Volume 37 Number 4 December 2013 ISSN 0350-5596

Informatica

An International Journal of Computing and Informatics

Editorial Boards, Publishing Council

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor from the Editorial Board can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the list of referees. Each paper bears the name of the editor who appointed the referees. Each editor can propose new members for the Editorial Board or referees. Editors and referees inactive for a longer period can be automatically replaced. Changes in the Editorial Board are confirmed by the Executive Editors.

The coordination necessary is made through the Executive Editors who examine the reviews, sort the accepted articles and maintain appropriate international distribution. The Executive Board is appointed by the Society Informatika. Informatica is partially supported by the Slovenian Ministry of Higher Education, Science and Technology.

Each author is guaranteed to receive the reviews of his article. When accepted, publication in Informatica is guaranteed in less than one year after the Executive Editors receive the corrected version of the article.

Fuzzy Logic Based Delamination Detection in CFRP Panels

Shanglei Li
Dept. of Electrical and Computer Engineering,
Southern Illinois University Carbondale, IL 62901, USA
E-mail: shanglei@siu.edu, http://www.engr.siu.edu/IMEL/

Anish Poudel and Tsuchin Philip Chu
Dept. of Mechanical Engineering and Energy Process,
Southern Illinois University Carbondale, IL 62901, USA

Keywords: fuzzy logic, NDE, ultrasonic testing

Received: January 31, 2013

This paper presents an intelligent interpretation of ultrasonic C-scan results for carbon-fiber-reinforced plastic (CFRP) panels by using fuzzy logic approach. Ultrasonic C-scan results have relatively low resolution and poor imaging quality in anisotropic composites due to the speckle noise produced by the interference of backscattered signals. In this study, fuzzy logic was implemented to accurately determine a defect's shape and size and to avoid over-segmentation and under-segmentation. For this, first, a 3×3 mask was considered to define the central value and the mean value within the C-scan amplitude data. Then, five linguistic labels for the central value and mean value were defined as: very low, low, neutral, high, and very high so as to determine fuzzy sets for the fuzzy inference system (FIS). Combined with 25 fuzzy rules, the FIS was capable of making decisions based on fuzzy sets and fuzzy rules. Experimental results demonstrated this fuzzy logic method can detect the size and shape of sub-surface delamination correctly, and restrain the noises effectively. The authors believe this approach for automatic defect detection and classification can be an integral part of the development of an intelligent NDE expert system for composite structures in the future, thus making defect evaluation process much easier and more accurate.

Povzetek: Predstavljena je inteligentna metoda mehke logike za analizo z ogljikom ojačane plastike.

1 Introduction

Carbon-fiber-reinforced plastic (CFRP) panels are now widely being used in many structural applications, especially in the aviation industry, due to their superior thermal and physical properties compared to metals. However, low velocity impacts, for instance, bird or hail strikes on an aircraft, can cause impact damage in CFRP structures. Such damage can take the form of cracking, delaminations, or fiber fractures [1, 2]. The damages in CFRP structures are usually complicated and highly dependent on the properties of the constituent materials, fiber orientation, stacking sequence, and nature of loading [3]. Therefore, a fast and reliable non-destructive evaluation (NDE) process is constantly required to economically ensure the integrity, safety, and reliability of these structures. Ultrasonic NDE is increasingly being used in composite inspection because of its large surface, speed, and non-contact testing capabilities [4-7]. However, due to the anisotropic properties and non-homogeneous behavior of these structures, they have brought a lot of challenges in the NDE industry. One critical problem is how aggressively to decide whether or not variations in the C-scan results are defects [8]. Another problem is the risk of under-segmentation or over-segmentation of the defect area. Both incidents will affect the size, location, and even features of defects, which are critical for defect evaluation. To meet these challenges, various imaging segmentation approaches [9, 10] have been reported to aid the inspection technique. Most segmentation algorithms are based on discontinuity and similarity. In the first category, an abrupt change in density is considered as the edge. Typical edge detection algorithms are Laplacian of a Gaussian (LoG) and Zero crossings. In the second category, segmentation is achieved by partitioning an image into similar density regions according to a set of predefined criteria. Image thresholding and region growing, splitting and merging are typical algorithms in this category. However image segmentation is still one of the most difficult tasks in image processing. Segmentation accuracy determines the eventual success or failure of computerized analysis procedures [10]. A study has demonstrated that rule based algorithms have better performance than the traditional image segmentation method in distinguishing defect areas [11, 12].

For this work, a rule based fuzzy logic approach was applied to solve this problem. The algorithm utilizes the fuzzy inference system of a center element and its eight neighboring elements to define a new objective function and determine the variance by classifying the function. The paper is organized as follows: a discussion of the basic theory, algorithm of fuzzy logic rules, and the experimental setup. They are then followed by the

experimental result. Finally, conclusions are provided at the end.

2 Fuzzy logic theory and application

Fuzzy logic originally developed by Zadeh [13] is not a logic that is fuzzy, but the logic that is used to describe fuzziness. It is an integral component of an expert system and has been widely implemented in many control and prediction systems because it can tackle many problems under various assumptions and approximations with greater accuracy. In addition, its extraordinary controlling and reasoning capabilities have also made it popular in many complex industrial systems. In a fuzzy system, it is possible to define expert knowledge even if statistical data is not available. A fuzzy rule is mathematically described as a fuzzy relation between the sets describing the antecedent and consequent. Each rule in a fuzzy logic is expressed by the following relation [14]:

$$R_i = \left\{ \begin{matrix} ((x,y),\mu_R(x,y)) \mid \\ (x,y) \in A_i \times B_i, \mu_R(x,y) \in [1,0] \end{matrix} \right\} \quad (1)$$

where, $x \in$ X and $y \in$ Y, A_i and B_i are fuzzy subsets of the domains X and Y associated with linguistic labels, $R_i(x,y)$ is a fuzzy relation defined on the Cartesian product universe X × Y.

A general fuzzy inference system (FIS) is shown in Figure 1. It consists of crisp input, fuzzifier, knowledge base, inference methods, deffuzifier, and a crisp output. The FIS takes a crisp input and determines the degree to which they belong to each of the appropriate fuzzy sets via membership functions. A membership function is a curve that defines how each point in the input space is mapped to a membership value. The fuzzifier then measures the value of input variables and performs a scale mapping that transfers the range of values of input variables into corresponding universes of discourse. The knowledge base consists of fuzzy sets and fuzzy rules. Fuzzy sets provide the necessary definitions which are used to define linguistic rules and fuzzy data manipulation, and fuzzy rules characterize the control goals and control policy of domain experts by means of a set of linguistic control rules.

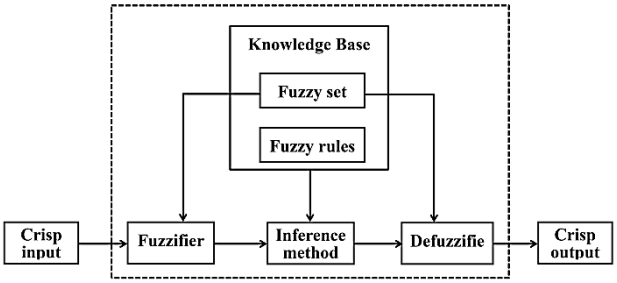


Figure 1: Fuzzy Inference System.

The inference method is the kernel of the FIS and it has the capability of making decisions based on fuzzy sets and fuzzy rules. Finally, defuzzifier performs a scale mapping that converts the range of values of output variables into corresponding universes of discourse.

3 Fuzzy logic algorithm

3.1 Central value and local mean value

The C-scan result obtained from ultrasonic testing is a 2D matrix corresponding to the plan-type view of the location and size of the test specimen. Each element of the matrix indicates the coordinate and the amplitude of received signals. Let $x(i,j)$ be the ultrasonic amplitude data of the element (i,j) in a two dimensional M × N matrix. The mask is defined as a (2*m*+1) × (2*n*+1) window centered at (i,j) where m and n are integers. This window will go through each element to obtain central amplitude and local mean amplitude values of every element in the M × N matrix. Note that the window's shape is not necessarily a square. The central amplitude value is:

$$C_x(i,j) = x(i,j) \quad (2)$$

The local mean of an element (i,j) can be computed as:

$$m_x(i,j) = \frac{1}{(2m+1)\times(2n+1)} \sum_{l=j-n}^{j+n} \sum_{k=i-m}^{i+m} x(k,l) \quad (3)$$

In equations (2) and (3), the parameters of the central value distribution and local mean value distribution for a given matrix are strongly dependent on the window size (2*m*+1) × (2*n*+1). For this, the data are assumed strongly correlated between the central element and its $m \times n - 1$ neighbors. Thus, the computed central amplitude and local mean amplitude will increase as the window size is increased. The window size also depends on the detail pattern as well as the C-scan data "resolution" (step increment of x and y axes). Higher resolution C-scans should use larger window sizes to facilitate the visualization of local details. However, a large window increases the computational requirement. Thus, there is a trade-off between the enhancement of local details and computational loading when determining the proper window size. In this study, to simplify the task, we choose $m = n = 1$, i.e. a 3×3 window as shown in Figure 2. For the given CFRP specimen A, the 2D matrix of C-scan result has 361×961 elements. A 3×3 window is large enough to carry sufficient detail and small enough to keep lower computational time in the whole 2D matrix area.

$i-1, j-1$	$i, j-1$	$i+1, j-1$
$i-1, j$	i, j	$i+1, j$
$i-1, j+1$	$i, j+1$	$i+1, j+1$

Figure 2: Applied 3 × 3 mask with $m = n = 1$.

3.2 Membership functions

The central value and local mean value for a 3 x 3 mask were used to define the membership functions. For this,

five linguistic labels for the central value and the mean value were defined as very low, low, neutral, high, and very high. These levels are based on threshold values for each element amplitude signal value. For example, low amplitude indicates less received ultrasonic signal, which has a greater probability to be classified as a defect. The central value classes are denoted as C_{VL}, C_L, C_N, C_H, and C_{VH}. Similarly, the local mean value is classified as M_{VL}, M_L, M_N, M_H, and M_{VH}. To separate different classes, 4 groups predefined thresholds $\alpha_1, \beta_1 \ \ldots\ldots\ \alpha_4, \beta_4$ are used as shown in Figure 3. These threshold values are determined experimentally. As an example, if a central amplitude value falls into the range of $[\beta_3, \alpha_4]$, it will be classified to "High" as C_H. If the value falls into $[\alpha_3, \beta_3]$, it partially belongs to both "Neural" and "High". In this case, 2 fuzzy rules are fired to determine the output linguistic label of this value. Similar functions and classes are determined for local mean values. Different values fall into different intervals and are classified into the corresponding linguistic labels (classes) appropriately. The linguistic labels and membership functions are depicted in Figure 3.

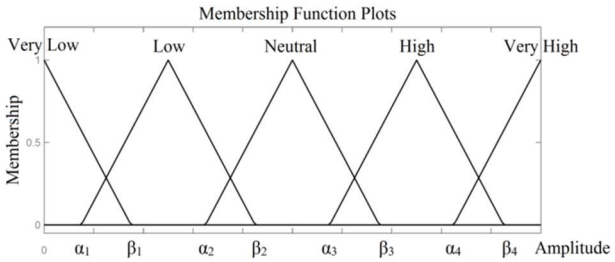


Figure 3: Input variable membership functions and thresholds.

For the output variable, 5 labels were attributed: VL (Very Low) indicating it is a positive defect, L (Low) indicating a potential defect, N (Neutral) indicating it may or may not be a defect, H (High) indicating a potential good area, and VH (Very High) indicating a positive good area. The inference method proposed by Sugeno was utilized in the output which is a constant value for each linguistic label of the variable in the range [0, 1]. The 5 output linguistic labels O_{VL}, O_L, O_N, O_H, and O_{VH} are shown in Figure 4.

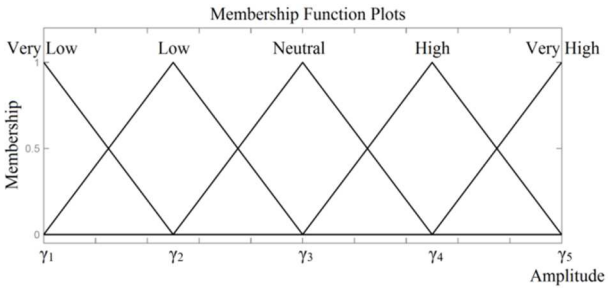


Figure 4: Output variable membership functions and thresholds.

3.3 Fuzzy logic rules

In the fuzzy inference system, fuzzy rules are usually elaborated arbitrarily based on experience and expert decision. It is impractical or impossible to find exact rule sets made by a mathematical formula or model. Especially for sufficiently complex problems, such as defect detection, mathematical methods cannot generate accurate sets of rules. In this case, mathematical methods can only support rules that have already been created. Thus, manual intervention based on expert knowledge is still required. Although most fuzzy rules cannot be accurately developed by a mathematical method, if one of the rules is wrong, even greatly wrong, the fuzzy inference system will compensate for the error just by firing the other correct rules. However, fuzzy rules should be decided carefully by using prior knowledge and NDE expert experience to avoid underperformance in the fuzzy inference system. These rules should be tested vigorously and refined if necessary.

For this study, two variables (central value and local mean value) are utilized as fuzzy inference system inputs; fuzzy logic rules are defined as the following:

Rule Number	**Inputs**		**Outputs (O)**	
	Central Value(**C**)	Mean Value(**M**)		
R1	VL	VL	L	Positive defect
R2	VL	L	VL	Positive defect
R3	VL	N	L	Potential defect
R4	VL	H	N	May or not be a defect
R5	VL	VH	H	Potential good area
R6	L	VL	VL	Positive defect
R7	L	L	L	Potential defect
R8	L	N	N	May or not be a defect
R9	L	H	H	Potential good area
R10	L	VH	VH	Positive good area
R11	N	VL	L	Positive defect
R12	N	L	L	Potential defect
R13	N	N	N	May or not be a defect
R14	N	H	H	Potential good area
R15	N	VH	VH	Positive good area
R16	H	VL	L	Potential defect
R17	H	L	N	May or not be a defect
R18	H	N	H	Potential good area

Rule Number	Inputs		Outputs (O)	
	Central Value(C)	Mean Value(M)		
R19	H	H	H	Potential good area
R20	H	VH	VH	Positive good area
R21	VH	VL	N	May or not be a defect
R22	VH	L	H	Potential good area
R23	VH	N	VH	Positive good area
R24	VH	H	VH	Positive good area
R25	VH	VH	VH	Positive good area

Table 1: Fuzzy logic rules.

