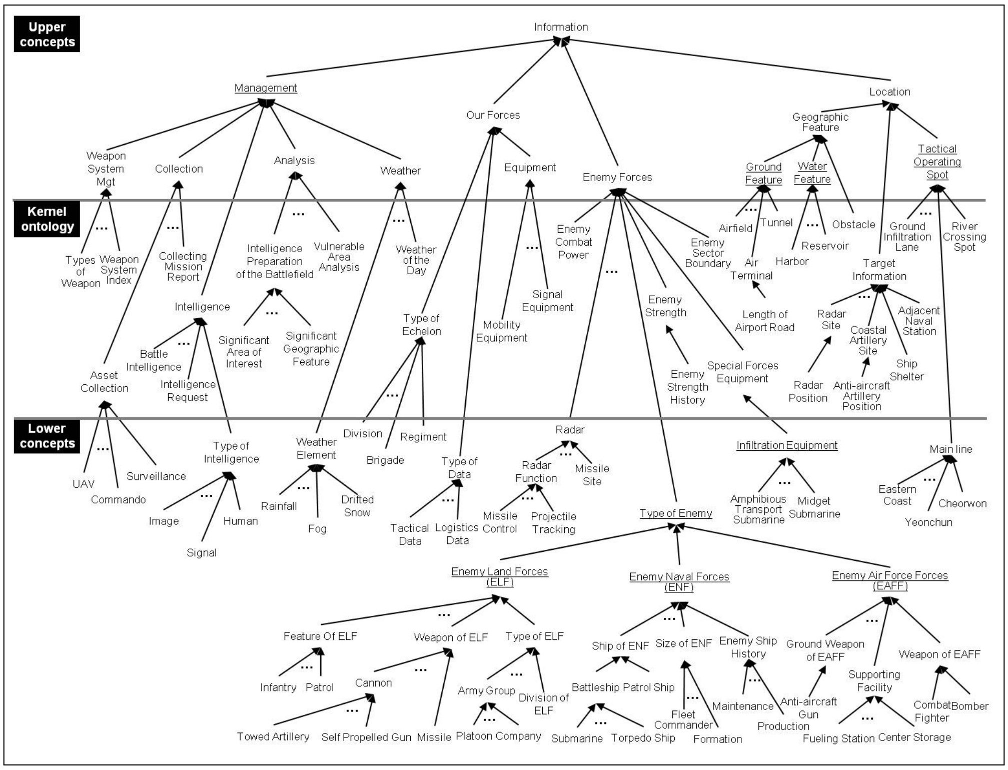


Figure 1: Example of class hierarchy in information domain

3.4 Creating class hierarchies from lower concepts

The third step was to specify the lower concepts of the kernel ontology. The additional information on database instances was utilized in addition to the terms referred to in the previous steps. The instances used in ATCIS can be divided into two types, instances for system implementation and instances for either real-time training or a real situation, and this step focused on specifying the lower concepts based on the instances for system implementation among the two types. As shown in the lower part of Figure 1, the collected terms were conceptualized by following the bottom-up approach, and the intermediate concepts were additionally defined to link the instances with the concepts in the kernel ontology through the MOBM. Here, the code information of the primary key in the ATCIS database directory was used for the intermediate terms, and they were closely related to the hierarchy information by function provided by ATCIS. For example, a user can select *Amphibious Transport Submarine* or *Midget Submarine* as *Infiltration Equipment* among *Special Forces Equipment* from the ATCIS system screen. When referring to this kind of information, *Infiltration_ Equipment* can be used as the intermediate term that links *Special_ Forces_ Equipment* in the kernel ontology to the instances *Amphibious_ Transport_ Submarine* and *Midget_ Submarine.* The fourth step completed the hierarchy of the core concepts in the military ontology by connecting the upper and lower concepts into the kernel ontology, as shown in Figure 1.

3.5 Enhancing the semantics and completing the ontology

The meaning of the military ontology was enhanced through the fifth, sixth, and seventh steps of the MOBM. Following the methods proposed in each step, the *equivalentClass* relationship, *disjointWith* relationship, and *intersectionOf* relationship were defined in addition to the *subClassOf* relationship to reinforce the hierarchy among the concepts. Moreover, the restrictions and axioms were additionally defined, and the domain rules were also proposed after collecting the military knowledge used in commanders' decision making. The final step implemented the enhanced military ontology in Web Ontology Language (OWL) [11] by using Protégé, and the military knowledge was defined as the form of rule by using Semantic Web Rule Language (SWRL) [12, 13].

4 Intelligent ATCIS

To illustrate how military ontology can be used to discover relevant military concepts or provide military knowledge to support commanders' decision making, we implemented intelligent ATCIS as a prototype system. In this section, we describe two services of the system, the *military concept navigation service* and the *commanders' decision support service.*

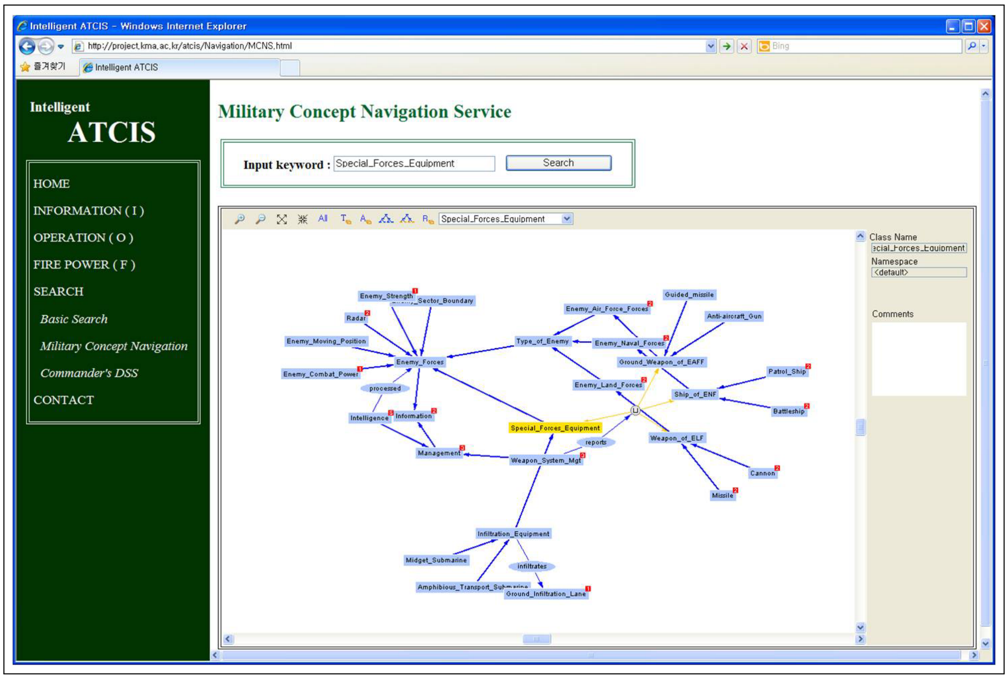
4.1 Military concept navigation service

To understand the relationship among the various military terms used in intelligent ATCIS, users can employ a Web-based military concept navigation service. As shown in Figure 2, users can enter a keyword (e.g., *Special_Forces_Equipment*) as a query in the upper part of the screen and then click the 'Search' button to start the military concept navigation. The result of the query is displayed in the center of the screen; the classes and individuals are represented as boxes and their corresponding properties are represented as circles connecting the boxes. The starting point for the navigation is the yellow box, which presents the entered query. Users can interactively navigate the underlying military concepts by moving from one box to another. To explore other classes or individuals, users right-click the menu on a selected box (e.g., the *show more* or *show less* function). This navigation process may be repeated until users find the military concepts they seek. This service was implemented in Java Server Pages (JSP) on an Apache Tomcat server and the visualization function related to concept navigation was developed using the GrOWL [14].

4.2 Commanders' decision support service

Commanders and staff officers require decision support services to effectively manage battlefields that are in flux. In particular, in the case of an emergency such as the occurrence of war, the system must provide relevant military knowledge that supports decision making that takes place in a short time. However, existing reasoning engines did not have sufficient inference capabilities for a significant amount of military knowledge. As military knowledge that consists of forms of SWRL rules increases, more reasoning time is necessary to create relevant military knowledge from the military ontology and various military data.

With the intelligent ATCIS, we have developed a commanders' decision support service based on the query rewriting method [15] to resolve the problem. Here, we describe the core function of the service, which is conducted via rewritten SPARQL queries instead of the reasoning engine to create relevant military knowledge in a short time. Figure 3 shows the overall process of the service. For example, when a commander enters queries that include military knowledge, the queries are translated into SPARQL queries. Then, the query rewriting engine divides the SPARQL queries into two parts: the military domain query (MDQ) and the military knowledge-related query (MKQ). In the system, the military knowledge as defined by SWRL is programmed as rule templates. Because the meaning of MKQ depends on military knowledge, the engine refers to military rule (MR) in the rule templates to understand the meaning of the military knowledge. Based on the rules, the triple pattern rewriter can change MKQ into Extended SPARQL Query



Input keyword : Special_Forces_Equipment

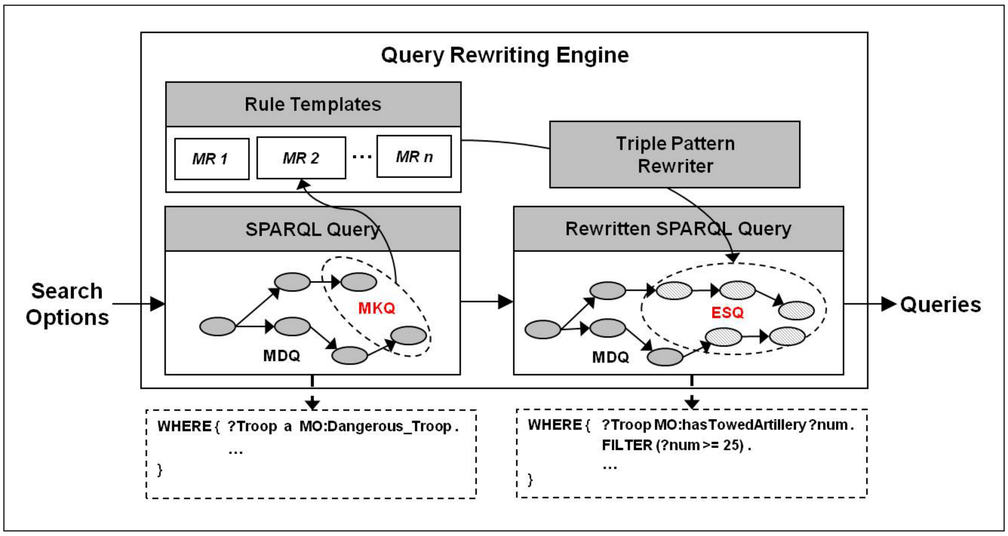
Special_Forces_Equipment

Class Name
ecial_Forces_Equipment

Namespace
<default>

Figure 2: Military concept navigation service

(ESQ). For example, Figure 3 shows that part of MKQ (e.g., *?Troop a MO:Dangerous_ Troop.*) changes in ESQ (e.g., *?Troop MO:hasTowedArtillery ?num . FILTER (?num >= 25).*) according to MR 2, which includes military knowledge (e.g., a dangerous troop is one with more than 25 towed artilleries). Then, the query rewriting engine submits a rewritten SPARQL query to the triple knowledge base and retrieves relevant information from the knowledge base.



```
WHERE { ?Troop a MO:Dangerous_Troop .
        …
}
```

```
WHERE { ?Troop MO:hasTowedArtillery ?num .
        FILTER (?num >= 25) .
        …
}
```

Figure 3: Overall process of the query rewriting method for decision supporting service

5 Conclusion

This paper presented the process of military ontology construction through the MOBM, and considerable additional aspects in each step were analyzed. Using the MOBM, the core concepts of the military ontology were quickly composed from the practical terms in ATCIS, and one can perceive that the more ISA relationships exist in the database, the more effective the kernel ontology constructed. Also, this paper described the implementation of intelligent ATCIS, which provides a military concept navigation service and a commanders' decision support service to show how to use the military ontology. In particular, we proposed a method for minimizing the time cost when inferring the military knowledge through the use of the query rewriting model. The results of this study may be used as the basic material for constructing a more practical military ontology in various military domains.