In the fuzzy inference system, multiple rules can fire at once. For instance, if a central value falls into the region of $[\alpha_3, \beta_3]$, the overlap part of linguistic labels "Neutral" and "High", both rules will fire. In case the value is more "High" than "Neutral", the "High" rule will generate a stronger response. The fuzzy algorithm will evaluate the result that fired based on fuzzy rules in Table 1, and use an appropriate defuzzification method to generate the output response.

To make the fuzzy rules easy to visualize, a fuzzy associate matrix is depicted in Table 2.

Mean / Central	VL	L	N	H	VH
VL	VL	VL	L	N	H
L	VL	L	N	H	VH
N	VL	L	N	H	VH
H	L	N	H	H	VH
VH	N	H	VH	VH	VH

Table 2: Fuzzy rules in associative matrix.

As shown in Table 2, more weight is attributed to the mean values than central values. For instance, if the mean value is VH, but the central value is L, the central value is "isolated" by its 8 neighbors. Such a point should be considered as an independent "mutation" point due to the possibility of noise or system error. Therefore, mean values are given more weight than central values to make sure the local defect information in 3×3 window does not contain a misjudged signal mutation caused by noise. Eventually the output result of this point will be VH.

3.4 Defuzzification

For this study, the central of area (COA) defuzzification method [15] was utilized to obtain a crisp output value from FIS. In the COA method, first the area under the scaled membership functions and within the range of the output variable is calculated. Then, the geometric center of this area is obtained by using the following equation:

$$y_c^* = \frac{\sum_{i=1}^{n} \mu_i \times \gamma_i}{\sum_{i=1}^{n} \mu_i} \quad (4)$$

where: y_c^* is the desired crisp defuzzification value with the COA method.

μ_i is the i_{th} membership degree of input variables.

γ_i is the i_{th} output class center (output variables membership function).

n is the number of elements in a fuzzy set.

The prod method is applied to both of the conjunction evaluation of the rule antecedents and fuzzy rules implication. The aggregation of the rule outputs is carried out by the sum method. Experiment values of input variables are pre-determined with expert knowledge as follows:

Membership functions of central value:
$C_{\alpha_1} = 0.09$, $C_{\beta_1} = 0.11$, $C_{\alpha_2} = 0.15$, $C_{\beta_2} = 0.17$, $C_{\alpha_3} = 0.19$, $C_{\beta_3} = 0.21$, $C_{\alpha_4} = 0.25$, $C_{\beta_4} = 0.27$

Membership functions of the mean value:
$M_{\alpha_1} = 0.13$, $M_{\beta_1} = 0.15$, $M_{\alpha_2} = 0.17$, $M_{\beta_2} = 0.19$, $M_{\alpha_3} = 0.19$, $M_{\beta_3} = 0.21$, $M_{\alpha_4} = 0.22$, $M_{\beta_4} = 0.23$

Corresponding to Figure 3, output class center γ_i are pre-determined as:
$\gamma_1 = 0$, $\gamma_2 = 0.25$, $\gamma_3 = 0.5$, $\gamma_4 = 0.75$, and $\gamma_5 = 1$

A simple demonstration is given below to briefly explain how the fuzzy logic algorithm works. For one certain element in a 2D matrix of C-scan result, its central amplitude value is 0.105 V and its mean amplitude value (in 3×3 window) is 0.227 V. According to the input membership sets and defined fuzzy rules, during the fuzzy-inference process, 4 fuzzy logic rules are fired in parallel:

Rule 4 $\mu_4 = \mu(C_{VL}) \times \mu(M_H) = 0.125 \times 0.15$
$= 0.18175 \rightarrow \mu_4 \; in \; N \quad (\gamma_3 = 0.5)$

Rule 5 $\mu_5 = \mu(C_{VL}) \times \mu(M_{VH}) = 0.125 \times 0.35$
$= 0.04375 \rightarrow \mu_5 \; in \; H \quad (\gamma_4 = 0.75)$

Rule 9 $\mu_9 = \mu(C_L) \times \mu(M_H) = 0.375 \times 0.15$
$= 0.05625 \rightarrow \mu_9 \; in \; H \quad (\gamma_4 = 0.75)$

Rule 10 $\mu_{10} = \mu(C_L) \times \mu(M_{VH}) = 0.375 \times 0.35$
$= 0.13125 \rightarrow \mu_{10} \; in \; VH \quad (\gamma_5 = 1)$

Based on equation (4), the fuzzy output y^* with COA method can be obtained:

$$y_c^* = \frac{\sum_{i=1}^{n} \mu_i \times \gamma_i}{\sum_{i=1}^{n} \mu_i}$$
$$= (0.18175 \times 0.5 + 0.04375 \times 0.75 + 0.05625 \times 0.7 + 0.13125 \times 1)/(0.5 + 0.75 + 0.75 + 1)$$
$$= 0.8625 \quad (5)$$

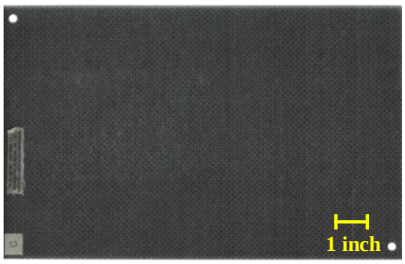
The fuzzy logic output y^* of C-scan data will be normalized to 0~255 and plot in 2D matrix corresponding to column and index, shown as a gray-level image.

4 Experimental setup

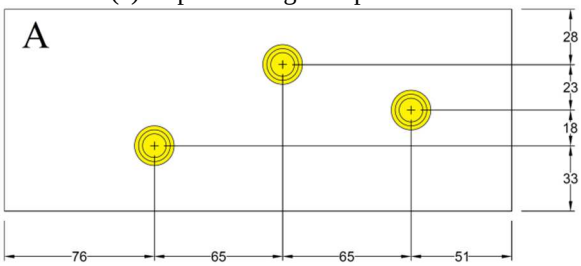
The immersion ultrasonic system with associated instrumentation used to inspect the CFRP panel is shown in Figure 5. A 5 MHz dual element Panametric transducer with a 2 inch focal length was utilized in a pulse-echo mode for the inspection. The standoff distance between the transducer and the panel was set to 2 inches and the scan was conducted at an increment of 0.01 inches.

Figure 5: Immersion ultrasonic testing system.

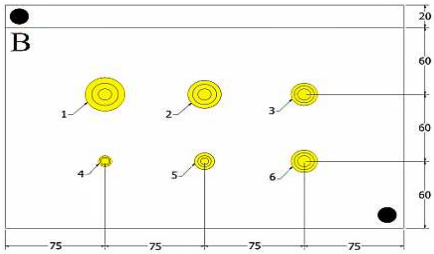
(a) Optical image of specimen A



(b) Optical image of specimen B



(c) Defect map of specimen A



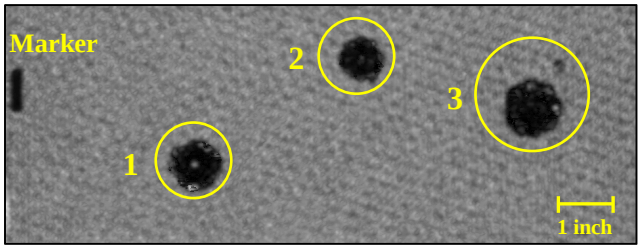
(d) Defect map of specimen B

Figure 6: Optical images and defect maps of specimen A and B. All dimensions are in mm.

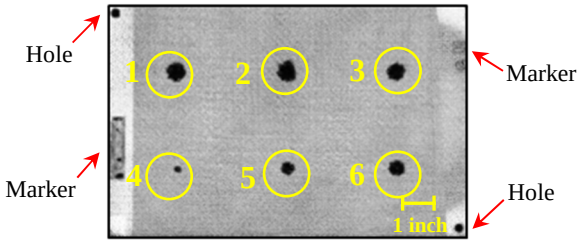
To verify the application of fuzzy logic defect detection, two different CFRP panels with predefined phantom defects, i.e. delamination defects due to impact damage were considered. These delamination defects were artificially simulated by impacting the panel with an external object of known energy. These defects are difficult to recognize by visual inspection, but have severely progressed within the panel. Specimen A is a 102 × 257 × 4.445 mm panel which consists of impact damage at three different locations. Similarly, specimen B measured 200 × 300 × 3.581 mm in dimensions. The optical images and defect maps of each specimen are shown in Figure 6.

5 Result and discussion

The fuzzy logic algorithm as described earlier, was applied to the ultrasonic C-scan results (maximum back wall amplitude data) obtained from both panels to verify the versatility and stability of the fuzzy inference system. The proposed method was implemented in MATLAB R2012b. The reconstructed raw C-scan results are presented in Figure7, where defect areas are represented by dark shade of gray i.e. significant drop in pulse-echo signal amplitude. The shaded area labeled “Marker” in Figure 7 is the marker that was attached to the panel for indication purposes.



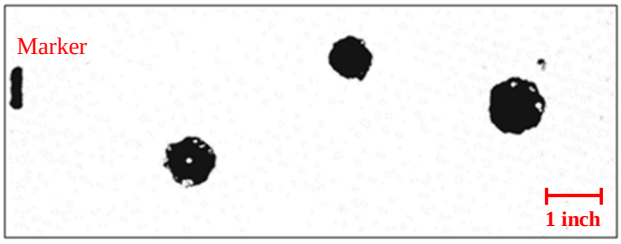
(a) Reconstructed C-scan image of specimen A



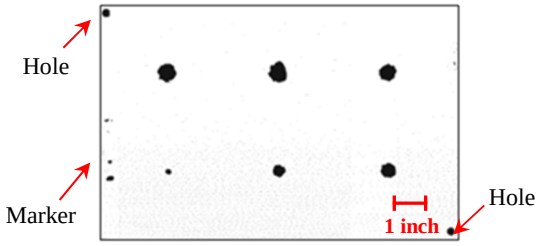
(b) Reconstructed C-scan image of specimen B

Figure 7: Reconstructed C-scan result for CFRP panels.

The fuzzy logic output of ultrasonic C-scan data for CFRP panels A and B are normalized and plotted in 8 bit grayscale images (256 gray-level) as shown in Figure 8. Figure 8 (a) and (b) are the fuzzy logic output results with the COA defuzzification method. From the results obtained, the fuzzy logic method is able to detect the defects with more confidence by eliminating the background compared to the raw C-scan image as in Figure 7 (a) and (b). The defect outline present is more distinct to recognize, allowing post-processing work such as measurement of defect size, shape, and location to be much easier.



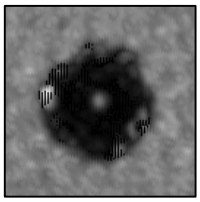
(a) Specimen A fuzzy logic output result image with the COA defuzzification method



(b) Specimen B fuzzy logic output results image with the COA defuzzification method

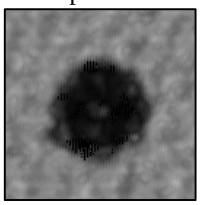
Figure 8: Fuzzy logic output results.

To demonstrate the effectiveness and robustness of the fuzzy logic method applied, defects in specimen A and 3 of 6 defects in specimen B are shown in Figure 9 and Figure 10, respectively. The experiment results indicate that the fuzzy logic method has satisfied performance on both CFRP panels (sample A and B), which have different carbon fiber orientation and laminates. As shown in Figure 9 and Figure 10, fuzzy logic results provide a clear and smooth edge area for all defects in sample A and B. The fuzzy logic method is able to remove the background noise in C-scan images to obtain high contrast and enhanced images.

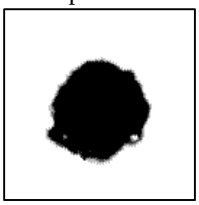


(a) C-scan image of sample A defect 1

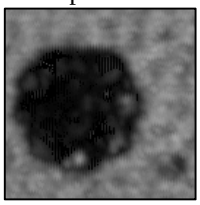
(b) Fuzzy result of sample A defect 1



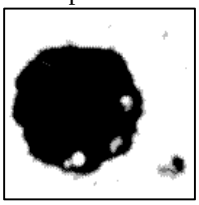
(c) C-scan image of sample A defect 2



(d) C-scan image of sample A defect 2

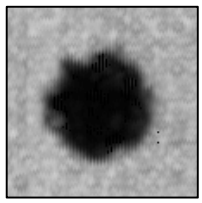


(e) C-scan image of sample A defect 3



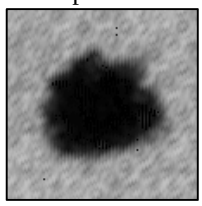
(f) C-scan image of sample A defect 3

Figure 9: Comparison of C-scan images and fuzzy logic results of 3 defects in specimen A.

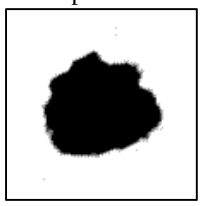


(a) C-scan image of sample B defect 1

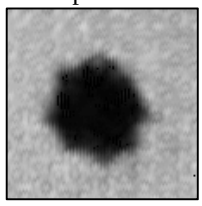
(b) C-scan image of sample B defect 1



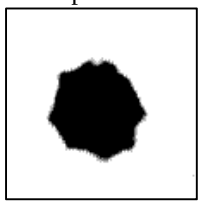
(c) C-scan image of sample B defect 2



(d) C-scan image of sample B defect 2



(e) C-scan image of sample B defect 3



(f) C-scan image of sample B defect 3

Figure 10: Comparison of C-scan images and fuzzy logic results of 3 defects in specimen B.

For subjective evaluation, fuzzy logic output results are compared to the reconstructed C-scan images side by side in Figure 9 and Figure 10. From the results obtained, the fuzzy logic method is able to detect the defects with more confidence by eliminating the noises seen in the C-scan images on the left side. The fuzzy logic output is capable of providing higher contrast of the defect area which allows NDE inspector make accurate decisions to identify the defect size and location.

In addition to the perceived image quality with human visual system (HVS), for objective evaluation, peak signal-to-noise ratio (PSNR) and contrast signal-to-

noise ratio (CNR) are employed for quantitative assessment. The fuzzy logic result and C-scan result are tested to demonstrate the image quality and robustness of the fuzzy logic method,
The PSNR is given as:

$$PSNR = 10 \cdot \log_{10}\left(\frac{{MAX_I}^2}{MSE}\right) \quad (6)$$

Where: MAX_I is the maximum possible pixel value of the image. In this study, all pixels are represented using 8 bits gray levels, here MAX_I is 255.
MSE is the mean squared error between two compared images.

The CNR is given as:

$$CNR = \frac{S_i - S_o}{\sqrt{\sigma_i^2 + \sigma_o^2}} \quad (7)$$

Where: S_i and S_o are the mean values inside and outside the ROI respectively
σ_i and σ_o are the standard deviations, respectively

	ROI		CNR (dB)	PSNR (dB)
Sample A	*Whole Sample*	C-scan result	5.63	6.17
		Fuzzy logic	8.97	
	Defect 1	C-scan result	5.09	7.18
		Fuzzy logic	6.20	
	Defect 2	C-scan result	4.38	7.08
		Fuzzy logic	5.07	
	Defect 3	C-scan result	3.83	7.47
		Fuzzy logic	4.69	
Sample B	*Whole Sample*	C-scan result	3.88	10.06
		Fuzzy logic	5.38	
	Defect 1	C-scan result	3.15	10.20
		Fuzzy logic	4.37	
	Defect 2	C-scan result	3.79	9.32
		Fuzzy logic	4.74	
	Defect 3	C-scan result	2.11	10.34
		Fuzzy logic	2.95	

Table 3: Experimental comparison of CNR & PSNR.

From Table 3, it can be verified that the CNR of the fuzzy logic output is higher than C-scan result. The PSNR has the close value in sample A and B, indicating the quality of the fuzzy logic method is robust and reliable. Further, note that the PSNR value of fuzzy logic output has an average increase of 6.98 dB in sample A and 9.98 dB in sample B compared to C-scan results. Since the applied method provides good performance in both CNR and PSNR, it can be seen that the proposed super-resolution reconstruction method is effective in resolution enhancement.

6 Conclusions

Analysis of the raw C-scan result of composites may not provide the reliable classification of different regions (defect, non-defect). A fuzzy logic methodology is applied to classify the defect and non-defect areas in CFRP panels with simulated delamination defects. The experimental results obtained for these panels have demonstrated the effectiveness of the applied method. It can be used as preprocessing of defect segmentation to reduce the computation complexity and time. However, membership function and fuzzy rules need to be adjusted for different types of CFRP materials to achieve better performance. An automated classification of defect and non-defect areas in composites remains a challenging job which requires a considerable amount of research work to be carried out in future. In addition, the performance of the system can also be improved by studying the correlation between the damage mechanism and the data distribution, and by applying more sophisticated algorithms. Further, better performance can be achieved by constantly updating the knowledge and rules so that the systems can adapt to new kinds of problems.

Acknowledgement

The authors would like to thank the Center for Advanced Friction Studies (CAFS) directed by Dr. Peter Fillip, Southern Illinois University Carbondale, IL for partially supporting this project. The authors would also like to thank Mr. Matt Lane and Mr. Caleb McGee for providing their assistance in performing the testing.

References

[1] Sihn, S., R.Y. Kim, K. Kawabe, S.W. Tsai, "Experimental studies of thin-ply laminated composites," *Composites Science and Technology*, 64 (6) 996-1008, 2007.

[2] Graham, D., P. Maasa, G.B. Donaldsona and C. Carr. "Impact damage detection in carbon fiber composites using HTS SQUIDs and neural networks," *NDT&E International* 37, 565-570, 2004.

[3] Chu, T.C., A. Leyte, A. DiGregorio, S. Russell and J.L. Walker. "Micro-Cracking Detection in Laminated Composites," *Proc. of ASNT Spring Conference and 11th Annual Research Symposium*, Portland, OR, 2002.

[4] Im, K. H., D.K. Hsu, I. Y. Yang, "Inspection of Inhomogeneities in Carbon/Phenolic Matrix Composite Materials Using NDE Techniques." *Key Engineering Materials*. Vols. 270 – 273, pp 1799-1805, 2004 .

[5] Lee, J.H., S.W. Choi, K.S. Kim, J.H. Park, J.H. Byun, "Nondestructive Characterisation of Carbon/Carbon Brake Disks Using Ultrasonics", http://www.ndt.net/article/apcndt01/papers/1109/1109.htm, 2001.

[6] Ruosi, A., "Nondestructive detection of damage in carbon fiber composite," *Journal of Physical stat.*, Vol, 2(5), pp 1153-1155, March 2005.

[7] Bray, D.E., McBride, D, *Nondestructive testing techniques*, Wiley-Interscience Publication, John Wiley & Sons, Inc., NewYork, 1992.

[8] Liu N., Q.M. Zhu, C.Y. Wei, N.D. Dykes, Irving PE. "Impact damage detection in carbon fiber composites using neural network and acoustic emission." *Key Eng Mater*,167-168: 45-54, 1999.

[9] Zennouhi, R., Lh. Masmoudi, "IEEE - Image segmentation using hierarchical analysis of 2D-histograms - Application to medical images." *Proc. of Multimedia Computing and Systems International Conference*, (s): pp 480- 483, Ouarzazate, 2009.

[10] Gonzalez, R.C. and Woods, R.E. *Digital Image Processing*, 3rd ed., Prentice Hall, Upper Saddle River, NJ, 2008.

[11] Poudel, A., S. Li, T.C. Chu, D. Palmer, and R. Engelbart, "An Intelligent Systems Approach for Detecting Delamination Defects due to Impact Damage in CFRP Panel by Using Ultrasonic Testing", *Proc. of ASNT Fall Conference and 21st Annual Research Symposium*, Palm spring, CA, Oct 2011.

[12] Poudel, A., S. Li, T.C. Chu, D. Palmer, and R. Engelbart, "Neural-Fuzzy Approach in Detecting and Classifying Foreign Object Inclusions in CFRP Panel by Using Ultrasonic Testing", *Proc. of ASNT Fall Conference and 21st Annual Research Symposium*, Palm spring, CA, Oct 2011.

[13] Zadeh, L.A., Fuzzy sets. *Information and Control*, 8, pp. 338-353, 1965.

[14] Negnevitsky, M., *Artificial Intelligence*, Harlow, England: Addison-Wesley, 2005.

[15] Yen, J, and Laungari, R., *Fuzzy logic intelligence, control and information*, Prentice-Hall, Inc., New Jersy, USA, 1999.

Network Topic Detection Model Based on Text Reconstructions

Zhenfang Zhu
School of computer science and technology, Shandong University, 250100, Jinan, China
School of Information Science and Electric Engineering, Shandong Jiaotong University, 250357, Jinan,China
E-mail: zhuzhfyt@163.com

Peipei Wang
Shandong management University, 250357, Jinan, China
E-mail:wpp870213@163.com

Zhiping Jia
School of computer science and technology, Shandong University, 250100, Jinan, China
E-mail:jzp@sdu.edu.cn

Hairong Xiao, Guangyuan Zhang and Hao Liang
School of Information Science and Electric Engineering, Shandong Jiaotong University, 250357, Jinan,China
E-mail:{hairong.xiao, xdzhanggy}@163.com, lianghao3141@126.com

Keywords: topic detection and tracking, single pass algorithm, text reconstruction, network topic detection

Received: January 25, 2013

Single pass clustering algorithm is widely used in topic detection and tracking. It is a key part of network topic detection model. In the process of single pass algorithm, clustering results are not satisfactory, and the similarity matching would be reduced. Focusing on these two defects, this paper physically reconstructs web information into a volume, in which every document contains "theme area" and "details area". To improve single pass clustering algorithm, this paper uses "theme area" to detect topics and apply the whole document to distinguish subtopics, while central vector model is used to denote topics. Experimental results indicate that the model based on text reconstruction performs well in detecting network topics and distinguishing subtopics.

Povzetek: Razvita je nova metoda za zaznavanje teme omrežja na osnovi tekstovne analize.

1 Introduction

Network public opinion inclines to express the public attitudes towards social problems of the world, which are described as hot topics in Internet. The netizens focus on the hot topics by reading, releasing and quoting information of kinds forms, such as web news, BBS posts, blog articles and so on. Therefore, detecting network hot topic quickly and efficiently is the key to grasp the rules of public sentiment changing. Topic detection and tracking (TDT) technology is to provide a core technology to identify a new topic in the web information and group stories on the same topic from huge volume of information that arrives daily. TDT automatically detects hot information of public opinions, is a kind of critical technology in the field of natural language processing and information retrieval.

2 Related work

The TDT technology is intended to explore techniques for detecting the appearance of new topics and tracking the reappearance and evolution of them, and is widely used to detect network hot topics. Researchers at home and abroad have done lots of researches on network topics.

Earlier researchers focus on selecting and combining clustering algorithms. Ron Panka and James Allan [1] use a single pass clustering algorithm and a novel thresholding model that incorporates the properties of events as a major component. Ref. [2] adopts Group Average Clustering (GAC) clustering techniques to detect a novel event. The task of TDT is to automatically detect novel events from a temporal-ordered stream of news stories. In addition, Ron Papka [3] makes comparisons among many clustering algorithms, and tries to solve problems of OTD by putting clustering algorithms together reasonably.

In above researches, all the stories and there related topic are at one level, and one of stories belongs to one topic. However, the whole topic may pivot on multiple points and one story also can cover more than one topics. In order to express these characteristics, hierarchical topic detection (HTD) is put forward in TDT2004. Participants in this task are no longer required to submit flat cluster partitions, but to generate a directed a cyclic graph (DAG). Each graph has a root node, which is an

ancestor of all other clusters. And each node represents a topic at a specific granularity, which can overlap or subsume each other. Hierarchical Agglomerative Clustering (HAC) is an effective method to generate hierarchical structures. Researchers, such as Trieschnigg [6], present a scalable architecture for HTD and compare several alternative choices for agglomerative clustering and DAG optimization in order to minimize the HTD cost metric.

Civil researches pay attention to topic's hierarchy and time sequence, combine natural language processing techniques to detect network topics. Ref. [7] divides all data into groups and clusters in each group to produce micro-clusters, and then groups all micro-clusters to final topics.

This paper proposes the thought of text reconstruction and applies it to improve single pass clustering algorithm. The method both increases processing speed in single pass clustering and considers hierarchical structures of topics.

3 Topic detection model

3.1 Topic/Story Model

Every story di in a topic is expressed as di=(item1,wi1,...itemj,wij,...,itemm,wim); where wij is computed by TF-IDF. In this paper, we considers the term's position in the story when we compute term frequency, as formula (1):

$$w_{ij} = \frac{(t_{ij} \times \log(N/n_i + 0.01))}{\sum_{k=1}^{m} [tf_{ik} \times \log(N/n_k + 0.01)]^2} \quad (1)$$

Where $tf_{ij} = 5 \times tf_{ij}(title) + tf_{ij}(text)$, $tf_{ij}(title)$ is the frequency of term termj occurrence in the title of the story, m is the number of terms, N is the number of stories in the topic, ni is the number of stories which contains term ti.

Many related stories make up a topic, this paper adopt a central vector to build topic model. The central vector is described as (item1,w1,...,itemj,wj,...,itemm,wm), the weight of the term is the average of all the stories, computed by formular (2):

$$w_j(t,T) = \sum_{d_i \in ST} w_{ij} \Big/ StoryNum(t,T) \quad (2)$$

Where wj(t,T) is the weight of term itemj in the t statistical time, StoryNum(t,T) is the number of all stories in topic T at the time.

The new coming story S is expressed as (item1,ws1,..., itemj,wsi,..., itemm,wsm) by vector space model, where itemj is the term, wsi is the weight of term si in story S.

The paper adopts classical cosine similarity to compute the similarity between story S and topic T.

3.2 Topic Detection Algorithm

Single pass incremental clustering algorithm is widely used in TDT; it has simple thought and faster processing. The algorithm sequentially process documents using a pre-specified order. The current document is compared to all existing topics, and it is merged with the most similar topics if the similarity exceeds a certain threshold. Single pass clustering is discussed in detail as Table 1.

Table 1: Single pass algorithm description.

Input: the new stories
Output: some clusters
Process:
Step1 Read in a new story S;
Step2 Compute similarity $Sim(S,T_i)$ between S and each cluster existing at its processing time.
Step3 The story S is assigned to the cluster T, when $Sim(S,T) = \arg\max_i Sim(S,T_i)$ and $Sim(S,T) > \theta$ （θ is a threshold）;
Step4 If the story S fails a certain similarity test it becomes a new cluster T';
Step5 If story S is not the last, go to Step1.

4 Text reconstruction–based hierarchy topic detection model

TDT can detect network hot information which reflects public opinion, and it is the basic work of public opinion analysis. This paper focuses on improving the results of single pass clustering algorithm in TDT and introduces the thought of text reconstruction. It separates information collected from internet into "theme area" and "details area". In addition, this paper adopts central vector to represent a topic, in order to increase processing speed when a story matches each topic.

4.1 Topic Structure

Events change continuously, a topic which is defined as an event or activity, along with all directly related events and activities is also in constant change. A topic can be described as a congregation which contains web news, BBS posts, blog articles and so on. They change along with public attentions and the course of events. Take Pakistani airliner crash for example, media and people focus on "air accident", "source of damage", "care-taking arrangement".

4.2 Text Reconstruction

4.2.1 Short Text Reconstruction.

Interactive BBS forum provides a platform for people to express their sentiment and opinions, which is an area with a high incidence of public opinions. The intercommunion in BBS forum by releasing and replying posts, public sentiment and opinions is merged in these posts. Therefore, how to organize posts effectively is the key to detect network hot topic.