Finally, future work will be continued in the direction of building the individual ontology, not only for the *information* domain but also for the *operation* and *firepower* domains, and constructing an integrated military ontology by combining the domains.

Acknowledgement

This research was supported by Agency for Defense Development (ADD) under the contract number UD110058MD.

Bibliography

[1] Republic of Korea Army, http://www.army.mil.kr/english/sub02/sub02_02_03.jsp.

[2] M. Ra, D. Yoo, S. No, J. Shin, C. Han, The Mixed Ontology Building Methodology Using Database Information, *IAENG Int. Conference on Artificial Intelligence and Applications (ICAIA'12)*, Hong Kong, 14-16 March 2012.

[3] M. Grüninger, M. S. Fox, Methodology for the design and evaluation of ontologies, *Proc. of the Workshop on Basic Ontological Issues in Knowledge Sharing held in conjunction with IJCAI-95*, Montreal, Canada, 1995.

[4] P. E. van der Vet, N. J. I. Mars, Bottom-Up Construction of Ontologies, *IEEE Transactions on Knowledge and Data Engineering*, 10(4):513-526, 1998.

[5] G. Schreiber, B. Wielinga, W. Jansweijer, The KACTUS View on the 'O' Word, *IJCAI Workshop on Basic Ontological Issues in Knowledge Sharing, Montreal*, Canada, 159-168, 1995.

[6] O. Corcho, M. Fernández-López, A. Gómez-Pérez, A. López-Cima, Building legal ontologies with METHONTOLOGY and WebODE, *Proc. of Law and the Semantic Web*, 142-157, 2003.

[7] F. J. Lopez-Pellicer, L. M. Vilches-Blázquez, J. Nogueras-Iso, O. Corcho, M. A. Bernabé, A. F. Rodriguez, Using a hybrid approach for the development of an ontology in the hydrographical domain, *Proceedings of 2nd Workshop on Ontologies for Urban Development: Conceptual Models for Practitioners*, 2007.

[8] J. Trinkunas, O. Vasilecas, Building ontologies from relatinoal databases using reverse engineering methods, *Proc. of Int. Conference on Computer Systems and Technologies*, 2007.

[9] N. Konstantinou, D. E. Spanos, N. Mitrou, Ontology and Database Mapping: A Survey of Current Implementations and Future Directions, *Journal of Web Engineering*, 7(1):1-24, 2008.

[10] S. S. Sane, A. Shirke, Generating OWL ontologies from a relational databases for the semantic web, *Proc. of Int. Conference on Advances in Computing, Communication and Control*, 143-148, 2009.

[11] D. L. McGuinness, F. V. Harmelen, OWL Web Ontology Language Overview, W3C Recommendation 10 February 2004, http://www.w3.org/TR/owl-features/.

[12] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, M. Dean, SWRL: A Semantic Web Rule Language Combining OWL and RuleML, W3C Member Submission 21 May 2004, http://www.w3.org/Submission/SWRL/.

[13] C. M. Bogdan, Domain Ontology of the VirDenT System, *Int J Comput Commun*, 6(1):45-52, 2011.

[14] S. Krivov, R. Williams, F. Villa, GrOWL: A Tool for Visualization and Editing of OWL Ontologies, *Journal of Web Semantics*, 5(2):54-57, 2007.

[15] D. Yoo, Hybrid Query Processing for Personalized Information Retrieval on the Semantic Web, *Knowledge-Based Systems*, 27:211-218, 2012.

INT J COMPUT COMMUN, ISSN 1841-9836
9(1):101-112, February, 2014.

Network Anomaly Detection based on Multi-scale Dynamic Characteristics of Traffic

J. Yuan, R. Yuan, X. Chen

Jing Yuan*, Ruixi Yuan, Xi Chen
Department of Automation, Tsinghua University
Beijing, China 100084
*Corresponding author: yuan-j05@mails.tsinghua.edu.cn
ryuan@mail.tsinghua.edu.cn, bjchenxi@mail.tsinghua.edu.cn

Abstract: This paper proposes a novel detection engine, called the Wavelet-Recurrence-Clustering (WRC) detection model, to study the network anomaly detection problem that is widely attractive in Internet security area. The WRC model first applies the wavelet transform and recurrence analysis to calculate the multi-scale dynamic characteristics of network traffic, and then identifies network anomalies through the clustering algorithm with those dynamic characteristics. The evaluation results on DARPA 1999 dataset indicate that the WRC detection model can effectively improve the detection accuracy with a low false alarm rate.

Keywords: network anomaly detection, multi-scale dynamic characteristics, recurrence analysis, WRC detection model.

1 Introduction

Currently, Internet suffers from a large number of different hacker attacks and threats frequently. In most cases, the hacking attacks cause anomalous traffic behaviors. Thus, how to accurately detect network anomalies by analyzing traffic behaviors becomes attractive in recent years. The basic idea of anomaly detection approaches is building the behavior profile of normal traffic and then identifying the observation of traffic as anomalous when it deviates from the normal profile. These methods can be basically categorized into the following two aspects. The first common approach to detection network anomalies is to establish normal traffic profiles with time series analysis, and then to identify the statistical observation of traffic as normal or anomaly, based on the deviation between the observation and the prediction value. Some traditional time series analysis for detecting anomalies, such as exponential smoothing and the auto-regressive process, were utilized in [1], [2]. These time series methods all require traffic series to be stationary. However, many prior studies indicated that traffic series are actually non-stationary and showed some nonlinear dynamic characteristics, such as self-similar, long range dependence and recurrence [3], [4]. Thus, statistical detection approaches by the use of traditional time series analysis may not be effective any more.

The other conventional approach for detecting network anomalies is built upon the machine learning theory. Based on statistical observations of traffic, such as the average packet size, the flow duration and the flow size, we can design and train effective classifiers to identify malicious traffic behaviors [5]- [7]. While these approaches do have a fast detection speed, they still have some problems, so that they are unreliable and may have high false alarm rates. For example, traffic statistical behaviors of the training dataset could be dissimilar to that of the testing dataset [8]. Also, traffic statistical observations have certain randomness and may vary along with network scales and application environments.

As the application of the two types of approaches above are both limited to some special scenarios, how to build new detection models especially with novel analysis tools is to effectively detect network anomalies with high detection accuracy and low false alarm rate, is definitely of interest to researchers and engineers working in this area. Before trying to propose a new detection model, let us look into the network traffic again to see if there is any new essential element that can be a hammer onto the nail. As a matter of fact, there are two important properties of the network traffic – the multi-scale property and the recurrence property. Since our work is motivated by these two properties, let us have a brief review on them.

1) Multi-scale property: the network traffic has different statistical behaviors at different time scales, i.e. during long time intervals observations show stable and periodic changes, while during short time intervals they are random and fluctuate sharply. Moreover, the prior study in [9] also indicated that different types of network attacks exist in different time scales. Thus, the multi-scale property of the network traffic is essentially important and should raise attention.

2) Recurrence property: the network traffic system is a complex dynamic system that exhibits some nonlinear and intrinsic features, such as recurrence. Recurrence is a fundamental property of dynamic systems, which indicates intrinsic evolution regularities of traffic states. Compared with statistical observations of traffic, recurrence patterns of traffic states are inherent and uninfluenced by network scales and application environments.

In this paper, we propose a novel model, called the Wavelet-Recurrence-Clustering (WRC) detection model, to detect network anomalies. To address the first property, we apply the wavelet transform to analyze the network traffic at different time scales respectively, to build the accurate traffic profile and highlight local and random variations of traffic behaviors. To address the second property, we calculate the dynamic characteristics of the network traffic at different frequencies, to reveal non-stationary transition patterns caused by anomalous events, based on the recurrence analysis. Our WRC detection model incorporates both these two properties.

The main contributions of this paper are summarized as follows. (1) a nonlinear analysis method to calculate the multi-scale dynamic characteristics of network traffic by employing the wavelet transform and the recurrence analysis; (2) a detection model for identifying network anomalies based on dynamic characteristics; (3) evaluations and comparisons of the WRC detection model with traditional detection methods on DARPA 1999 dataset.

The rest of this paper is organized as follows. Section 2 overviews the WRC detection model. In Section 3, the implementation of the WRC model is illustrated in detail. Section 4 shows experiments to evaluate the performance of the WRC model. Section 5 concludes our work.

2 WRC detection model

As shown in Fig. 1, our WRC detection model consists of two components, i.e., the dynamic characteristic extraction module and the anomaly identification module. In this section, we illustrate each module in detail.

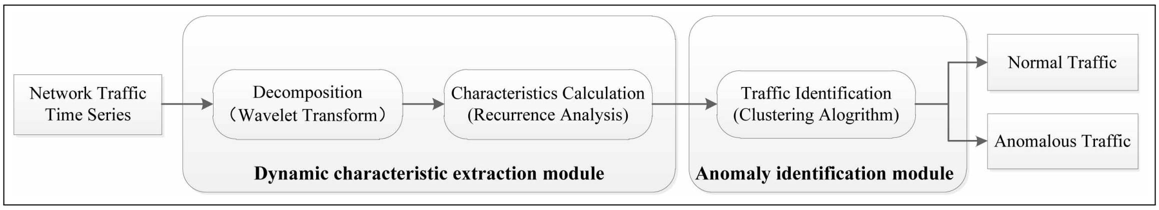

Network Traffic Time Series
Decomposition （Wavelet Transform）
Characteristics Calculation (Recurrence Analysis)
Dynamic characteristic extraction module
Traffic Identification (Clustering Alogrithm)
Anomaly identification module
Normal Traffic
Anomalous Traffic

Figure 1: The Framework of WRC detection model

2.1 Dynamic characteristic extraction module

Based on the recurrence property of network traffic states, we propose a multi-scale recurrence characteristic extraction method. First, the wavelet transform is used to decompose and reconstruct the traffic at different frequencies, and then the recurrence analysis is employed to calculate the dynamic characteristics of traffic, which can effectively highlight the non-stationary transition patterns caused by malicious events and improve the performance of network anomaly detection.