Ref. [8] defines "one clue" as title, main post and all responding posts. They consider that main post and responding posts revolve around the title. They introduce the idea of reconstruction to solve the problem of sparseness of the short text and get a better clustering performance.

Borrowing ideas from the above references, this paper introduces "text reconstruction". It gets typical features of a topic together, namely "theme area" and the remainder, namely "details area".

Posts in "one clue" of BBS forum usually contains plentiful information, such as title, author, main post, posts in responding to original and so on. Therefore, this paper puts title and main post together to form "theme area". "Details area" are made up of random selected responding posts. Other short texts, such as instant message, commenting on blogs, and online chat log, can be processed by the similar method.

4.2.2 Web news reconstruction.

News title contains plentiful categories information. It is the summary of the web news. The title has simple syntax and structure, and the accuracy is higher when it is used to classify web news. Ref. [9] adopts 2003 text corpus of People's Daily to test, up to 93.7% titles contains category information.

Topic is the support point of title structure. Under the same topic, every report's titles are the same or similar. Title information has significant ability of topic distinguish in TDT. But with the events' development, the distance between follow-up report's title and initial event's title may become farther and farther. The longer the time from the initial event is, the greater possibility of title drifting is. Therefore, the accuracy of utilizing title similarity to identify topic will certainly decline.

The first paragraph of news webpage outlines the basic information which includes time, place, events, characters and so on. And also numerous of category information was included in this paragraph. According to the idea of text reconstruction, combining the news webpage's first paragraph and title information which effectively gathers the typical features of topic.

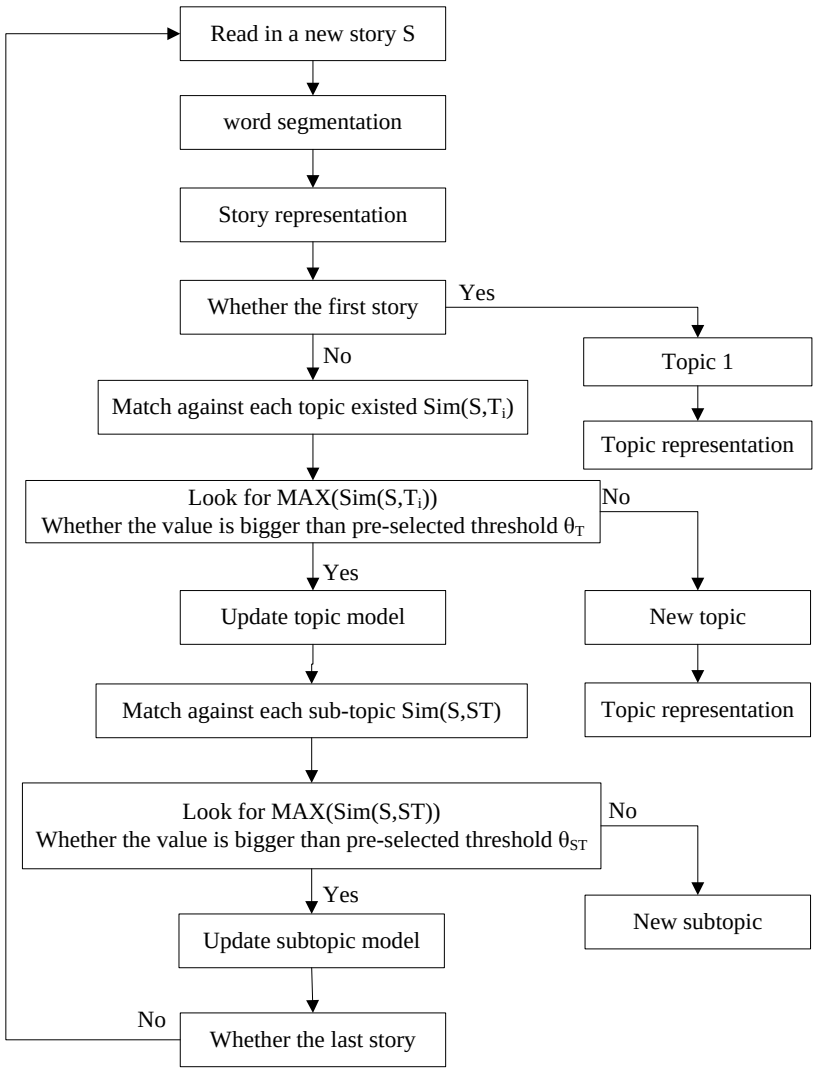


Figure 1: The improved Single pass clustering algorithm.

In TDT system, “theme area” is composed by news webpage’s first paragraph and title information. Considering the effect of news commentary in public opinion analysis, this paper uses the remaining paragraphs’ descriptive information and news comments to compose “details area”.

4.3 Hierarchy topic detection algorithm

Single pass clustering algorithm can detect network hot topics, which satisfies public opinion analysis system basically, but here lists several shortcomings:

(1) The effective of the algorithm is dependent on the order in which documents are processed. This is not a problem when the documents are temporally ordered, because the order is fixed.

(2) The common strategies for combining similarity values are known as single-link, complete-link, and average-link clustering. All comparison strategies need to compare with all documents in each existed topics. If the number of documents in a topic is on a large scale, processing speed reduces.

(3) Single pass clustering algorithm can group documents with similar content. It collects stories which belong to one topic, but it ignores topical hierarchical structure.

In order to overcome the above shortcomings, this paper uses the “theme area” to gather similar stories which belong to one topic, and adopts “details area” to divide subtopics. It also adopts central vector model to denote topics to increase processing speed, then the improved single pass clustering algorithm is described as Fig. 1.

5 Experiments and Analysis

5.1 Experiments corpus

The paper takes web news to validate effectiveness of text reconstruction–based network topic detection model. We collect thematic information and much-talked-about topic web news from sina.com.cn, 163.com.cn, sohu.com, ifeng.com, people.com.cn, xinhuanet.com. We select and clean up eight topic information, such as Tang Jun, "fake door" (TJFD), Dalian oil pipeline explosion (DOPE), Hubei officials wife incident (HOWI), Qian Wei-chang's death (QWCD), 10-year goal of developing the western region (GDWR), Luanchuan Bridge collapse incident (LBOI), Nanjing plant Explosion (NJPE), Zijin Mining Pollution (ZJMP), Table 2 gives description in detail.

To verify the effectiveness of text reconstruction in TDT, we construct two data sets: data set one contains the above eight topics, and each story is original documents; data set two still contains eight topics, but each story is reconstructed as “theme area” and “details area”.

5.2 Evaluation indexes

In the TDT setting, we chose the miss rate Pmiss and false alarm rate Pfa to measure the effectiveness of topic detection model based on text reconstruction. Pmiss is the probability that a model produces a miss, and Pfa is the probability that a model produces a false alarm. The method for calculating the measures are summarized below using the following table3:

Where the retrieved texts in the table are those that have been classified by the system as positive instances of a topic, and the relevant texts are those that have been manually judged relevant to a topic. The measures used in this paper can be computed from the table as follows:

$$P_{miss} = C/(A+C) \tag{3}$$

$$p_{fa} = B/(B+D) \tag{4}$$

A cost function (Cdet) is usually used to analyze detection effectiveness. The general form of the TDT cost function is as formular (5):

$$C_{\det} = C_{miss} * P_{miss} * P_{t\arg et} + C_{fa} * P_{fa} * P_{non-t\arg et} \tag{5}$$

Where Cdet is a cost function, Cmiss is lost cost, Cfa is false alarm cost, Ptarget is the prior probability that a document is relevant to a topic. In our expriment, we

Table 2: The experiment corpus.

Number	1	2	3	4	5	6	7	8
Topic	TJFD	DOPE	HOWI	QWCD	GDWR	LBOI	NJPE	ZJMP
Collected stories	135	119	14	42	30	139	54	94
Selected stories	100	100	14	42	30	100	54	94
Sub-topics	5	4	3	2	2	4	3	4

Table 3: The related parameter.

	Relevant	Non-relevant
Retrieved	A	B
Not Retrieved	C	D

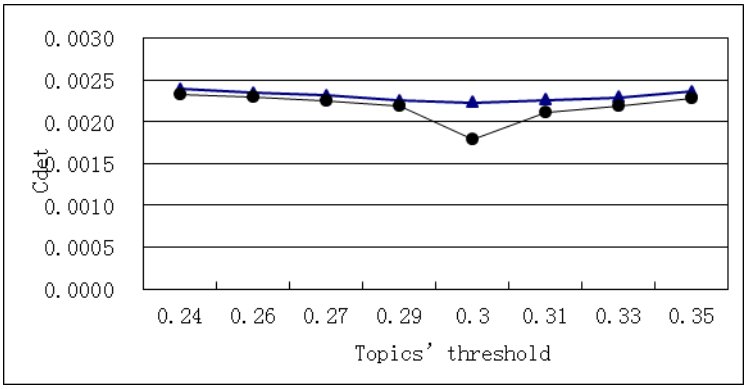


Figure2: The comparison in C_{det}.

preset Cmiss= Cfa=1.0, Ptarget=0.2.

For one topic, we use formular (5) to compute Cdet, but for all topics, we adopt (Cdet)norm, which is defined as the weighted Cdet average value of all topics, the weigh is that the number of all stories in a topic divided by all stories in all topics [10].

5.3 Experiments and results analysis

5.3.1 Topic detection and interpretation of results.The experiment is to check the ability of improved single pass clustering algorithm in detecting topics. A reasonable threshold θ_T is the key to identify topic correctly. Stories that have similarity exceeding the threshold are classified into one topic. If the θ_T is too big, the granularity of a topic is too large, in contrast, if the θ_T is too small, there will be too many topics. So it is more difficult to determine a good score that can be used as a threshold.

We select 10 stories randomly from each topic, reconstruct each story. We select cosine similarity between every story and the topic in which it belongs to. According to the values of similarities, we can conclude that:

(1)In data set one, the similarities between story and its topic are above 0.30, and the similarities with the other topics are under the 0.24.

(2) In data set two, the similarities between story and its topic are above 0.35, and the similarities with the other topics are under the 0.28.

Therefore, the topics' similarity threshold θ_T should range from 0.24 to 0.35.

We use the above corpus to compare the original topic detection model which use single pass clustering algorithm and the text reconstruction–based hierarchy topic detection model which improve the single pass clustering algorithm. We adopt 10-fork method and adopt average Cdet of all experiments to evaluate performance. The Cdet changes along with the different threshold. Fig.2 gives the detail description.

Fig.2 tells that, given the topics' similarity threshold θ_T, compared with original topic detection model, the topic detection model based on text reconstruction has smaller Cdet. It shows that the improved topic detection model performs better in identifying and tracking topics. The average cost function Cdet fluctuates along with the different similarity threshold θ_T. θ_T ranges from 0.24 to 0.30, Cdet keep in decreasing, but Cdet is in upswing when θ_T is more than 0.30. So θ_T =0.3 is reasonable in experiments.

5.3.2 Topic Structure Identification and Results Analysis.The purpose of this experiment is to check the ability of improved single pass clustering algorithm in detecting topics. We determine subtopic threshold θ_{ST} in a similar way as θ_T, it ranges from 0.4 to 0.6, and θ_{ST} =0.48 is the best in the experiment.

Take topic "DOPE" for an example, "DOPE" covers few subtopics, such as "Event overview (EO)", "Accident cause and responsibility (ACR)", "Deal with pollution (DP)", "Accident impact and compensation (AIC)" and so on. We collect five subtopics and its stories of "DOPE" artificially.

In the experiment, we present $\theta_T = 0.3$, $\theta_{ST} \in [0.4,0.6]$, and adopt text reconstruction–based hierarchy topic detection model to test. The model identifies the topic "DOPE" correctly, and separate subtopics at a certain extent. It detects five categories, which is consistent with the results of artificial markers in principle. We count numbers of stories in each subtopic collected both by hand and by the model, Fig.3 gives the details.

Fig.3 shows that the results of topic detection model based on text reconstruction are similar with the results of artificial markers. It indicates that the improved model is able to identify topic structure to some extent.

6 Conclusion

The basic work of analysing public opinion is to detect hot topics on internet and find out what people concerns, what people meet with, and what people dissatisfy. TDT groups stories to a topic automatically from huge volume of updating information in technology. This paper proposes a network topic detection model based on text reconstruction and improves the usual detection algorithm of the model. Text reconstruction makes every document into two parts: "theme area" and "details area".

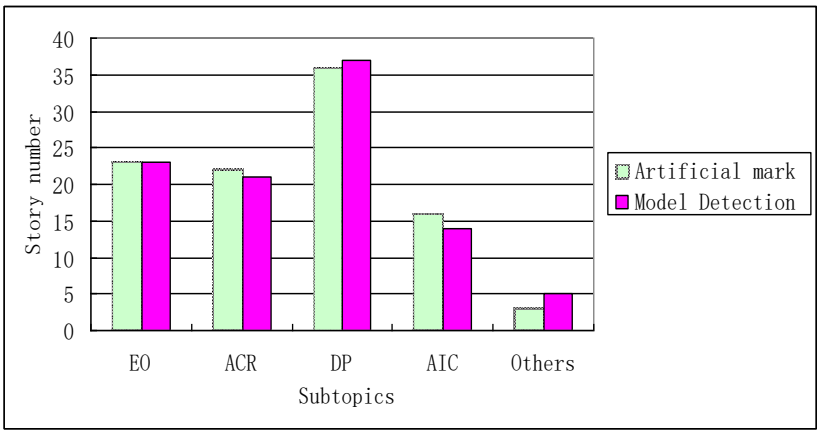


Figure 3: Subtopic Detection.

We use theme area to detect topics and apply the whole document to distinguish subtopics. Experimental results indicate that the model performs well in detecting network topics and distinguishing subtopics.

Text reconstruction–based network topic detection model shows the hierarchical structure of a topic to a certain extent, but increases the complexity of computing. In the next study, we will reform similarity calculation to improve the computational efficiency.

7 Acknowledgement

This work is supported by Shangdong Province Young and Middle-Aged Scientists Research Awards Fund (BS2013DX033), National Nature Science Foundation of China (61373148), Nature Science Foundation of Shandong Province (ZR2012FM038).

References

[1] Ron Papka and James Allan (1998). On-Line New Event Detection using Single Pass Clustering. UMASS Computer Science Technical Report UM-CS-1998-021, Amherst.

[2] Y Yang, T Pierce, J Carbonell (1998). A study on Retrospective and On-Line Event detection. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. CMU, USA, ACM, pp.28-36.

[3] Papka R (1999). On-line New Event Detection, Clustering and Tracking. Amherst: Department of Computer Science, UMASS.

[4] The 2004 Topic Detection and Tracking (2004) Task Definition and Evaluation Plan. version 1.2, http:/ /www.nist.gov.

[5] HONG Yu, ZHANG Yu, LIU Ting etc (2007). Topic Detection and Tracking Review. Journal of Chinese in Information Processing, 21(6), pp. 71-87.

[6] D Trieschnigg and W Kraaij (2004). TNO Hierarchical topic detection report at TDT 2004. The 7th Topic Detection and Tracking Conference.

[7] LUO Wei-hua, YUMan-quan, XU Hong-bo etc (2006). The Study of Topic Detection Based on Algorithm of Division and Multi-level Clustering with Multi-strategy Optimization. Journal of Chinese in Information Processing, 20 (1), pp. 29-36.

[8] SUN Cheng-jie, ZHU Wen-huan, LIN Lei, etc (2009). Research on BBS Short Text Clustering. CCIR 2009, pp.470-479.

[9] MIAO Jian-ming, ZHANG Quan, ZHAO Jin-fang (2008). Chinese Automatic Text Categorization Based on Article Title Information. Computer Engineering, 34(20), pp.13-14.

[10] ZHANG Xiao-yan, WANG Ting, CHEN Huowang (2007). Story Link Detection Based On Multi-Vector Model with Support Vector Machine. Chinese Computing Technologies and Related Linguistic Issues---Proceedings of the 7th International Conference on Chinese Computing, pp.390-95.

Informal Multilingual Multi-domain Sentiment Analysis

Tadej Štajner[1,2], Inna Novalija[1] and Dunja Mladenić[1,2]
[1]Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773900
E-mail: {firstname.secondname}@ijs.si

[2]Jožef Stefan International Postgraduate School
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773100

Keywords: sentiment analysis, social media, news sentiment, opinion mining

Received: March 6, 2013

This paper addresses the problem of sentiment analysis in an informal setting in multiple domains and in two languages. We explore the influence of using background knowledge in the form of different sentiment lexicons, as well as the influence of various lexical surface features. We evaluate several different feature set combination strategies. We show that the improvement resulting from using a two-layer meta-model over the bag-of-words, sentiment lexicons and surface features is most notable on social media datasets in both English and Spanish. For English, we are also able to demonstrate improvement on the news domain using sentiment lexicons as well as a large improvement on the social media domain. We also demonstrate that domain-specific lexicons bring comparable performance to general-purpose lexicons.

Povzetek: Ta članek obravnava problem analize naklonjenosti v neformalnem besedilu v različnih domenah in v dveh različnih jezikih.

1 Introduction

Sentiment analysis is a natural language processing task which aims to predict the polarity (usually denoted as positive, negative or neutral) of users publishing sentiment data, in which they express their opinions. The task is traditionally tackled as a classification problem using supervised machine learning techniques. However, this approach requires additional effort in manual labelling of examples and often has difficulties in transferring to other domains.

One way to ameliorate this problem is to construct a lexicon of sentiment-bearing words, constructed from a wide variety of domains. While some sentiment-bearing cues are contextual, having different polarities in different contexts, the majority of words have unambiguous polarity. While this is a compromise, research shows that lexicon-based approaches can be an adequate solution if no training data is available. In practice, sentiment dictionaries or lexicons are lexical resources, which contain word associations with particular sentiment scores. Dictionaries are frequently used for sentiment analysis, since they allow in a fast and effective way to detect an opinion represented in text. While there exists a number of sentiment lexicons in English [1] [2] , the representation of sentiment resources in other languages is not as developed. The first problem this paper focuses on is integrating external knowledge in the form of general-purpose sentiment lexicons.

The second problem this paper focuses on is detecting sentiment in specific domains, such as social media. Besides being domain-specific, it can also be grammatically less correct and contain other properties, such as mentions of other people hash-tags, smileys and URL, as opposed to traditional movie and product review datasets.

This paper explores various combinations of methods that can be used to incorporate out-of-domain training data, combined with lexicons in order to train a domain-specific sentiment classifier.

2 Related work

Sentiment classification is an important part of our information gathering behaviour, giving us the answer to what other people think about a particular topic. It is also one of the natural language processing tasks which is well suited for machine learning, since it can be represented as a three-class classification problem, classifying every example into either positive, neutral, or negative. Earlier work applied sentiment classification to movie reviews [10] , training a model for predicting whether a particular review rates a movie positively or negatively. While in the review domain all examples are inherently either positive or negative, other domains may also deal with non-subjective content which does not carry any sentiment. Furthermore, separating subjective

from objective examples has proven to be an even more difficult problem than separating positive from negative examples [13] . Another difficult problem in this area is dealing with different topics and domains: models, trained on a particular domain do not always transfer well onto other domains. While the standard approach is to use one of widely used classification algorithms such as multinomial Naïve Bayes or SVM, explicit knowledge transfer approaches have been proven to improve performance in these scenarios, such as using sentiment lexicons [1] or modifying the learning algorithm to incorporate background knowledge [9] . Some challenges are also domain-specific. For instance, while a lot of sentiment is being expressed in social media, the language is often very informal, affecting the performance by increasing the sparsity of the feature space. On the other hand, the patterns arising in informal communication, such as misspellings and emoticons, can be themselves used as signals [13] . It has also been shown that within social media, using different document sources, such as blogs, microblogs and reviews, can improve performance compared to using a single source. [12] .

This paper also explores the integration of multiple data representations for a specific task of text classification. This sort of approach was also successful in the case where several combination strategies were used for the task of authorship detection [14] , such as feature set concatenation or majority voting of classifiers, trained on only subsets of features. While these are known general strategies, a lot of aspects of selecting sensible feature subsets are very domain specific.

3 Sentiment Lexicons

SentiWordNet [1] is the most known English-language sentiment dictionary, in which each WordNet [3] synset is represented with three numerical scores – objective *Obj(s)*, positive *Pos(s)* and negative *Neg(s)*. However, SentiWordNet does not account for domain specificity of the input textual resources. In addition to addressing English language, this paper also discusses applications of sentiment dictionaries in Spanish. For this purpose, we have used the sentiment dictionaries published by Perez-Rosas et al. [6] .

Expressing sentiment and opinion varies for different domains and document types. In such way, sentiments carried in the news are not equivalent to the sentiments from the Twitter comments. For instance, the word “turtle” is neutral in a zoological text, but in informal Twitter comment “connection slow as a turtle”, “turtle” has negative sentiment. This paper also evaluates a method for construction of dictionaries as domain specific lexical resources, which contain words, part of speech tags and the relevant sentiment scores. We have chosen the topic of telecommunication services within social media as the domain of primary interest, and the corpus, used for dictionaries development, was composed out of Twitter comments referring to services of telecommunication companies. We have started with a number of positive and negative seeds for different part-of-speech words (adjectives, nouns, verbs). These sentiment dictionaries are built in English and Spanish languages. As discussed in [3] , there are a number of approaches to develop the sentiment dictionary. In our research on developing sentiment dictionaries we were following the work of Bizau et al. [4], where, the authors suggested a 4-step methodology for creating a domain specific sentiment lexicon. We have modified the methodology in order to generalize to other languages and provide sentiments for different parts of speech.

We have created dictionaries not only in English, but also in Spanish. Our dictionaries were built not only for adjectives as done in [4], but also for nous and verbs. For the English dictionary, we have additionally provided several extra features, such as the number of positive links and number of negative links for a particular word. The English sentiment dictionary for the Telecommunication domain is composed out of around 2000 adjectives, 1700 verbs and 8000 nouns, while the Spanish counterpart contains around 650 adjectives, 2000 verbs and 4100 nouns.

4 Feature construction

We have used different feature sources to represent individual opinion data points. In news and review datasets, every data point is a sentence, while in social media datasets, every data point is a single microblog post. We preprocess the textual contents by replacing URLs, numerical expressions and the names of opinions’ targets with respective placeholders. We then tokenize this text, lower-casing and normalizing characters onto an ASCII representation, filtering for stopwords and weigh the terms using TF-IDF weights. The words were stemmed using the Snowball stemmer for English and Spanish [17] . The punctuation is preserved.

To accommodate social media, we have also used other text-derived features that can carry sentiment signal in informal settings, as commonly done in representation of social media text:

- count of fully capitalized words
- count of question-indicating words
- count of words that start with a capital letter
- count of repeated exclamation marks
- count of repeated same vowel
- count of repeated same character
- proportion of capital letters
- proportion of vowels
- count of negation words
- count of contrast words
- count of positive emoticons
- count of negative emoticons
- count of punctuation
- count of profanity words[1]

[1] Obtained from http://svn.navi.cx/misc/abandoned/opencombat/misc/multilingualSwearList.txt

We use lexicons in the form of features, where every word has assigned one or more scores. For instance, our dictionaries, described in Section 3, as well as SenticNet, provide a single real value in the range from -1 to 1, representing the scale from negative to positive. For these lexicons, we generate the sum of sentiment scores and the sum of absolute values of sentiment scores for every part of speech tag, as well as in total. SentiWordNet scores are represented as a triple of positive, negative and objective scores, having a total sum of 1.0. We have used a similar feature construction process as in [7] :

- Sum of all positive sentiments of all words.
- Sum of all negative sentiment of all words.
- Total objective sentiment of all words (where *obj = 1.0 - (pos + neg)*) score
- Ratio of total positive to negative scores for all words

Besides providing total sums, we also generate these features for nouns, verbs, adjectives and adverbs separately.

For Spanish, we have used the UNT sentiment lexicon [6] . Since each entry is labelled only as either positive or negative, we use the count of detected positive words and count of detected negative words as features.

5 Models

The data is composed of three main modalities: bag-of-words features, lexicon features, and surface features.. In order to take differing distributions, dimensionality and sparsity properties into account, we use two different approaches: either concatenating the features into a single features space, or using different models for each set of features. While this situation has been solved by extending the Naïve Bayes classifier with pooling multinomials [9] , we chose to implement it with a two-step model. We experiment with different feature combination approaches that are better suited for integration of background knowledge and other learning algorithms.

5.1 Feature combination

We therefore compare three feature combination approaches and a baseline, illustrated in Figures 1 through 4. The concatenating model simply stacks all feature spaces together and performs learning on the joint feature space. While this approach is simple, it is sensitive to different feature distributions. Therefore, we pre-emptively scale the features, so every feature has a standard deviation of 1.0. We don't standardize the mean, since the features themselves may be sparse, and complete standardization would densify the data. The concatenation approach from Figure 1 is considered as the baseline.

The second approach, as shown in Figure 2, is using a separate learning model for the bag-of-words feature set, and feeding the output of that model as features into the final classifier, together with the less sparse lexicon and surface features, in order to 'compress' the bag-of-words signal.

The third approach is related to the well-known attribute bagging [16] meta-learning strategy, with the crucial difference that the feature sub-sets are already defined in advance via domain knowledge.

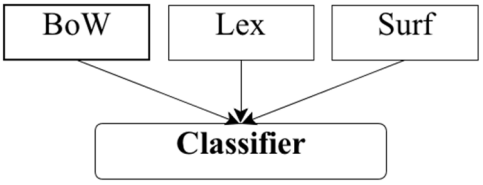


Figure 1: Feature concatenation diagram.

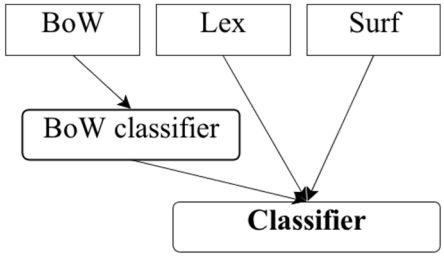


Figure 2: Separate bag-of-words model, denoted as "Words and features"

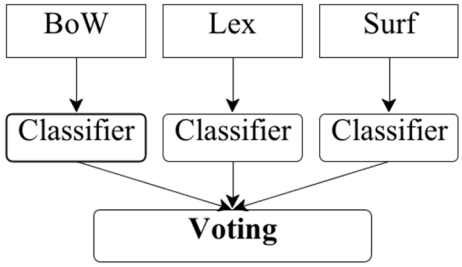


Figure 3: Separate model for every feature set, aggregated by voting.

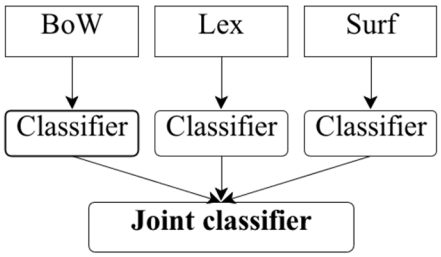


Figure 4: Meta-classifier, using class probabilities from the inner classifier predictions as its features.

The fourth approach extends the voting by employing a separate classifier model that operates on the output of the output probabilities of the inner models, in order to minimize bias of individual feature sets.

We experiment by varying the training algorithm used: For the approaches using multiple models, we use the same algorithm for all the models.