A. Multi-scale analysis

Wavelet transform [10] is a multi-scale analysis method. It has a good time-frequency resolution, i.e., gets the good frequency resolution at low frequencies and gets the good time resolution at high frequencies, which can help us to capture the traffic behaviors at different time scales. In this paper, we adopt the discrete wavelet transform (DWT) to reconstruct the traffic at different frequencies. DWT is a multi-stage algorithm that decomposes traffic time series into a coarse approximation and a series of detail information, which are used to reconstruct the traffic, by employing a scaling function (low pass filters, LP) and a wavelet function (high pass filters, HP).

B. Recurrence analysis

Network traffic system is a dynamic system. The trajectory of traffic state shows recurrence phenomena in the phase space. Recurrence is a fundamental property of dynamic systems, which indicates the evolution regularities of the state trajectory, i.e., after a period time, the system state is identical or similar to the former states and the evolution patterns are repeating. Thus, based on this inherent property, our WRC detection model calculates the dynamic characteristics of network traffic to accurately describe traffic behaviors and reveal the non-stationary transition patterns.

1) Recurrence plot

In order to intuitively explore the recurrence phenomena, WRC detection model employs recurrence plot (RP) [11] to visualize the recurrence property of traffic states in the high-dimensional phase space into a two-dimensional plane.

Given a network traffic time series $x = \{x_i\}, i = 1, 2, ..., n$, the traffic system state can be expressed as follows:

$$\boldsymbol{X}_j = [x_j, x_{j+\tau}, ..., x_{j+(m+1)\tau}] \quad j = 1, 2, ..., N \tag{1}$$

where m is the embedding dimension and τ is the time delay, $N = n-(m-1)\tau$. After obtaining traffic system states, we use RP to investigate the recurrence phenomena of traffic states. The mathematical expression of RP is shown as follows:

$$R_{i,j} = \Theta(\varepsilon - \|\boldsymbol{X}_i - \boldsymbol{X}_j\|) \quad j = 1, 2, ..., N \tag{2}$$

where $R_{i,j}$ is an element of the recurrence matrix, ε is the threshold, $\boldsymbol{X}_i$ is a system state in the m-dimensional phase space, $\|\cdot\|$ is a norm, N is the number of states, $\Theta(\cdot)$ is the Heaviside function defined as :

$$\Theta(y) = \begin{cases} 0 & y \leqslant 0 \\ 1 & y > 0 \end{cases} \tag{3}$$

RP gives an intuitive description of the recurrence phenomena of traffic states in the phase space. If the distance between the states $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ is smaller than ε, then the value of $R_{i,j}$ is 1 and there is a black dot at (i, j) in the RP; otherwise, the value of $R_{i,j}$ is 0 and there is a white dot at (i, j). Fig. 2 gives an example of RPs for the normal and anomalous traffic series.

Compared with the normal traffic, the RP of the anomalous traffic exists wide white bands obviously (non-stationary transition of states), which indicates some malicious events happened.

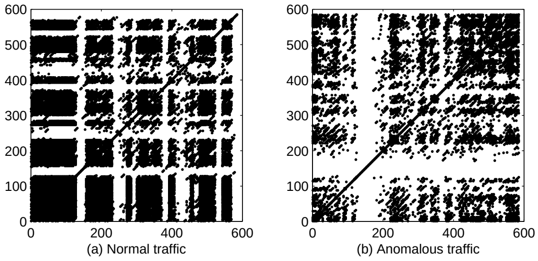


Figure 2: The RPs of normal and anomalous traffic series

2) Recurrence quantification analysis

Recurrence quantification analysis (RQA) [12] is employed by our WRC detection model to quantify the structure textures shown in RP to reveal the evolution patterns of traffic states. The proper quantification variables studied are as follows.

(1) Recurrence Ratio (RR) is the percentage of recurrence points in the phase space, which measures the density of recurrence points in RP. If dynamic systems are periodic, the value of RR will be high.

$$RR = \frac{1}{N^2} \sum_{i,j=1}^{N} R_{i,j} \tag{4}$$

where $R_{i,j}$ is the recurrence point calculated by formula (2).

(2) Determinism (DET) is the ratio of recurrence points that form diagonal line structures in RP to all recurrence points. It measures the determinism of traffic systems, i.e., it can tell us how deterministic and predictable the system is. If the traffic system is periodical, the value of DET will be high. That is because the period system states can form long diagonal lines.

$$DET = \frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{i,j=1}^{N} R_{i,j}} \tag{5}$$

where $P(l)$ is the frequency distribution of diagonal line with length l. l_{min} is the minimal length of diagonal lines.

(3) Entropy (ENT) is the Shannon entropy of the probability of diagonal line lengths. It measures the complexity of traffic systems.

$$ENT = -\sum_{l=l_{min}}^{N} p(l) \log_2 p(l) \tag{6}$$

where $p(l) = \frac{P(l)}{\sum_{l=l_{min}}^{N} P(l)}$ is the probability of frequency distribution of diagonal line l .

The dynamic characteristic extraction module calculates the above three quantification variables of network traffic at different frequencies, in order to form the feature vectors to reveal the inherent traffic behaviors and discover the non-stationary transition caused by anomalous events.

2.2 Anomaly identification module

In our WRC detection model, a clustering algorithm, i.e., k-means [13] is employed to identify normal and anomalous traffic based on the feature vectors obtained from the dynamic characteristic extraction module. Detailed steps of k-means method for identifying anomalies are as follows.

Step 1: randomly selects k instances from the traffic training dataset to represent the centroids of k clusters $c_1, c_2, ..., c_k$;

Step 2: for each instance x in training dataset, calculates its distance to the centroids of all k clusters, $d(c_i, x), i = 1, 2, ..., k$. If the value of $d(c_m, x)$ is smallest, then assigns x into c_m;

Step 3: when all instances in the training dataset are assigned, recalculates the centroids of all k clusters;

Step 4: repeats step 2 and step 3 until the centroids of all k clusters no longer change;

Step 5: for each instance y in the testing dataset, first calculates its distance to the centroids of all k clusters, $d(c_i, y), i = 1, 2, ..., k$, and then finds the cluster c_n with the closest distance;

Step 6: according to the following threshold rule, identifies y as a normal traffic or an anomalous traffic.

$$\begin{cases} y = 1 & ifP(\omega_{1n}|y \in c_n) > th \\ y = 0 & ifP(\omega_{1n}|y \in c_n) \leq th \end{cases} \tag{7}$$

where, “1” and “0” represent the types of anomalous and normal traffic respectively. ω_{1n} is the anomalous traffic in cluster c_n, $P(\omega_{1n}|y \in c_n)$ is the probability of the anomalous traffic instances in c_n. th is a threshold and its value is 0.5, which means if and only if the majority of the cluster c_n are anomalous traffic, then y is identified as an anomaly.

3 Implementation of WRC detection model

In this section, we first preprocess DARPA 1999 traffic traces to obtain the statistical traffic time series as the input of WRC detection model, and then describe the implementation process of WRC in detail.

3.1 Data preprocessing

A. DARPA 1999 dataset

DARPA 1999 intrusion detection dataset [14] has been widely used for evaluating the performance of intrusion detection systems. This dataset includes five weeks tcpdump packet traces collected from two sniffers: “inside sniffer” between the gateway and the simulated air force network, and “outside sniffer” between the gateway and the simulated Internet. Among the five weeks, traffic of the first and third weeks are anomaly-free, while traffic of the rest include anomalous traffic. For each week, packet traces are captured from Monday to Friday, and for each day, they are collected from 8:00am to 6:00am in the next day. In addition, we find the traffic volume of each day is very small after 6:00pm, thus we analyze the traffic traces only during workday from 8:00am to 6:00pm (10 hours).

DARPA 1999 dataset includes five types of attacks, i.e., Denial of Service Attacks (DoS), User to Root Attacks (U2R), Remote to Local Attacks (R2L), Probes and Data. U2R and Data only exploit computer system vulnerabilities and do not have bad impact on the performance of network. Their traffic behaviors cannot be differed from the normal traffic, i.e., they are stealthy for network. Thus, in our experiments, we do not consider these two types of attacks.

B. Network traffic series acquistion

We preprocess DARPA 1999 packet traces and obtain the flow-based statistical observations to form different traffic time series to be analyzed in the WRC detection model. First, the tcpdump packet traces are converted into flow based on five tuples (source IP, destination IP, source Port, destination Port, transportation protocol), and then five flow-based statistics listed in Table 1 are selected to form the traffic time series (the time interval is one minute).

Table 1: Flow-based statistics of network traffic

	Statistic	Description
S_1	flownum	The number of flow per minute
S_2	avepktnum	The average number of packet per flow
S_3	avebyte	The average byte per flow (average flow size)
S_4	avepktsize	The average byte per packet (average packet size)
S_5	ratio	The ratio of flownum to avepktsize

In Table 1, the first four statistics are directly calculated from the flow traffic and give us a detailed picture about the volume of the traffic. The fifth statistics is the ratio of the number of flow to the average packet size, which measures the interaction communication behaviors of traffic. Fig. 3 gives an example of the normal and anomalous traffic time series based on the S_2 statistic. In this paper, the anomalous traffic time series means the traffic series that contains anomalies. From Fig. 3, we can see that the statistic of avepktnum cannot distinguish anomalous traffic from normal traffic. Thus, our WRC detection model calculates the dynamic characteristics of network traffic, which are more sensitive to the small-scale variations and transitions of traffic than statistical features.

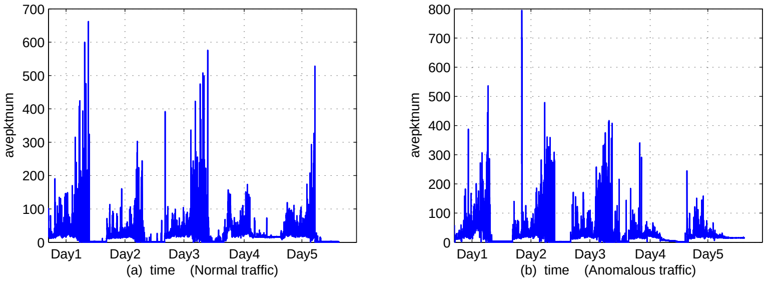


Figure 3: The normal and anomalous traffic time series (S_2: avepktnum)

3.2 Feature vectors extraction for clustering identification

Once the traffic time series have been obtained, WRC detection model analyzes these traffic series to extract the multi-scale dynamic characteristics to compose the feature vectors of traffic, which will be used to identify network anomalies.