All in all, we evaluate four feature set combination strategies, corresponding to Figures 1-4:

- Concatenation (**Concat**)

- Two-layer words and features (**W+F**)
- Voting model (**Voting**)
- Meta-classifier (**Meta**)

6 Experiments

Furthermore, we focus our experiment onto performance on our target datasets. We use the following datasets:

- Pang & Lee review dataset (PangLee), English [10] , consisting of movie reviews, gathered from IMDB.
- JRC news dataset (JRC-en), English [11] , consisting of statements from news articles on the topic of global politics.
- JRC news dataset, translated to Spanish using Microsoft Translator (JRC-es)
- RenderEN, English. 134 Twitter posts about a telecommunications provider (48 positive, 84 negative)
- RenderES, Spanish, 891 Twitter posts about a telecommunications provider (388 positive, 445 negative, 58 objective)

Besides our lexicons introduced in section 3 (denoted "RenderLex" and "RenderLexLinks"), we also evaluate performance of using the Spanish lexicons from Perez-Rosas et al [6] (denoted FullUNT and MedUNT for the full and medium variant respectively), as well as SenticNet [8] and SentiWordNet[1] for English. The label "Lex" indicates usage of all lexicons. Our key indicators are performance metrics on RenderEN and RenderES, as they represent our use case. We perform experimental evaluation for all of these datasets on various combinations of classifiers and features construction schemes. The experiments cover various learning algorithms, as well as different modelling pipelines. We explore various combinations of feature sets: surface, bag-of-words, lexicons, as well as performance contributions of individual lexicons.

The first evaluation deals with observing the applicability of various sentiment lexicons, as described in Section 3. First, we evaluate the lexicons in isolation, followed by a combination of lexicons together with surface features. We train a L1-regularized logistic regression classifier on lexicon features. The performance is measures using averaged F_1-score [18] in a 10-fold cross-validation setting.

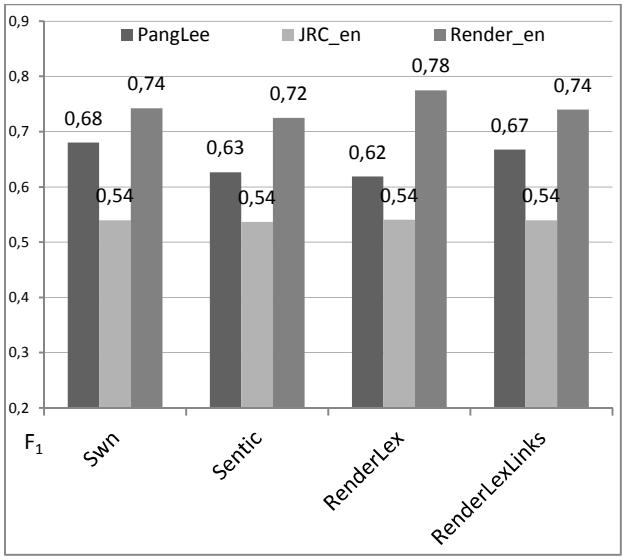


Figure 5: Sentiment F_1 scores with various sentiment lexicons for English.

Figure 5 shows the results, obtained performing sentiment classification on the basis of sentiment lexicon features alone. We observe that performance across the news dataset is constant, since the expression of sentiment in news doesn't directly correspond to sentiment meaning of individual words, but more to the domain-specific political statements. For the social media dataset, we observe improved performance when using a telecommunications domain-specific lexicon, compared to using a general domain sentiment lexicon.

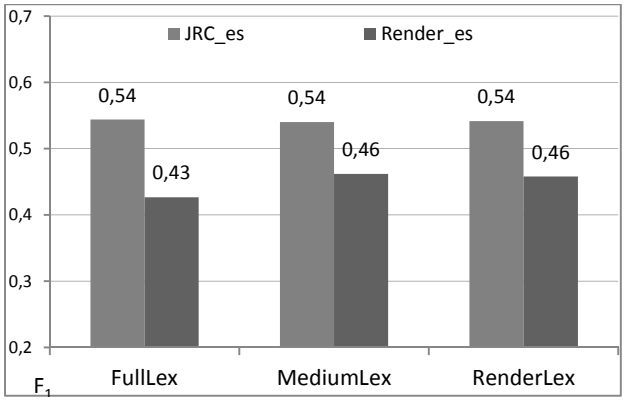


Figure 6: Sentiment F_1 scores with various sentiment lexicons for Spanish.

While Figure 6 confirms the same behaviour for news, the benefit of using lexicons is much lower in Spanish social media content. Given these results, we establish that a custom-built lexicon can give better results than a general purpose one. To continue, we evaluate various feature combination techniques on different learning algorithms.

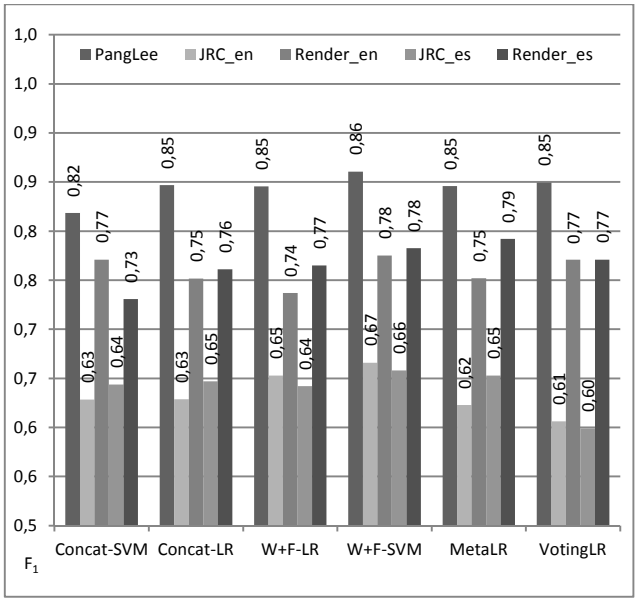


Figure 7: F_1 scores with various feature combination approaches across both languages and two learning approaches.

Figure 7 displays the performance across different feature combination approaches across all datasets. Looking into individual models, we observe that the W+F model, having the bag-of-words feature set on a separate layer, consistently works best for the purpose of combining all the three feature sets and masking the differences in the distribution of their features. While the W+F model consistently outperforms concatenation by a small but statistically significant margin, the Voting or Meta-classifier model only outperform concatenation on some occasions, and perform worse on the news dataset in both languages. We report the results on scenarios where LR was used as the learning algorithm on the Meta and Voting models due to the fact that they obtain comparable performance.

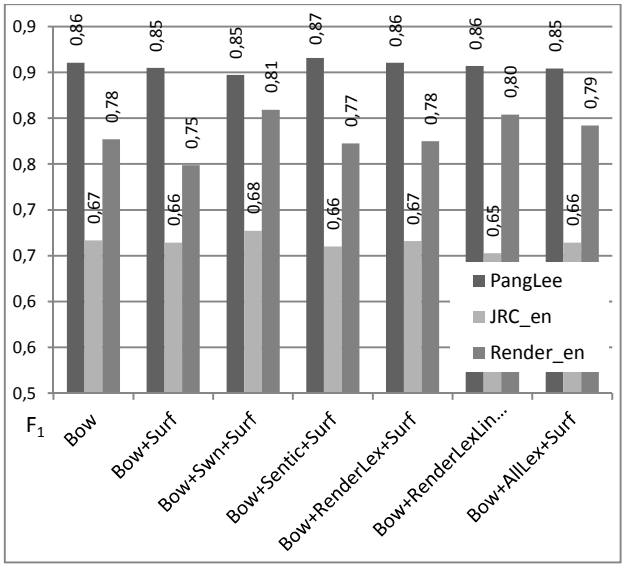


Figure 8: Using various feature sets on English datasets, using W+F-SVM.

Figure 8 shows the results on English reviews, news, and social media. On reviews, none of the additions significantly beat the bag-of-words baselines on reviews. On news, while adding SentiWordNet marginally improves the performance from 0.67 to 0.68, surface features don't give any improvement, mostly due to the formal language used in reporting, which leads to the fact that the text is written without informal cues. On other hand, results on the *Render_en* social media dataset, demonstrate the performance improvements in combining all three feature sets in a two-layer model. The best performing model is able to obtain a F_1 score of 0.87. While the dataset is small, this demonstrates the feasibility of using generalized external knowledge and surface features in a social media setting, especially with insufficient training data.

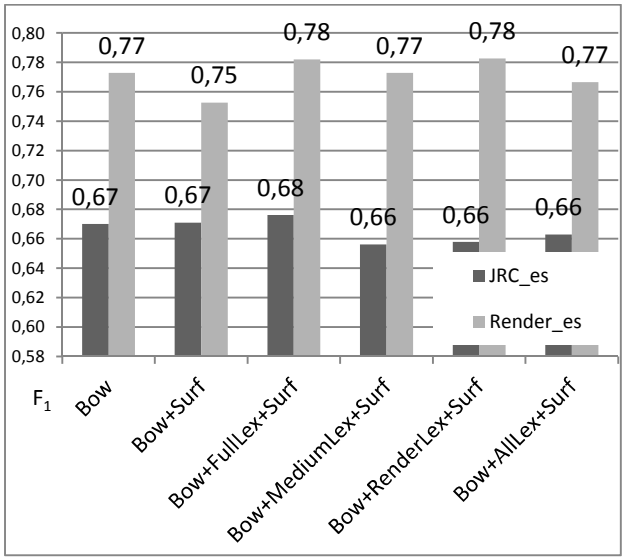


Figure 9: Sentiment F_1 scores on Spanish datasets, using W+F-SVM.

Figure 9 shows the results on both Spanish datasets when combining different feature sets in a W+F setting and a SVM model. We observe that on the news dataset, adding the Full UNT Lexicon slightly improves the F_1 score, while surface features alone don't give any improvement. On Render-ES, the variant combining all additions and running on a two-layer SVM model improves over the bag-of-words model by a small margin, resulting in an F_1 score of 0.78. Looking at usage of various lexicons alone, it shows that the lexicons themselves only slightly improve over the surface features. In many cases, the difference is not significant, although we observe that the domain specific lexicon RenLex does not improve over a general domain lexicon neither in news nor in social media.

7 Model analysis

In order to better understand the obtained models, we visualized the decision trees as hierarchical diagrams, produced in the output of CLUS [15] . To ensure better interpretability of the models, we have constructed them in the following way: using a 10% pruning and 10%

testing dataset, we have used the F-test stopping criterion for splitting nodes. A node was split only when the test indicated a significant reduction of variance inside the subsets at the significance level of 0.10. The tree was then pruned with reduced error pruning using the validation dataset.

For clarity, we have only attempted to interpret the models using the lexicon and surface features. Bag-of-words features were omitted, since they resulted in deep one-branch nodes, which are difficult to visualize.

```
full_unt_pos > 0.0
+--yes: [OBJ]
+--no:  renderlex_noun_sum_neg > 0.0
  +--yes: [NEG]
  +--no:  numcaps > 0.0386
    +--yes: renderlex_adjective_abs > 0.4069
    | +--yes: h1w5 > 0.0312
    | | +--yes: [POS]
    | | +--no:  [OBJ]
    | +--no:  renderlex_all_sum > 3.866
    |   +--yes: [OBJ]
    |   +--no:  h1w5 > 0.0833
    |     +--yes: [OBJ]
    |     +--no:  full_unt_neg > 0.0
    |       +--yes: [OBJ]
    |       +--no:  repeat_vowel > 0.0244
    |    +--yes: [POS]
    |    +--no:  numvowel > 0.3429
    |      +--yes: [OBJ]
    |      +--no:  renderlex_all_abs > 2.1249
    |        +--yes: renderlex_all_sum > 2.7152
    |        | +--yes: [OBJ]
    |        | +--no:  [NEG]
    |        +--no:  [OBJ]
    +--no:  [OBJ]
```

Figure 10. Model constructed from training on Spanish news data (JRC-ES).

Figure 10 shows the tree, constructed by training the lexicon and surface feature representation of the news dataset. It shows that lexicon indicators are closest to the root, covering the most examples. The negative sum of noun scores has proven to be a good indicator for negative sentiment, suggesting that nouns are the more sentiment-bearing words in the news domain. Also, capitalization plays an important role in the model. While it is most likely a proxy for appearance of named entities, it shows that subjective statements tend to have more capitalized phrases. Also, the presence of questions (denoted as *h1w5*) tended to indicate a positive sentiment.

```
numvowel > 0.3246
+--yes: numcaps > 0.8462
| +--yes: [POS]
| +--no:  renderlex_all_sum_neg > 0.2682
|   +--yes: [POS]
|   +--no:  numvowel > 0.3566
|     +--yes: [NEG]
|     +--no:  renderlex_adverb_sum_neg > 0.4899
|       +--yes: [POS]
|       +--no:  repeat_letter > 0.0588
|    +--yes: [POS]
|    +--no:  [NEG]
+--no:  renderlex_adverb_abs > 0.52
  +--yes: renderlex_adverb_abs > 0.5964
  | +--yes: [POS]
  | +--no:  [NEG]
```

```
      +--no:  negation > 0.0
        +--yes: repeat_letter > 0.0357
        | +--yes: [NEG]
        | +--no:  [POS]
        +--no:  full_unt_neg > 0.0
          +--yes: [NEG]
          +--no:  length > 27.0
      +--yes: renderlex_noun_abs > 4.4911
      | +--yes: sad_face > 0.0
      | | +--yes: [POS]
      | | +--no:  [NEG]
      | +--no:  [OBJ]
      +--no:  [POS]
```

Figure 11. Model, constructed from training on Spanish social media (Render_es).

Figure 11 shows the model, trained with a Spanish social media dataset. Here, the primary features were the number of vowels, capitalized characters, along with letter repetition, reflecting how sentiment is typically expressed in social media and other forms of informal communication. Also, adverbs were shown to be the most important sentiment-bearing words, along with presence of negation words and emoticons.

```
renderlex_adjective_sum > 0.1096
+--yes: senticnet > 15.509
| +--yes: renderlex_adverb_abs > 8.1989
| | +--yes: swn_posneg_ratio > 5.2202
| | | +--yes: [POS]
| | | +--no:  numpunc > 0.0313
| | |   +--yes: renderlex_pos_links > 8025.0
| | |   | +--yes: renderlex_adjective_sum > 1.1693
| | |   | | +--yes: [POS]
| | |   | | +--no:  [NEG]
| | |   | +--no:  [NEG]
| | |   +--no:  [POS]
| | +--no:  [POS]
| +--no:  numvowel > 0.2808
|   +--yes: renderlex_adjective_abs > 0.3998
|   | +--yes: [NEG]
|   | +--no:  [POS]
|   +--no:  swn_total_pos > 17.0
|     +--yes: [NEG]
|     +--no:  renderlex_noun_sum > 7.8051
|       +--yes: [POS]
|       +--no:  [NEG]
+--no:  senticnet > 27.085
  +--yes: [POS] [98.0]: 182
  +--no:  repeat_letter > 0.1193
    +--yes: senticnet > 13.511
    | +--yes: [POS]
    | +--no:  [NEG]
    +--no:  numpunc > 0.0306
      +--yes: repeat_letter > 0.0626
| +--yes: renderlex_neg_links > 317.0
      | | +--yes: swn_total_obj > 272.5
      | | | +--yes: repeat_letter > 0.1001
      | | | | +--yes: [NEG]
      | | | | +--no:  renderlex_adjective_abs >
[.. omitted for brevity ..]
      | | | +--no:  [POS]
      | | +--no:  [NEG]
      | +--no:  swn_total_neg > 16.75
      |   +--yes: [NEG]
      |   +--no:  [POS]
      +--no:  [NEG]
```

Figure 12. Model, constructed from training on English review data (PangLee).

Figure 12 shows the same model, trained on the movie review dataset. Here, almost the entire model is dominated by various lexicon features – total scores, absolute scores, positive-negative ratios. To a minor extent, surface features such as vowel and letter repetition appear.

```
numcaps > 0.0345
+--yes: senticnet_neg > 1.113
| +--yes: [NEG]
| +--no:  renderlex_adjective_sum_neg > 0.2178
|   +--yes: [POS]
|   +--no:  senticnet_neg > 0.084
|     +--yes: swn_total_neg > 3.0
|     | +--yes: [POS]
|     | +--no:  numcaps > 0.037
|     |   +--yes: [OBJ]
|     |   +--no:  [NEG]
|     +--no:  renderlex_all_abs > 1.5025
|       +--yes: senticnet_abs > 0.816
|       | +--yes: renderlex_adverb_sum > 0.8143
|       | | +--yes: [POS]
|       | | +--no:  swn_total_neg > 4.0
|       | |   +--yes: renderlex_adjective_sum > 0.0
|       | |   | +--yes: [NEG]
|       | |   | +--no:  [OBJ]
|       | |   +--no:  [OBJ]
|       | +--no:  [NEG]
|       +--no:  [OBJ]
+--no:  [OBJ]
```

Figure 13. Model, constructed from training on English news (JRC-en).

Figure 13 shows a similar picture than its Spanish counterpart in Figure 8, showing the importance of lexicon features, followed by surface features. In English, although all words were sentiment-bearing, adjectives and adverbs seem to be more informative, compared to nouns in Spanish.

Figure 14 shows the social media sentiment model for English. Here, lexicons seem to be the most indicative, followed by vowel repetition and proportion, presence of negation and capitalization. These models also demonstrate that in English, lexicon features tend to be closer to the root than in its Spanish counterparts. This could be explained either by the quality and coverage of lexicons for the respective language or even cultural differences, where the sentiment expression is present not only in the choice of words, but also in the capitalization, use of punctuation and phrasing.

```
senticnet_neg > 0.007
+--yes: numvowel > 0.2963
| +--yes: negation > 0.0
| | +--yes: [POS]
| | +--no:  renderlex_all_abs > 0.1811
| |   +--yes: [NEG]
| |   +--no:  [POS]
| +--no:  [NEG]
+--no:  swn_total_neg > 1.5
  +--yes: numcaps > 0.0439
  | +--yes: [POS]
  | +--no:  [NEG]
  +--no:  repeat_letter > 0.125
    +--yes: numpunc > 0.0299
    | +--yes: [POS]
    | +--no:  numcaps > 0.0368
    |   +--yes: [POS]
    |   +--no:  [NEG]
    +--no:  renderlex_all_sum > 0.1013
      +--yes: numvowel > 0.2727
      | +--yes: renderlex_all_sum > 0.419
      | | +--yes: renderlex_pos_links > 442.0
      | | | +--yes: numpunc > 0.044
      | | | | +--yes: [POS]
      | | | | +--no:  [NEG]
      | | | +--no:  renderlex_adjective_sum > 0.0949
      | | |   +--yes: [POS]
      | | |   +--no:  [NEG]
      | | +--no:  [POS]
      | +--no:  [POS]
      +--no:  [NEG]
```

Figure 14**.** A model, constructed from training on English social media (Render_en).

8 Conclusions

The obtained results confirm that social media content is the domain which benefits from external knowledge. Topic-specific lexicons can bring some minor improvement over general purpose lexicons, but the best-performing approaches use a combination of bag-of-words and lexicons training data. We reported improvement on two English datasets, especially on social media, which benefited significantly from pre-processing, surface features, as well as lexicons.

Moreover, having a two-layer model brings the most consistent performance across all domains and languages. In terms of comparison against state-of-the art studies, the best result on the Pang and Lee datasets scores at 0.90 F1, while ours was slightly lower at 0.88. However, on the news domain, our best approach even improves the performance on the JRC-EN dataset from the original authors' 0.65 to our result of 0.68 F1. On the other hand, the voting and meta-models did not show any improvement over the W+F model, and only improved the concatenation on some datasets, while performance was even reduced on the other datasets.

The analysis of the models shows that there are major differences between domains on which features are considered important: while news and review domains benefited from lexicons, surface features were important only in social media. On the other hand, both languages exhibited similar behavior across the same domains in news. By interpreting the models trained on social media we show that, for Spanish, surface features were more important than lexicons, while the opposite was observed for English.

This paper also demonstrates the feasibility of using machine translation to obtain a training corpus in another language, showing that the performance obtained for JRC-ES was the same as in the original version - JRC-EN. Other research [10] shows promising approaches to facilitate the knowledge transfer via lexicons using specifically tailored machine learning approaches. In future work we will explore cross-lingual learning, demonstrating approaches for training sentiment models using language resources from other languages.

Acknowledgements

This work was supported by the Slovenian Research Agency and the IST Programme of the EC under PASCAL2 (ICT-216886-NoE), XLike (ICT-STREP-288342), and RENDER (ICT-257790-STREP).

References

[1] Esuli, A. and Sebastiani, F. 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In Proceedings of the 5th LREC.

[2] Janyce Wiebe and Ellen Riloff. 2005. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In Proceeding of CICLing-05, pages 486–497, Mexico City, Mexico.

[3] Fellbaum, Ch. 1998. WordNet: An Electronic Lexical Database. MIT Press.

[4] Bizau, A., Rusu, D., Mladenic. D. 2011. Expressing Opinion Diversity. In Proceedings of the 1st Intl. Workshop on Knowledge Diversity on the Web (DiversiWeb 2011), Hyderabad, India.

[5] Hatzivassiloglou, V. and McKeown, K. 1997. Predicting the semantic orientation of adjectives. In Proceedings of the 35th Annual Meeting of the ACL.

[6] Perez-Rosas, V., Banea, C., Mihalcea, R: Learning Sentiment Lexicons in Spanish. In Proceedings of the LREC 2012

[7] Ohana, B. and Tierney, B: Sentiment classification of reviews using SentiWordNet, In Proceedings of 9th. IT & T Conference, 2009

[8] E. Cambria, C. Havasi, and A. Hussain. SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis. In: Proceedings of FLAIRS, pp. 202-207, Marco Island (2012)

[9] Melville, P. and Gryc, W. and Lawrence, R.D.: Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification. Proceedings of the 15th ACM SIGKDD, 2009

[10] Pang, B., Lee, L., and Vaithyanathan, S: Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of EMNLP 2002.

[11] Balahur, A. and Steinberger, R. and Kabadjov, M. and Zavarella, V. and Van Der Goot, E. and Halkia, M. and Pouliquen, B. and Belyaeva, J:. Sentiment Analysis In the News. Proceedings of LREC, 2010

[12] Yelena Mejova, Padmini Srinivasan: Crossing Media Streams with Sentiment: Domain Adaptation in Blogs, Reviews and Twitter. In Proceedings of the 6th ICWSM, ACM, 2012

[13] Bo Pang, Lillian Lee: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), pp. 1–135, 2008.

[14] Kaster, A. and Siersdorfer, S. and Weikum, G.: Combining text and linguistic document representations for authorship attribution, SIGIR workshop: Stylistic Analysis of Text For Information Access, 2005

[15] D. Kocev, C. Vens, J. Struyf and S. Džeroski, *Ensembles of multi-objective decision trees*. In J. Kok, J. Koronacki, R. de Mántaras, S. Matwin, D. Mladenic and A. Skowron, editors, Machine Learning: ECML 2007, 18th European Conference on Machine Learning, Proceedings. Lecture Notes in Computer Science, volume 4701, pages 624-631, Springer, 2007

[16] Bryll, R. and Gutierrez-Osuna, R. and Quek, F.: Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. Pattern recognition, vol 36., no.6., pp. 1291-1302, Elsevier, 2003

[17] Porter, M. F.: Snowball: A language for stemming algorithms, 2001

[18] Yang, Yiming and Liu, Xin: A re-examination of text categorization methods. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 42-49, ACM, 1999

Mining Web Logs to Identify Search Engine Behaviour at Websites

Jeeva Jose
Department of Computer Applications, BPC College,
Mulakkulam North P.O, Piravom- 686664, Ernakulam District, Kerala, India,
E-mail: vijojeeva@yahoo.co.in

P. Sojan Lal
School of Computer Sciences, Mahatma Gandhi University,
Kottayam, Kerala, India
E-mail: padikkakudy@gmail.com

Keywords: web logs, web usage mining, search engines, crawler

Received: April 3, 2013

Web Usage Mining also known as Web Log Mining is the extraction of user behaviour from web log data. The log files also provide immense information about the search engine traffic at a website. This search engine traffic is helpful to analyse the ethics of search engines, quality of the crawlers, periodicity of the visits and also the server load. Search engine crawlers are automated programs which periodically visit a website to update information. Crawlers are the main components of a search engine and without them the websites will not be listed in the search results. The visibility of the web sites depends on the quality of the crawlers. Different search engines may have different behaviour at web sites. We intend to see the differences in behaviour of search engines in terms of the number of visits and the number of pages crawled. The hypothesis was tested and it was found that there is a significant difference in the behaviour of search engines.

Povzetek: Analizirano je obnašanje različnih spletnih iskalnih algoritmov.

1 Introduction

Web Usage Mining is the extraction of information from web log files generated when a user visits the website [1]. Web mining tasks include mining web search engine data, analysing web's link structures, classifying web documents automatically, mining web page semantic structures and page contents, mining web dynamics (mining log files), building a multilayered and multidimensional web. Web log data is usually mined to study the user behaviour at websites. It also contains immense information about the search engine traffic. The user traffic is removed by pre processing tasks, otherwise it may bias the search engine behaviour. The crawler is an important module of a web search engine. The quality of a crawler directly affects the searching quality of web search engines.

The process of identifying the web crawlers is important because they can generate 90% of the traffic on websites [2]. Commercial search engines play a vital role in accessing web sites and wider information dissemination [3, 4]. Search engines use automated programs called web crawlers to collect information from the web. These web crawlers are also known as spiders, bots, robots etc. These crawlers are highly automated and seldom regulated manually [5, 6, 7]. The crawlers periodically visit the websites to update the content. Certain web sites like stock market sites or online news may need frequent crawling to update the search engine repositories. Web crawlers access the websites for diverse purpose which includes security violations also. Hence they may lead to ethical issues like privacy, security and blocking of server access. Crawling activities are regulated from server side with the help of Robots Exclusion Protocol. This protocol is present in a file called robots.txt. Usually ethical crawlers first access this file which will be present at the root directory of the website and follow the rules specified by robots.txt [8, 9]. But it is also possible to crawl the pages at a website without accessing the robots.txt. Certain crawlers seems to disobey the rules in robots.txt after its modification because crawlers like "Googlebot", "Yahoo! Slurp" , "MSNbot" cache the robots.txt file for a website [8]. The web site monitoring software Google Analytics does not track crawlers or bots. This is because Google Analytics tracking is activated by a JavaScript that is placed on every page of the website. A crawler hardly recognizes these scripts and hence the visits from search engines are not recognized. In this work we intend to see whether all the search engines are behaving in the same way when it accesses a website.

The most widely used log file formats are Common Log File Format and Extended Log File Format. The Common Log File format contains the following information: a) user's IP address b) user's authentication name c) the date-time stamp of the access d) the HTTP request e) the URL requested f) the response status g) the size of the requested file.

The Extended Log File format contains additional fields like a) the referrer URL b) the browser and its version and c) the operating system [11, 12]. Usually there are three ways of HTTP requests namely GET, POST and HEAD. Most HTML files are served via GET method while most CGI functionality is served via POST or HEAD. The status code 200 is the successful status code. Like the user access the website using a browser, the search engines also deploy user agents to access the web.

2 Background literature

Most of the works in Web Usage Mining is related to user behaviour. This is because websites like e-commerce websites will be interested in studying user behaviour for marketing, online sales and personalization. Several data mining tasks like clustering, classification, association rule mining etc. has been done for web log data of user behaviour. The web crawler ethics are measured to discover the ethicality of commercial search engine crawlers [9]. A survey of the use of the Robots Exclusion Protocol on the web through statistical analysis of a large sample of robots.txt files is done [10]. An empirical pilot study on the relationship between JavaScript usage and web site visibility was carried out to identify whether JavaScript based hyperlinks attract or repel crawlers resulting in an increase or decrease in web site visibility [6]. Another study is done with commercial search engines to find whether there is a significant difference in their coverage of commercial web sites [4]. A report on search engine ratings in United States is also available [3].