A. Multi-scale analysis for traffic series

Given a network traffic time series $x = \{x_i\}, i = 1, 2, ..., n$, we use the daubechies wavelet to decompose and reconstruct the traffic signals at different frequencies, i.e., low-frequency and high-frequency, which are expressed as below.

$$l = \{l_i\} \quad i = 1, 2, ..., n \tag{8}$$

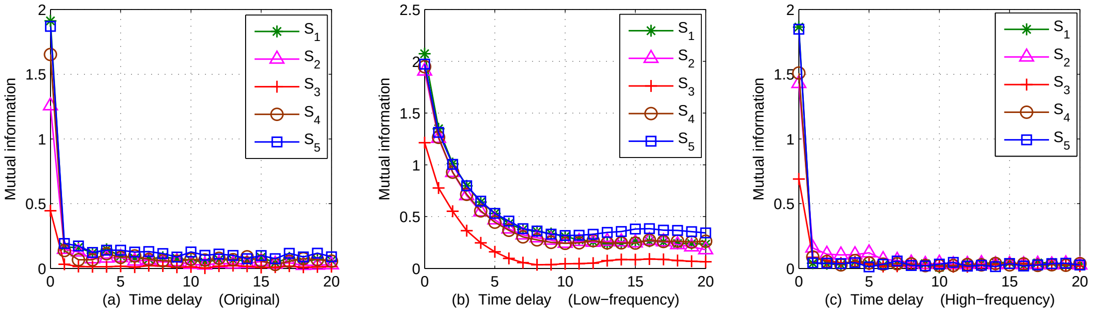
$$h = \{h_i\} \quad i = 1, 2, ..., n \tag{9}$$

B. Recurrence dynamic characteristics extraction

After obtaining the traffic series at different frequencies, we employ the recurrence quantification analysis (RQA) method to calculate the dynamic characteristics.

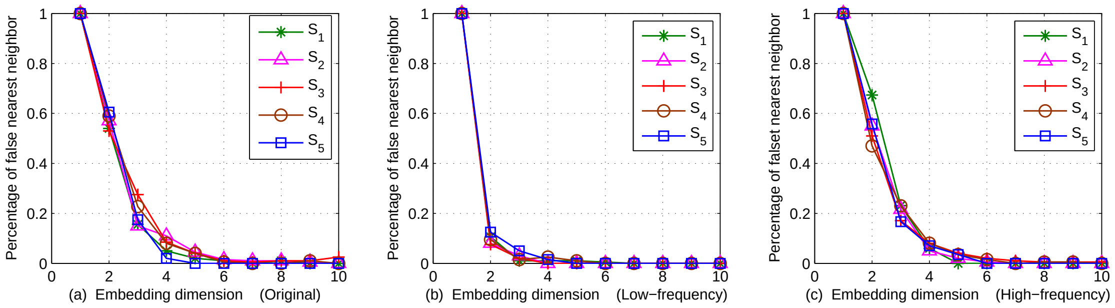
1)Parameter selection for RQA

Time delay τ, embedding dimension m and threshold distance ε are three fundamental parameters for RQA to extract the recurrence characteristics accurately. In this paper, we use the mutual information method to select the proper time delay τ and adopt the false nearest neighbors method to determine the correct embedding dimension m [15]. Fig. 4 and Fig. 5 show the results of the mutual information and the percentage of false nearest neighbors for traffic series at different frequencies based on the five statistics, respectively.



(a) Time delay (Original) (b) Time delay (Low−frequency) (c) Time delay (High−frequency)

Figure 4: Time delay of network traffic series



(a) Embedding dimension (Original) (b) Embedding dimension (Low−frequency) (c) Embedding dimension (High−frequency)

Figure 5: Embedding dimension of network traffic series

From Fig. 4, we can find that for the original traffic and high-frequency traffic, the mutual information get the first minimum value at time unit 3, while for the low-frequency traffic, the first minimum value is obtained around time unit 10. From Fig. 5, we can see that the percentage of false nearest neighbors is down to zero after $m = 5$. The previous study indicated that the embedding parameters (including τ and m) has less influence on recurrence analysis [16]. Therefore, we chose $\tau = 3, m = 5$ for recurrence quantification analysis without loss of generality. In addition, based on the rule of thumb, we set the threshold distance to 10% of the maximum diameter of phase space [17].

2) Dynamic characteristics calculated using RQA

Based on the traffic time series at different frequencies and the determined parameters, we employ RQA to calculate the multi-scale dynamic characteristics of network traffic so as to reveal the inherent traffic behaviors and discover the non-stationary transition patterns caused by malicious attacks.

In order to identify traffic anomalies in time, WRC detection model employs RQA method within a sliding window to analyze traffic series. First, the whole traffic time series is divided into several subseries by a sliding window, and then the recurrence characteristics (RR, DET and ENT) introduced in section 2.1.2 are calculated for these subseries. If the size and the shift of the sliding window are W and W_s, then the start time and the end time of subseries r are $t_{start} = (r-1)W_s + 1$ and $t_{end} = (r-1)W_s + W$. In our experiment, we set W to thirty minutes and W_s to six minutes.

For each subseries r, after calculating its recurrence characteristics (RR, DET and ENT), we can use a feature vector $[f_{RR}, f_{DET}, f_{ENT}]$ to describe the traffic behaviors in it, which will be used to identify traffic anomalies by k-means algorithm. Fig. 6 shows the recurrence characteristics of the normal and anomalous traffic time series based on the S_1 statistic.

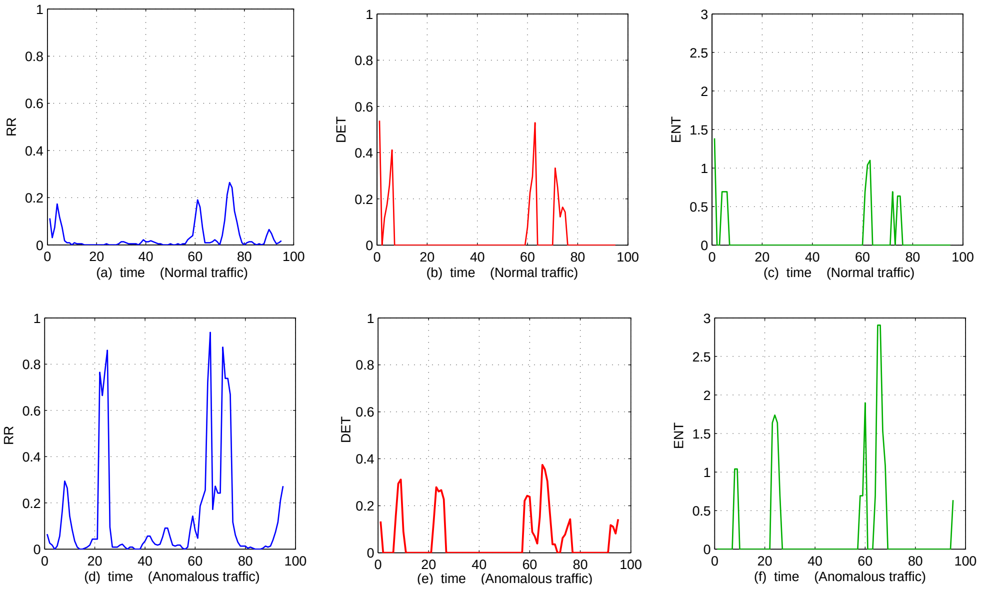


Figure 6: The dynamics characteristics of the normal and anomalous traffic series

As shown in Fig. 6, the dynamic characteristics of the anomalous traffic series are distinguished from that of the normal traffic series. In Fig. 6-(a) and Fig. 6-(d), the value of RR significantly increases when malicious events happened in the traffic series. Thus, RR is an effective feature for detecting anomalies. From Fig. 6-(c) and Fig. 6-(f), we find ENT is another good discriminator to identify anomalous attacks. The entropy measures the uncertainty of systems. The larger the ENT is, the more uncertain the system is. In normal traffic series the daily fluctuations are stable, so the value of ENT is small. On the contrary, in anomalous traffic series, the fluctuations are unstable and the value of ENT is increasing due to the malicious events. Therefore, from Fig. 6, we can conclude that the feature vectors that consist of these three recurrence characteristics can accurately describe the traffic behavior patterns and differ the anomalous behaviors from the normal.

3.3 Implementation process of WRC

Based on the dynamic feature vectors, this section illustrates the detailed implementation process of the WRC detection model as follows:

Input: network traffic time series (five statistics) $x = \{x_i\}, i = 1, 2, ..., n$

Output: normal traffic and anomalous traffic

Step 1: for a traffic time series x, employs the wavelet transform to reconstruct the low-frequency traffic series l and the high-frequency traffic series h;

Step 2: based on the sliding window, uses RQA method to calculate the recurrence characteristics of different traffic series l and h, respectively;

$$l = \{\boldsymbol{Fl_r}\} = \{[f_{r,RR}, f_{r,DET}, f_{r,ENT}]\} \quad r = 1, 2, ..., N_w \tag{10}$$

$$h = \{\boldsymbol{Fh_r}\} = \{[f_{r,RR}, f_{r,DET}, f_{r,ENT}]\} \quad r = 1, 2, ..., N_w \tag{11}$$

where r is the r^{th} subseries, N_w is the number of subseries, $N_w = \frac{n-W}{W_s} + 1$. The traffic series x can be expressed as follows:

$$x = \{\boldsymbol{F_r}\} = \{[\boldsymbol{Fl_r}, \boldsymbol{Fh_r}]\} \tag{12}$$

Step 3: for each traffic series (five statistics), repeats step 1 and step 2 and then combines the dynamic characteristics of the five traffic series together to describe the traffic behavior patterns. The expression is as follows:

$$\boldsymbol{X} = \{\boldsymbol{X_r}\} = \{[\boldsymbol{F_r^1}, \boldsymbol{F_r^2}, \boldsymbol{F_r^3}, \boldsymbol{F_r^4}, \boldsymbol{F_r^5}]\} \tag{13}$$

Step 4: uses k-means algorithm to classify each $\boldsymbol{X_r}$ into different clusters and identify the anomalous traffic based on the threshold rule.

4 Evaluation of WRC detection model

In this section ,we evaluate the performance of WRC detection model on DARPA 1999 dataset and compare it with existing detection methods.

4.1 Evaluation metrics

We use the following three criterions to evaluate the performance of the WRC detection mode.

- Detection accuracy rate (DAR): the ratio of the anomalous traffic that are truly detected by WRC model in traffic series to all anomalous traffic.
- False negative rate (FNR): the ratio of the anomalous traffic that are missed by WRC model in traffic series to all anomalous traffic.
- False positive rate (FPR): the ratio of the normal traffic that are incorrectly identified as anomalies by WRC model in traffic series to all normal traffic.

4.2 Evaluation results

DARPA 1999 dataset includes five weeks traffic. We choose the first three weeks traffic as the training dataset and the rest as the testing dataset. During the training phase, based on the dynamic characteristics, traffic are classified into different clusters by our WRC detection model. The number of clusters is 10. In the testing phase, the traffic of week4 and week5 are identified as the normal or anomalous traffic based on the threshold rule. Table 2 lists the detection results of WRC model for week4 and week5. The results show that week4day5 has the highest detection accuracy, and the average detection accuracy and false alarm rate are 92.62% and 8.91%.

From the evaluation results on DARPA 1999 dataset, we can conclude that based on the multi-scale dynamic characteristics of network traffic, WRC detection model can accurately describe the traffic behavior patterns and effectively detect the anomalous traffic in time.

Table 2: Detection results for week4 and week5

	DAR	FNR	FPR
Week4Day1	91.24%	8.76%	8.68%
Week4Day2	92.08%	7.92%	9.09%
Week4Day3	95.73%	4.27%	7.23%
Week4Day4	90.29%	9.71%	9.69%
Week4Day5	96.84%	3.16%	6.54%
Week5Day1	92.76%	7.24%	9.82%
Week5Day2	91.57%	8.43%	8.39%
Week5Day3	92.68%	7.32%	9.26%
Week5Day4	92.42%	7.58%	9.61%
Week5Day5	90.54%	9.46%	10.83%
Average	92.62%	7.38%	8.91%

4.3 Comparison results

WRC detection model adopts the wavelet transform method for extracting multi-scale characteristics and the k-means algorithm for identifying traffic anomalies. In the previous studies, these two methods have already been employed to detect anomalies. In order to validate that our model indeed improves the detection performance, this section compares the WRC detection model with the existing detection methods.

While, in previous studies, the wavelet transform was usually combined with other methods to detect network anomalies. In the purpose of indicating the impact of wavelet transform on the detection performance through a fair comparison, we propose the Recurrence-Clustering (RC*) detection model that is similar to WRC model except employing the wavelet transform method. Thus, this section performs comparisons with k-means and RC* detection model. Table 3 shows the results.

Table 3: Comparison results

	DAR	FNR	FPR
k-means	76.19%	23.81%	25.74%
RC*	83.54%	16.46%	12.91%
WRC	92.62%	7.38%	8.91%

From Table 3, we can see that among these three methods, k-means has the worst detection performance. That is because k-means uses the statistical observations of network traffic to detect anomalies, which have certain randomness and may be unreliable to characterize the traffic behavior patterns, leading to the high false positive rate.

Like WRC, the RC* detection model also employs the dynamic characteristics to describe the traffic inherent behavior patterns, which are sensitive to the non-stationary transition caused by anomalies. Thus, compared with k-means, its detection performance is indeed improved, especially the false positive rate is significantly reduced. However, its detection performance is still poorer than that of WRC model. That is because WRC model calculates the dynamic characteristics of network traffic at different time scales based on the wavelet transform, which can highlight the local and short-timescale variations of traffic behavior caused by anomalies, resulting in the good detection performance for WRC detection model.

From Table 3, we can conclude that compared with the traditional statistical detection methods, WRC model can accurately detect traffic anomalies in time and obviously improve the detection performance, i.e., it has the high detection accuracy with a low false alarm rate.

5 Conclusions

The statistical observations of network traffic have certain randomness, which may vary along with network scales or application environments. Thus, it is difficult to accurately describe the traffic behaviors by adopting traditional statistical detection methods. This paper proposes a novel network anomaly detection model based on the multi-scale dynamic characteristics of traffic, i.e., the Wavelet-Recurrence-Clustering (WRC) detection model. The WRC detection model identifies traffic anomalies based on the inherent dynamic features of the traffic at different frequencies. Evaluation results on DARPA 1999 dataset show that WRC model has better detection performance, compared with existing methods. More specifically, based on the multi-scale recurrence dynamic characteristics, our WRC detection model can accurately describe the traffic behaviors and timely discover the non-stationary transition caused by malicious events, which leads to the good detection performance.

Bibliography

[1] Kim, H. J.; Na, J. C.; Jang, J. S.; Network traffic anomaly detection based on ratio and volume analysis, *International Journal of Computer Science and Network Security*, 6(5): 190-194, 2006.

[2] Wu, Q.; Shao Z.; Network anomaly detection using time series analysis, *Proc. of the Joint Int. Conference on Autonomic and Autonomous Systems and International Conference on Network and Services*, Papeete, Tahiti, 42-47, 2005.

[3] Willinger, W.; Paxson, V.; Taqqu, M. S.; Self-similarity and heavy tail: structural modeling of network traffic, *A Pratical Guide to Heavy Tails: Statistical Techniques and Applications*, BirkhRăuser, Boston, USA, 1998.

[4] Grossglauser, M.; Bolot, J. C.; On the relevance of long-range dependence in network traffic, *IEEE/ACM Transactions on Networking*, 7(5): 629-640, 1999.

[5] Tsai, C. F.; Hsu, Y. F.; Lin, C.; Lin, W.; Intrusion detection by machine learning: a review, *Experts Systems with Applications*, 36(10): 11994-12000, 2009.

[6] Shon, T.; Moon, J.; A hybrid machine learning approach to network anomaly detection, *Information Science*, 177: 3799-3821, 2007.

[7] Gaddam, S. R.; Phoha, V. V.; Balagani, K. S. ; K-Means+ID3: a novel method for supervised anomaly detection by cascading K-Means clustering and ID3 decision tree learning methods, *IEEE Transactions on Knowledge and Data Engineering*, 19(3): 345-354, 2007.

[8] Sabhnani, M.; Serpen, G.; Why machine learning algorithms fail in misuse detection on KDD intrusion detection dataset, *Intelligent Data Analysis*, 8(4): 403-415, 2004.

[9] Barford, P.; Kline, J.; Plonka, D.; Ron, A.; A signal analysis of network traffic anomalies, *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet Measurement*, Marseille, France, 71-82, 2002.

[10] Polikar, R.; Wavelet tutorial, http://users.rowan.edu/ polikar/WAVELETS/ WTtutorial.html, 2001.

[11] Eckmann, J. P.; Kamphorst, S. O.; Ruelle, D.; Recurrence plots of dynamical systems, *Europhysics Letters*, 4(9): 973-977, 1987.

[12] Zbilut, J. P.; Webber, C. L.; Embedding and delays as derived from quantification of recurrence plots, *Physics Letter A*, 171: 199-203, 1992.

[13] Duda, R. O.; Hart, P. E.; Stork, D. G.; Pattern classification, 2rd edn., Wiley-intersicence, New York, USA, 2000.

[14] DARPA 1999; http://www.ll.mit.edu/mission/communications/cyber/CSTcorpora/ideval/data/1999data.html, 1999.

[15] Ohira, T.; Schreiber T.; Nonlinear time series analysis, 2rd edn., Cambridge University Press, New York, USA, 2004.

[16] Chen, W. (2006); Study on the identification of two-phase flow patterns, *Master Thesis.*

[17] Marwan, N.; Romano, M. C.; Thiel, M.; Kurths, J.; Recurrence plots for the analysis of complex systems, *Physics Reports*, 438: 237-329, 2007.

INT J COMPUT COMMUN, ISSN 1841-9836
9(1):113-124, February, 2014.

A Novel Entity Type Filtering Model for Related Entity Finding

J. Zhang, Y. Qu, S. Tian

Junsan Zhang
1.College of Computer & Communication Engineering,
China University of Petroleum,
Qingdao 266580, P.R.China
2.School of Computer and Information Technology,
Beijing Jiaotong University,
Beijing 100044, P.R.China
zhangjunsan@upc.edu.cn

Youli Qu*, Shengfeng Tian
School of Computer and Information Technology,
Beijing Jiaotong University,
Beijing 100044,P.R. China
*Corresponding author: ylqu@bjtu.edu.cn
sftian@bjtu.edu.cn

Abstract: Entity is an important information carrier in Web pages. Searchers often want a ranked list of relevant entities directly rather a list of documents. So the research of related entity finding (REF) is a meaningful work. In this paper we investigate the most important task of REF: Entity Ranking. To address the issue of wrong entity type in entity ranking: some retrieved entities don't belong to the target entity type. We propose a novel entity type filtering model in which the target types are composed of the originally assigned type and the new type which is automatically acquired from the topic's narrative to filter wrong-type entities. For the query, we propose a method to process the original narrative to acquire a new query which is composed of noun and verb phrases. The results of experiments show our novel type filtering model gets a better result than the traditional filtering model at whatever precision and recall. Also the experiment shows the method that we acquire a new query is feasible.

Keywords: related entity finding, entity, entity ranking, type filtering.

1 Introduction

Along with the rapid development of internet, the number of web pages becomes more and more, so mass information is being produced now. Search engine became an important tool to query information from web in people's life. If a user is looking for entities, which have a specific relationship to some entity, he has to scan the documents retrieved by a Search Engine system to look for entities. For instance, when searchers submit a query *Michael's teammates while he was racing in formula 1* [1], searchers want some related entities actually. Related entity finding task can meet searchers requirements. According to the definition of TREC2009 Entity track, related entity finding(REF): given a source entity, a relation and a target type, identify homepages of target entities that enjoy the specified relation with the source entity and that satisfy the target type constrain [1]. REF provides a new way of information searching through entities. Entity ranking is an important issue of REF .Two elements can affect the result of entity ranking, target entity type and entity relation between source entity and target entity. In this paper we focus on the effect of target entity type to entity ranking. Because wrong type entities pollute the result of entity ranking, we try to filter the entities of wrong type. However the common type

type filtering method is too coarse to filter wrong entities exactly. Therefore, we propose a novel type filtering model to filter wrong entities. We utilize the Wikipedia category information as the source of entities types. Also we observe carefully the effect of experiment and we see using the novel type filtering model can get a better result at both recall and precision . To the issue of extracting query , we propose an approach in which parsing the narrative's syntactic structure and rewriting the query. The paper is organized as follows: Section 2 provides an overview of related work, Section 3 gives a description of the basic architecture of REF, Section 4 gives a detailed description of the proposed method, in Section 5 we utilize data set to implement our proposed method and analyze the experimental results, in Section 6 we draw the conclusion and propose our works in the future.

2 Related Work

The entity retrieval originate natural language processing, specifically IE (information extraction). Finding all entities for a certain class is the target of IE, i.e., extracting entities based patterns or learned from examples or created manually [2]. QA (question answering) is the intersection of natural language processing and IR which combines IE and IR. It looks like the REF, yet it differs from REF: (i) an entity is not always been contained in QA query list [3] (ii) REF task add a special relation between target entities and source entity [1]. As an important issue of REF task, entity ranking begin with ranking a specific type entities, e.g. persons in expert search [4]. The task of expert finding is finding experts either by modeling an expert's knowledge by its associated documents or collecting topic related documents first and then modeling experts [5]. Now it develops to rank more general types, e.g. persons, products, locations, organizations etc. The goal of entity ranking is retrieving entities as answers to a query. It is primarily focused on returning a ranked list of relevant entities [6]. What our concern are precision and recall. Type filtering can demote the wrong type entities and improve recall and precision. The novelty of our approach is we use co-occurrence model which widely used to estimate the strength of association between terms to estimate the associations between source entity and target entities and using a novel type filtering model filters wrong entities. We apply Wikipedia category information to as the source of entities types. Also we carefully analyze the effect of using the novel filtering model and the traditional filtering model for entity ranking . TREC has run an entity ranking track in 2009 aiming at performing entity-oriented search task on the web [1]. The definition of entity track is: given a source entity, a relation and a target type, find the relevant entities. It makes use of 20 topics, finds three types entities (persons, products, organizations). A query topic is defined as follows [1]:
$< query >$
$< num >$ 1 $< /num >$
$< entity\_name > Blackberry < /entity\_name >$
$< entity_url >$clueweb09-en0004-50-39593
$< /entity\_url >$
$< target\_entity > organization < /target_entity >$
$< narrative >$Carriers that Blackberry makes phones for.
$< /narrative >$
$< /query >$

A general approach of REF task is: (i) collecting text snippets from relevant documents (ii) obtaining entities by performing named entity recognition (iii) ranking relevant entities (iv) finding homepage [1]. Researchers propose several approaches to perform the REF task. Some researchers use different language modeling approaches where the entity model is constructed from text snippets and relation is utilized ad a query [7], [8]. Y.Wu et al. [9] develop an effective approach to rank entities via measuring the "similarities" between supporting snippets of entities and input query. Y.Fang et al. [10] propose a hierarchical relevance retrieval model for entity ranking. Three levels of relevance are examined which are document, passage and entity, respectively. R. Kaptein et al. [11] propose an approach using Wikipedia as a pivot for finding entities on the web, reducing the hard web entity ranking problem to easier problem of Wikipedia entity ranking.

3 The Basic Architecture of REF

The basic architecture of related entity finding is shown in Figure 1. The REF task can be divided into three main parts: (i) relevant documents retrieving (ii) candidate entities extracting and entity ranking (iii) homepage finding. We will describe the three parts in the following paragraph.

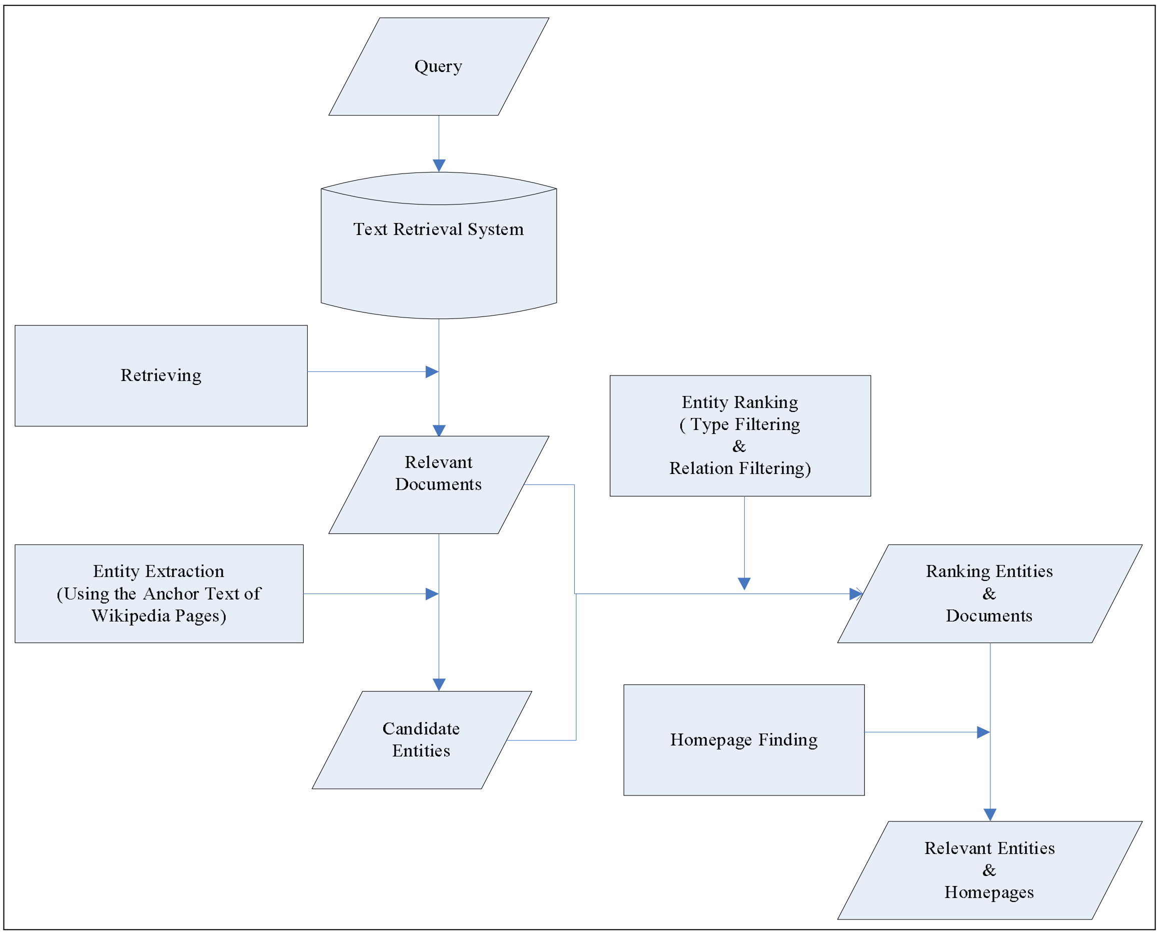

Query
Text Retrieval System
Retrieving
Relevant Documents
Entity Ranking (Type Filtering & Relation Filtering)
Entity Extraction (Using the Anchor Text of Wikipedia Pages)
Ranking Entities & Documents
Candidate Entities
Homepage Finding
Relevant Entities & Homepages

Figure 1: The Basic Architecture of REF

(1) Relevant documents retrieving. Retrieving relevant documents is the basic component of REF task. The first step is using corpus (here, we use the ClueWeb09 Category B as the documents repository) to build a full text retrieval system. Because our computing resource is limited, we make use of ”The Lemur Project” [12] which provides an online service of ClueWeb09 Category B as our source. Secondly, we send a query to the retrieval architecture and preliminary generate some candidate answers. For the selection of query, we chose the noun phrases and verb phrases as the query. We are interested in noun groups and verb groups because the

noun groups often qualify the target entity and source entity in more detail and can be seen as a kind of a selection criterion. Through extracting the part-of-speech elements from the narrative, we get the noun phrases and verb phrases of each topic. For example, the topic19's (Entity Track09) narrative is "Companies that John Hennessy serves on the board of". After the parsing the syntactic structure, the query is obtained: "Companies John Hennessy serves serves". For detecting the feasibility of acquiring query through this approach, we respectively utilize the "pure narrative" and the "noun and verb phrases" as the query to retrieve documents in experimental part.

(2) Entity Ranking. Entity ranking is the focus in this paper. After generating the relevant documents, the traditional following step is named entity recognition (NER), yet NER is not our emphasis in this paper. We handle NER by considering only anchor texts as entity occurrences in Wikipedia pages [13], [14] . When we get some candidate entities, we hope to find the most relevant entities.So we need to rank the candidate entities. There are two factors will effect the entity ranking: entity type and entity relation. The wrong type entities and the entities which do not conform to the relation between source entity and target entity will pollute the ranking result. In this paper,We only focus on the issue of "filtering wrong type entities".

(3) Homepage Finding. An entity is uniquely identified by its homepage, according to the definition of REF. Three homepages and a Wikipedia page at most can be returned for each entity in 2009 Entity Track. Homepage finding can be seen a document retrieval problem which employs a standard language modeling [15] to ranks homepages according to the query likelihood: $p(q = e/d)$, using the entity's name as a query. This issue is also not the emphasis in this paper.

4 Entity Ranking Model

According to the definition of REF, given a $Q(E^s, T, R)$, return a ranked list of relevant entities. In the paper, we use E^s to indicate the source entity, E^t indicate the target entity, T indicate the target type, R indicate a relation between E^s and E^t. Using the conditional probability formula $P(E^t|Q)$ estimates REF task. Due to the condition of $P(E^t|Q)$ is complex and difficult to estimating. Next we rewrite $P(E^t|Q)$ to:

$$P(E^t|Q) = \frac{P(E^t, Q)}{P(Q)} \tag{1}$$

Considering the denominator $P(Q)$ does not influence the ranking of entities, we derive the ranking formula as follows:

$$\begin{aligned}
P(E^t, Q) &= P(Q|E^t) \cdot P(E^t) && (2)\\
&= P(E^s, T, R|E^t) \cdot P(E^t) \propto P(E^s, R|E^t) \cdot P(T|E^t) \cdot P(E^t) && (3)\\
&= P(E^s, R, E^t) \cdot P(T|E^t) = P(R|E^s, E^t) \cdot P(E^s, E^t) \cdot P(T|E^t) && (4)\\
&= P(R|E^s, E^t) \cdot P(E^t|E^s) \cdot P(E^s) \cdot P(T|E^t) && (5)\\
&= P(R|E^s, E^t) \cdot P(E^t|E^s) \cdot P(T|E^t) && (6)
\end{aligned}$$

We assume that type T is independent of source entity and relation R in (3). Assuming $P(E^s)$ is a uniform value in (5), we drop it. Now the ranking task is converted to three conditional probability question: $P(R|E^s, E^t)$, $P(E^t|E^s)$, $P(T|E^t)$. In this paper, our goal is to address the issue of wrong type polluting entity ranking. So we only discuss $P(E^t|E^s) \cdot P(T|E^t)$ in this paper.

(1) Co-occurrence model

We see $P(E^t|E^s)$ as a co-occurrence issue expresses the association between source entity E^s and target entity E^t. We use a formula to estimate $P(E^t|E^s)$ as flows:

$$P(E^t|E^s) = \frac{Co(E^t, E^s)}{\sum_{E^{t'}} Co(E^{t'}, E^s)} \tag{7}$$

$E^{t'}$indicates an entity co-occurrence with source entity E^s in documents. We use two approach to estimate $C(E^t, E^s)$: (1) maximum likelihood estimate, (2) χ^2 hypothesis test [13], [16].

Maximum likelihood estimate(MLE):

$$Co_{MLE}(E^t, E^s) = C(E^t, E^s)|C(E^s) \tag{8}$$

Where $C(E^t, E^s)$ indicates the number of documents in which E^t and E^s co-occurrence, $C(E^s)$ indicates the number of documents in which $C(E^s)$ occurrence.

χ^2 hypothesis test:

$$Co_{\chi^2}(E^t), E^s)) = \frac{N \cdot (C(E^t, E^s) \cdot C(E^{\bar{t}}, E^{\bar{s}}) - C(E^t, E^{\bar{s}}) \cdot C(E^{\bar{t}}, E^s))^2}{C(E^s) \cdot C(E^t) \cdot (N - C(E^t)) \cdot (N - C(E^s))} \tag{9}$$

Where$E^{\bar{t}}$, $E^{\bar{s}}$ indicate the E^t and E^s don't appears respectively, and N indicates the total number of documents. For example, $C(E^{\bar{t}}, E^{\bar{s}})$ expresses the number of documents in which both E^t and E^s don't appear.

(2) Entity type filtering model

The co-occurrence model preliminary ranks entities. But it can not resolve the problem of wrong type entities pollute the ranking result. To address the issue of wrong type entities will pollute the ranking result. Traditional type filtering model deal with $P(T|E^t)$: the relation between target entity type and candidate entity type as flows:

$$P(T|E^t) = \begin{cases} 1 & \text{if } C(E^t) \cap C(T) \neq \phi \\ 0 & otherwise \end{cases} \tag{10}$$

Here, the $C(T)$ indicates the expected target entity type and the $C(E^t)$ indicates the type of candidate entity. The former is previously defined, although the latter is acquired via the Wikipedia category information of candidate entity. If they have an intersection we think the probability is 1, otherwise the probability is 0.

Although utilizing traditional entity filtering model can filter some wrong type entities, it is not enough accurate sometimes. According to the definition of REF, the types of target entities are divided into several types which are too wide to a certain extent. Such as, for entity track 2009, there are only three types of target entities which are assigned to 20 topics: person, organization and product. Yet, we see the exact target type of each topic should be different through the observation of topics' narratives. For example, there are two topics which have same target type (person). But they have completely different narratives: "Authors awarded an Anthony Award at Bouchercon in 2007" , "Chefs with a show on the Food Network". From the narratives, the exact type which the former want is Authors but the latter want is Chefs. Certainly, authors and chefs are both persons, yet they are also two different kinds of persons. So if we can refine the target type according to the topic's narrative, it may filter wrong type entity more accurately. We propose a novel entity filtering model to estimate $P(E^t|E^s)$ as flows:

$$Score(T) = Score(T_t) + Score(T_n) = \frac{F_{c,t}}{F_t} + \frac{F_{c,n}}{F_n} \tag{11}$$

Where: $Score(T)$ - score of type that a candidate entity get on the whole; $Score(T_n)$- score of type that a candidate entity get through the type acquired from topic's narrative; $Score(T_t)$-score of type that a candidate entity get through the topic's assigned target type; $F_{c,t}$- number of category features that a candidate entity type and the topic's assigned target entity type have in common; F_t- number of category features that the topic's assigned target entity type has; $F_{c,n}$-number of category features that the candidate entity type and the type acquired from topic's narrative have in common; F_n- number of category features that acquired from topic's narrative.

In order to calculate the $Score(T_t)$, we first get $Cat(t)$ - category information of target entity type that a topic is assigned and its sub-categories (one level down) and $Cat(c)$ - category information of a candidate entity. Then we can acquire $F_{c,t}$ and F_t through the category information of themselves. In this paper, we make use of Wikipedia category information as the criterion of judgment. However, The Wikipedia category structure is not a strict hierarchy and the category assignments are imperfect [17]. So we must process the category information further. The number of features that $Cat(t)$ has is too much to use directly. We find many categories of $Cat(t)$ often have a common structure (starting with "something" and ending with "something"). We select the top five categories or structures that appear in $Cat(t)$ as the selected features. For example, the selected features of a assigned target entity type - "Person" are:

- 'Living People'
- Ending with 'births'
- Ending with 'deaths'
- Starting with 'People'
- Ending with 'People'

For $Score(T_n)$, we first get $Cat(n)$ - the category information of target entity type that we get from the topic's narrative and its sub-categories and $Cat(c)$ - the category information of a candidate entity. Then we can acquire $F_{c,n}$ and F_n through the category information of themselves. To get $Cat(n)$, the category names in the narrative must extract first. We process the topics' narratives using Brill's Part-of-Speech [18] tagger and a Noun-Phrase chunker [19]. The first noun phrase in the narrative is the category that we want. For example, the narrative "Chefs with a show on the Food Network" is processed as follows:

```
(ROOT
(NP
(NP (NNS Chefs))
(PP (IN with)
(NP (DT a) (NN show)))
(PP (IN on)
(NP (DT the) (NNP Food) (NNP Network)))))
```

The noun Chefs is the type that is extracted from the narrative. Then we utilize the Wikipedia category information to get $Cat(n)$. Next, the category features of $Cat(n)$ is selected and the process is same as that is described in the previous paragraph. For example the selected features of "Chefs" are:

- 'Chefs'
- Ending with 'Chefs'
- Ending with 'Characters'
- Ending with 'births'
- 'Living People'

5 Experiment

5.1 Dataset and Evaluation Measures

In this paper, we use ClueWeb09 Category B subset as our corpus, including about 50 million documents. TREC 2009 Entity Track has three kind of basic types which are Person, Product and Organization [1]. We restrict the scope of entities only in Wikipedia pages. So we drop 5 topics in which no Wikipedia pages are retrieved and 15 topics are left.The Wikipedia is an excellent structured resource of entities. The title of the page is the name of the entity, the content of the page is the representation of the entity and each Wikipedia page is assigned to a number of categories. In our experiment, we get about 100 thousand relevant entities totally. After culling duplicated entities we get about 70 thousand relevant entities. We also make use of the DBpedia category information to get each entity's category information. DBpedia is a project aiming to extract structured content from the information created as part of the Wikipedia project. This structured information is then made available on the World Wide Web. DBpedia allows users to query relationships and properties associated with Wikipedia resources, including links to other related datasets. We use precision and recall as our estimation measures. Using $P@10$ to express precision and $R@N$ express recall where N taken to be 100, 2000.

In order to evaluate the effects that the different model produces for entity ranking and the feasibility that we acquire query from the original topic narrative. We divide the experiment into five steps: (1) Using the original topic narrative as the query retrieves relevant documents.
(2) Using the noun and verb phrases which is chosen from the original topic narrative as the query retrieves relevant documents.
(3) We only make use of pure co-occurrence model to rank candidate entities.
(4) We make use of traditional entity type filtering method to rank candidate entities based the result of using pure co-occurrence model.
(5) Utilizing our novel entity type filtering model to rank candidate entities based the result of using pure co-occurrence model.

5.2 Experimental Result

The precision and recall are our estimation measures. Using $P@10$ to express precision of top 10 entities retrieved and $R@N$ express recall where N taken to be 100, 2000. We utilize chart and table to describe the experimental data, showing the effects of different methods rank entities.

(1) For detecting the feasibility that using noun and verb phrases which are chosen from the narrative as the query, we make use of the narrative and the chosen phrases as the query to retrieve candidate entities respectively. The number of right entities which are retrieved from the documents are shown in Figure 2.

(2) We make use of pure co-occurrence (MLE), traditional type filtering model and our proposed novel type filtering model (based the MLE) to estimate the effect of entity ranking respectively. The result ($P@10$) is shown in Figure 3.

(3) We make use of pure co-occurrence (χ^2), traditional type filtering model and our proposed novel type filtering model (based the χ^2) to estimate the effect of entity ranking respectively. The result ($P@10$) is shown in Figure 4.

(4) We take topic14 as an example and list the top ten entities in Table 1. In order to observe the variation of the top ten ranked entities using different methods, more intuitively.

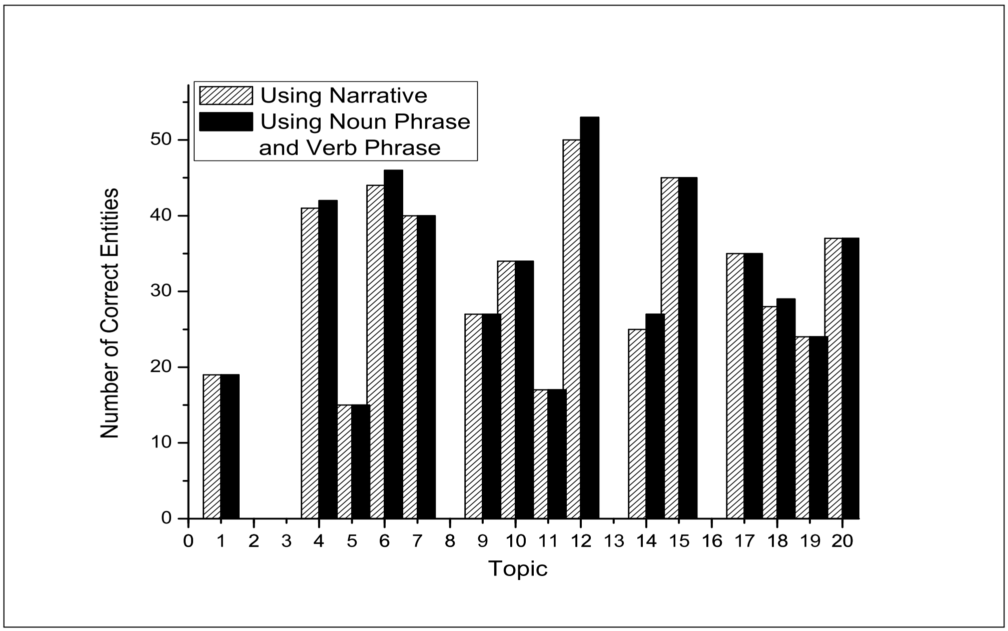


Figure 2: Number of correct entities using different query

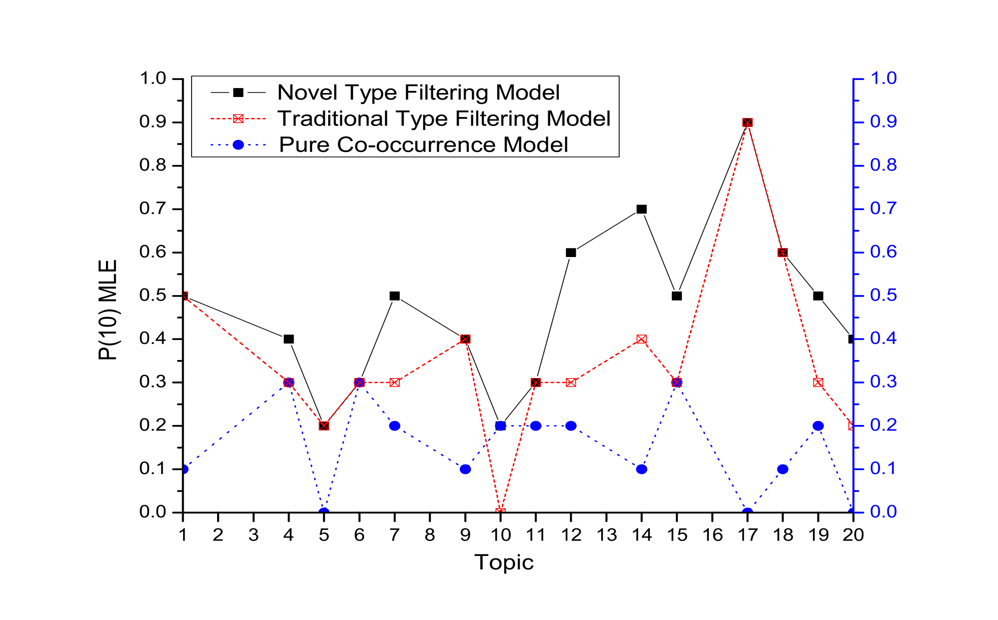


Figure 3: $P@10$: using MLE co-occurrence model, traditional type filtering model and novel type filtering model filters entities respectively.

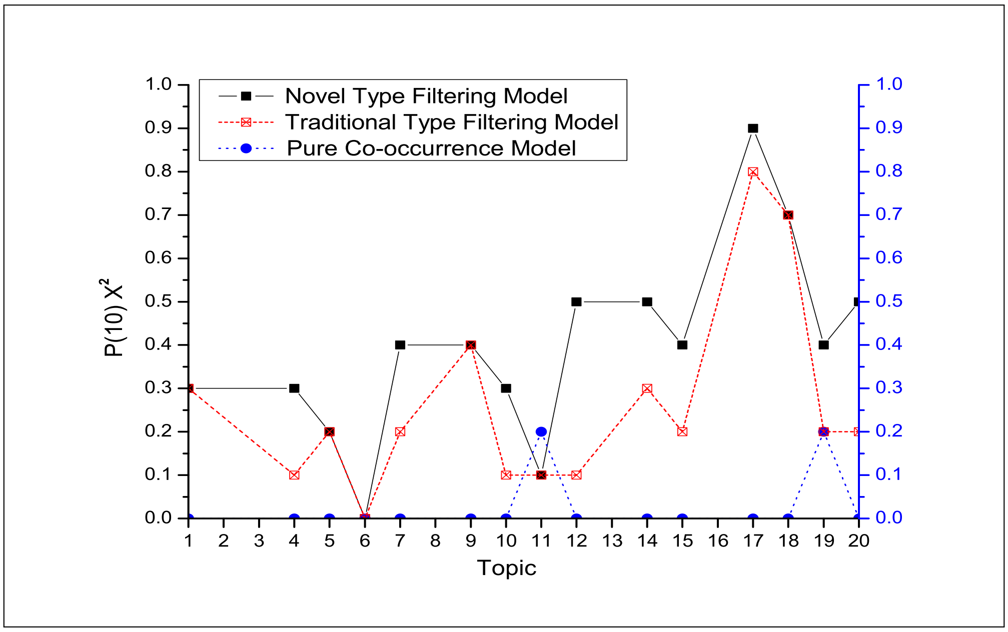


Figure 4: P@10: using χ^2 co-occurrence model, traditional type filtering model and novel type filtering model filters entities respectively.

(5) In order to evaluate the experiment effect in the whole. We list the results (the average of precision and recall of all topics) of using different models which are shown in Table 2.

5.3 Analysis

Figure 2 shows the number of correct entities which are retrieved from the documents using original narrative and phrases as queries respectively. We see using noun and verb phrases as queries retrieves more correct entities(increased by 2.07%) than using original narratives as queries. It certifies the approach that we extract noun and verb phrases as query is feasible.

Figure 3 shows the type filtering (based on MLE) whatever using traditional type filtering model or using our novel type filtering model causes a better result than using pure MLE co-occurrence model in P@10. The Figure 4 show the type filtering (based χ^2) whatever using traditional type filtering model or using our novel type filtering model causes a better result than using pure co-occurrence model in P@10. Through the Figures we find that utilizing type filtering can improve the effects of P@10. And our proposed novel type filtering model can filter wrong type entities better than traditional type filtering model. Table 1 demonstrates intuitively the variation of topic14's top ten ranked entities. Our novel type filtering model effectively removes the wrong type entities from the ranking.

As to recall, Table 2 shows the R@100 and R@2000 of using pure co-occurrence model, traditional type filtering model and using our novel type filtering model respectively. We see using the novel type filtering model (based on MLE) increases by 27.35% than using traditional type filtering model (based on MLE) and increases by 54.43% than using pure MLE co-occurrence model in R@100. Also we see using the novel type filtering model (based on χ^2) increases by 50% than using traditional type filtering model (based on χ^2) in R@100. The result illustrates than our novel type filtering model can filter wrong type entities than traditional type filtering model in R@100.

Pure Co-occurrence Model		Traditional Type Filtering Model		Novel Type Filtering Model	
MLE	χ^2	MLE	χ^2	MLE	χ^2
Bouchercon	Music	**Sue Grafton**	**Sherlock Holmes**	**Sue Grafton**	Lawrence Block
Book	Publisher	Agatha Christie	**Sue Grafton**	**Ross Macdonald**	Sue Grafton
Navigation	Husband	**Edgar Allan Poe**	**Edgar Allan Poe**	**Edgar Allan Poe**	Bill Pronzini
History	Television	**Sherlock Holmes**	Danny Boyle	George Washington	**Lee Child**
US	Crime	**Robert Crais**	Hercule Poirot	Agatha Christie	Marcia Muller
Son	Books	George Washington	George Washington	**Robert Crais**	**Ross Macdonald**
Organization	Performance	Hercule Poirot	Agatha Christie	Dennis Lehane	George Washington
Ant	Homicide	Stieg Larsson	Laurence Olivier	**Michael Connelly**	**Max Allan Collins**
Cher	Detective	Marcia Muller	Lawrence Block	**Lee Child**	Laura Lippman
Pub	Big	Ross Macdonald	Bill Pronzini	**Val McDermid**	**Val McDermid**

Table 1: The top ten ranked entities of topic14. The correct entities are indicated in bold.

		P@10	R@100	R@2000
Pure Co-occurrence Model	MLE	0.1533	0.1646	0.4287
	χ^2	0.0267	0.0438	0.3968
Traditional Type Filtering Model		P@10	R@100	R@2000
	MLE	0.3533	0.1996	0.2674
	χ^2	0.2600	0.1492	0.2674
Novel Type Filtering Model		P@10	R@100	R@2000
	MLE	0.4667	0.2542	0.3632
	χ^2	0.3933	0.2238	0.3627

Table 2: Recall and Precision of using different type filtering model filters entities respectively.

However, the data illustrates some different things in R@2000. We see using novel type filtering model (based on MLE) reduces by 15.28% than using pure MLE co-occurrence model and using traditional type filtering model (based MLE) reduces by 38.25% than using pure MLE co-occurrence model. And using novel type filtering model (based on χ^2) reduces by 8.47% than using pure χ^2 co-occurrence model and using traditional type filtering model (based on χ^2) reduces by 32.61% than using pure χ^2 co-occurrence model. The result illustrates that the type filtering model is not accurate enough. In other words, the model may remove some correct entities incorrectly. But it is encouraging to see the novel type filtering model gets a better result than traditional type filtering model (it increases by 26.38% in MLE and 26.27% in χ^2).

6 Conclusions and Future Works

For the issue of related entity finding, entity ranking is still an important issue. While some entities do not confirm to the required entity type and will affect the ranking result, filtering wrong type entities is essential. First, we parse the original narrative and acquire the noun and verb phrases as the new query. Then we make use of a novel type filtering model and the traditional type filtering model to filter entities respectively. In the experiment section, we choose 15 topics which target entity types are Person, Product and Organization as our test topics. We compare the experiment results and find: (i) the approach that we acquire a new query is feasible (ii)using our novel type filtering model gets a better result than using the traditional type filtering model whatever in precision or recall. We also see the problem of our type filtering model: compare to the pure co-occurrence model(in R@2000),it reduces the recall.It indicates some correct entities are removed incorrectly and the model need to be improved further.

In this paper we only argue the problem of wrong entity type filtering, while the wrong relation also affects the entity ranking result. As future work, we plan to investigate two issues: (i) how can we optimize our type filtering model to improve recall and precision further (ii) how can we use relation filtering to further optimize the result of entity ranking.

Acknowledgement

This work is supported by the Fundamental Research Funds for the Central Universities of China (Program No.2011JBM231).

Bibliography

[1] K. Balog, A. P. de Vries, P. Serdyukov, P. Thomas, and T. Westerveld, Overview of the TREC 2009 entity track, *TREC 09*, 2009.

[2] E. Riloff, Automatically generating extraction patterns from untagged text, *AAAI*, 2:1044-1049, 1996.

[3] E. M. Voorhees, Overview of the TREC 2002 Question Answering Track, *TREC 02*, 115-123, 2009.

[4] K. Balog, People Search in the Enterprise. PhD thesis, University of Amsterdam, 2008.

[5] K. Balog, L. Azzopardi, and M. de Rijke, A language modeling framework for expert finding, *Inf. Proc. and Man.*, 45(1): 1-19, 2009.

[6] Jovan Pehcevski, James A. Thom, Anne-Marie Vercoustre ,Vladimir Naumovski, Entity ranking in Wikipedia: utilising categories, links and topic difficulty prediction, *Information Retrieval*, 13(5):568-600, 2010.

[7] W. Zheng, S. Gottipati, J. Jiang, and H. Fang, UDEL/SMU at TREC 2009 Entity Track, *TREC 09*, 2009.

[8] Q. Yang, P. Jiang, C. Zhang, and Z. Niu, Experiments on related entity finding track at TREC 2009, *TREC 09*, 2009.

[9] Y.Wu and H. Kashioka, NiCT at TREC 2009: Employing three models for Entity Ranking Track, *TREC 09*, 2009.

[10] Y. Fang et al, Entity retrieval with hierarchical relevance model, *TREC 09*, 2009.

[11] R. Kaptein, P. Serdyukov, A. de Vries, and J. Kamps, Entity ranking using Wikipedia as a pivot, *CIKM*, 2010.

[12] http://lemurproject.org.

[13] M. Bron and K. Balog and M. de Rijke, Ranking relfated entities: components and analyses, *CIKM*, 2010.

[14] P. Serdyukov and A. de Vries, Delft university at the TREC 2009 Entity Track: Ranking wikipedia entities, *TREC 09*, 2009.

[15] F. Song and W. B. Croft, A general language model for information retrieval, *CIKM 99*, 77-82, 1999.

[16] C. D. Manning and H. Schuetze, Foundations of Statistical Natural Language Processing, *The MIT Press*, 1999.

[17] A. de Vries et al, Overview of the INEX 2007 Entity Ranking Track, 245-251, 2007.

[18] E. Brill, Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging, *Computational Linguistics*, 21(4): 543-565, 1995.

[19] L. Ramshaw, and M. Marcus, Text Chunking Using Transformation-Based Learning, *Proc. of the Third ACL Workshop on Very Large Corpora*, MIT, 1995.

Author index