2.1 Preprocessing

The two data sets were extracted and it was found that the dataset 1 consists of 5,29,175 records for 8 weeks and dataset 2 consists of 2,60,775 records. The entries with unsuccessful status code 400 were eliminated. The HTTP requests with POST and HEAD was also removed. In addition all the user requests were removed to get the search engine requests. This is required as a user request in the input file may bias the results of search engine behaviour. After pre processing the resultant file contained only the successful search engine requests. Various search engine crawlers were identified. Some crawlers were identified from the IP address field. It contained substrings like “googlebot”, “baiduspider”, “msnbot” etc. The user agents were also helpful in identifying the bots or crawlers like Ezooms, discobot etc. Certain search engine crawlers with number of visits less than 5 per week was removed as it was considered irrelevant.The bots Ahrefbot, Seexie.com_bot, Turnitinbot, Yrspider were some of the bots in data set 1 whose number of visits were less than 5 in a week. For data set 2 the Alexabot was considered irrelevant. The crawlers in dataset 1 like Baiduspider, Discobot, Exabot, Feedtetcher-Google, Feedseeker, Gosospider, Ichiro, Magpie, MJ12bot, MSNbot, Seexie.com_bot, Slurp, Sogou, Sosospider, SpBot, Turnitinbot, Yahoo, Yeti, Yodao, Youdao and YrSpider were not present in dataset 2. After pre processing there were 22 crawlers for data set 1 and 5 crawlers for data set 2. The results for the number of visits made by various search engines of data set 1 is given in Table 1 and for data set 2 is given in Table 2.

We also intend to see the number of pages crawled by various search engines to see the dynamic behaviour of different search engines. Most of the search engines initially accessed the robots.txt file before crawling other pages except a few. Certain search engines crawled more pages compared with other bots or crawlers. For example the crawlers like Googlebot, Slurp, Bingbot, Feedfetcher-google, MJ12 etc crawled more number of pages and showed consistency in their behaviour. Table 3 shows the number of pages crawled by various search engines for data set 1 and Table 4 shows the result for data set 2.

2.2 Kruskal Wallis H test

Kruskal Wallis H Test detects if n data groups belong or not to the same population [13, 14]. This statistic is a non parametric test suitable to distributions that are not normal such as the exponential distributions observed in web usage mining or web log analysis [15]. The formula for H static of Kruskal- Wallis test is given below where K is the number of samples.

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{\mathrm{Rj2}}{\mathrm{nj}} - 3(N+1) \qquad (1)$$

where Rj is the sum of the ranks of the sample j, n_j is the size of the sample j, j=1, 2, 3, ...K and N is the size of the pooled sample ($n_1+n_2+.....n_K$). The calculated H value is to be compared against the chi-square value with (K-1) degrees of freedom at the given significance level α.

Case I

H_0: There is no significant difference between the number of visits made by various search engine crawlers.
H_1: There is significant difference between the number of visits made by various search engine crawlers.

From the test statistic in Table 5, both the data sets show a clear evidence of rejecting the null hypothesis. For data set 1, the p-value shows a strong evidence of rejecting the null hypothesis and for data set 2 shows a moderate evidence of rejecting the null hypothesis. The result of H test shows that there is a significant difference in the number of visits made by various search engines.

Case II

H_0: There is no significant difference between the number of pages crawled by various search engine crawlers.
H_1: There is significant difference between the number of pages crawled by various search engine crawlers.

Table 1: No: of visits by various crawlers for data set 1.

		Week										
No	**Crawler**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total**	**μ**	**σ**
1	Alexa	1	5	10	1	2	0	2	3	24	3.00	3.207
2	Baiduspider	128	222	65	89	124	67	66	47	808	101.00	56.87
3	Bingbot	157	166	159	175	126	100	118	96	1097	137.13	30.94
4	Discobot	113	33	0	21	24	52	5	69	317	39.63	37.42
5	Exabot	1	1	2	1	5	3	3	3	19	2.38	1.408
6	Ezooms	50	48	40	22	0	23	38	41	262	32.75	16.74
7	Feedfetcher-Google	179	170	167	223	192	191	187	188	1497	187.13	17.28
8	Googlebot	211	226	238	273	212	207	200	207	1774	221.75	23.99
9	Gosospider	26	10	1	0	0	0	0	0	37	4.63	9.303
10	Ichiro	117	81	122	146	0	42	21	33	562	70.25	53.8
11	Magpie	20	17	13	15	13	15	14	18	125	15.63	2.504
12	MJ12bot	38	36	37	50	37	37	37	41	313	39.13	4.643
13	MSNbot	24	17	11	19	15	12	18	15	131	16.38	4.138
14	Slurp	149	114	144	190	144	145	160	145	1191	148.88	21.07
15	Sogou	48	34	37	54	40	44	43	60	360	45.00	8.701
16	Sosospider	28	31	42	38	31	32	30	28	260	32.50	4.957
17	SpBot	3	3	3	4	2	2	1	1	19	2.38	1.061
18	Yandex	51	71	57	72	102	44	51	74	522	65.25	18.64
19	Yahoo	22	0	0	0	0	1	1	0	24	3.00	7.69
20	Yeti	3	4	1	4	3	2	4	4	25	3.13	1.126
21	Yodao	16	59	26	100	72	42	10	32	357	44.63	30.6
22	Youdao	2	4	1	1	18	1	3	0	30	3.75	5.898

Table 2: No: of visits by various crawlers for data set 2.

		Week										
No	**Crawlers**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total**	**μ**	**σ**
1	Ahrefsbot	79	0	1	19	37	66	31	48	281	35.13	28.6
2	Bingbot	31	41	27	43	23	30	28	17	240	30	8.64
3	Ezooms	3	20	26	38	26	24	9	28	174	21.75	11.1
4	Googlebot	42	49	42	44	42	49	35	60	363	45.38	7.41
5	Yandex	35	10	67	88	6	7	3	12	228	28.5	32.3

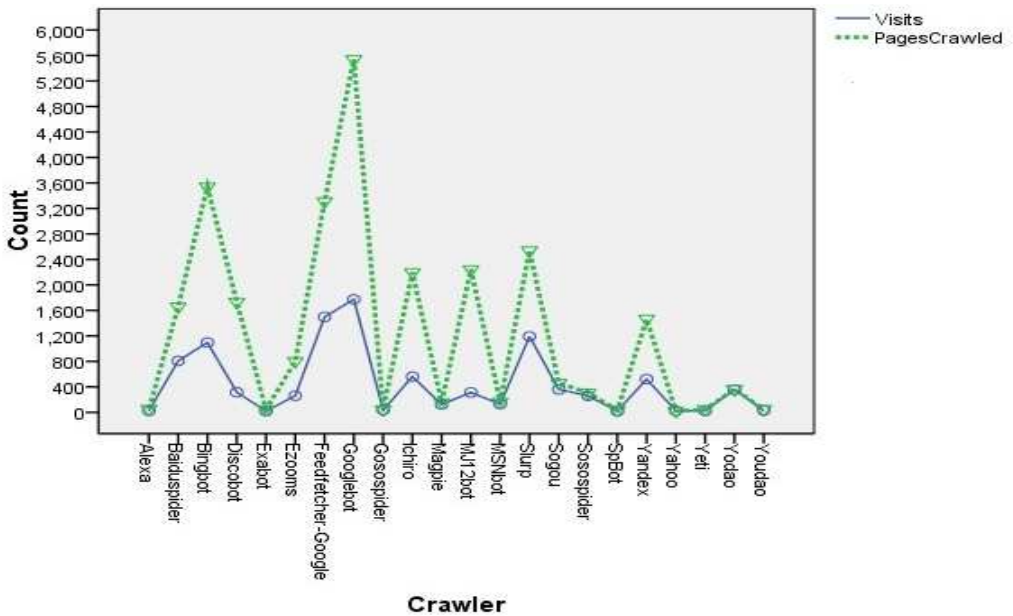


Figure 1: Time series sequence plot for data set 1.

Table 3. No: of pages crawled by various crawlers for data set 1

		Week										
No	**Crawler**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total**	**μ**	**σ**
1	Alexa	2	13	27	2	4	0	4	4	56	7.00	8.96
2	Baiduspider	219	674	102	124	260	98	94	90	1661	207.63	199.03
3	Bingbot	368	559	519	526	404	232	287	647	3542	442.75	143.30
4	Discobot	889	161	0	119	92	289	6	178	1734	216.75	287.42
5	Exabot	2	11	4	2	11	6	5	6	47	5.88	3.52
6	Ezooms	235	160	77	57	65	59	83	67	803	100.38	63.79
7	Feedfetcher-Google	386	343	340	493	442	447	443	417	3311	413.88	53.81
8	Googlebot	841	895	682	847	655	525	540	556	5541	692.63	150.42
9	Gosospider	34	11	1	0	0	0	0	0	46	5.75	12.03
10	Ichiro	230	277	387	414	320	234	45	291	2198	274.75	113.86
11	Magpie	23	21	18	23	16	16	18	22	157	19.63	2.97
12	MJ12bot	174	304	224	392	255	285	294	316	2244	280.50	65.06
13	MSNbot	31	24	13	28	17	15	18	18	164	20.50	6.44
14	Slurp	367	253	297	410	310	264	308	331	2540	317.50	51.79
15	Sogou	72	42	47	61	52	54	51	80	459	57.38	12.89
16	Sosospider	32	38	57	42	36	36	35	33	309	38.63	8.03
17	SpBot	6	6	6	8	4	4	2	2	38	4.75	2.12
18	Yandex	140	250	99	171	216	102	212	276	1466	183.25	66.20
19	Yahoo	22	0	0	0	0	0	0	0	22	2.75	7.78
20	Yeti	6	9	2	7	7	4	7	7	49	6.13	2.17
21	Yodao	16	59	27	102	75	43	10	34	366	45.75	31.29
22	Youdao	4	8	2	2	25	2	7	2	52	6.50	7.86

Table 4: No: of pages crawled by various crawlers for data set 2.

		Week										
No	**Crawler**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total**	**μ**	**σ**
1	Ahrefsbot	282	0	1	19	108	119	46	74	649	81.13	93.08
2	Bingbot	66	172	158	251	102	90	78	48	965	120.63	68.03
3	Ezooms	3	23	35	51	32	36	9	40	229	28.63	16.08
4	Googlebot	74	92	83	99	90	95	65	83	681	85.13	11.33
5	Yandex	39	18	123	199	6	7	4	13	409	51.13	71.65

The test statistic in Table 6 also shows that there is significant difference in the number of pages crawled by various search engines. The p-value for both the datasets is a strong evidence of rejecting the null hypothesis. A time series sequence plot was done for both data sets with total number of visits and total number of pages crawled. The result for data set 1 is shown in Figure 1 and for data set 2 is shown in Figure 2. We also intend to see whether there exists any correlation between the number of visits and number of pages crawled. The Karl Pearson's Correlation Coefficient [14] was calculated for both data sets. The data set 1 showed a strong positive correlation of 0.932 whereas the data set 2 showed a moderate positive correlation of 0.505.

3 Conclusion

The obtained results point to the differences in the behaviour of web crawlers by various search engines. The more the number of search engines accessing a website, the more will be its visibility when searching for a particular web site. The observed results show that all search engine crawlers are not visiting all the websites. In our experiment the data set 1 was accessed by more number of search engines compared to data set 2. Certain search engines were consistent in the number of visits and number of pages crawled while a few were not consistent or irregular in their visits and pages crawled. It is found that data set 1 is more visible to search engine crawlers as it is crawled by more number of search engines compared to data set 2. The results also showed a positive correlation between the number of visits and number of pages crawled. A better search engine optimization policy can be followed to make the websites visible to different search engines so that the websites will be listed top in the search engine rankings.

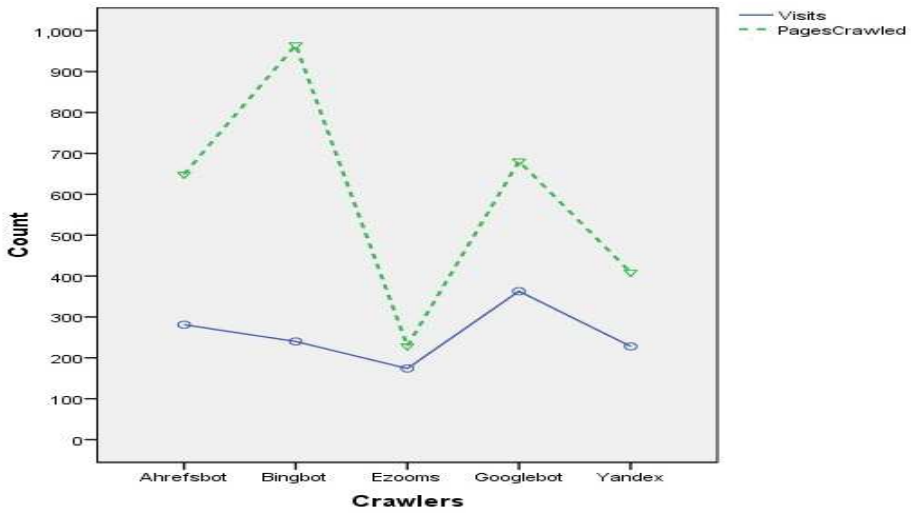


Figure 2: Time series sequence plot for data set 2.

Table 5: Test Statistic for Case I.

Kruskall Wallis Test		
	Data Set 1	**Data Set 2**
α	0.01	0.01
p-value	0.0001	0.044
Chi-square	148.734	9.799
df	21	4

Table 6: Test Statistic for Case II.

Kruskall Wallis Test		
	Data Set 1	**Data Set 2**
α	0.01	0.01
p-value	0.0001	0.013
Chi-square	154.85	12.714
df	21	4

Acknowledgement

This research work is supported by Kerala State Council for Science Technology and Environment, Kerala State, India as per Order No.009/SRSPS/2011/CSTE .

References

[1] Kosala, R. And Blockeel, H., Web Mining Research: A Survey. ACM SIGKDD Explorations. 2(1), pp. 1-15, 2000.

[2] Mican, D. And Sitar-Taut, D., Preprocessing and Content/Navigational Pages Identification as Premises for an Extended Web Usage Mining Model Development. *Informatica Economica*, 13(4), pp. 168-179, 2009.

[3] Sullivan, D.2003, Webspin : Newsletter [online]. Available from: http://contentmarketingpedia.com/Marketing-Library/Search/industryNewsSeptA1.pdf .Accessed December 4, 2012.

[4] Vaughan, L. And Thelwall, M., Search Engine Coverage Bias: Evidence and Possible Causes, *Information Processing and Management*, 40(4), pp. 693-707, 2004.

[5] Bhagwani, J. And Hande, K., Context Disambiguation in Web Search Results Using Clustering Algorithm. International Journal of Computer Science and Communication, 2(1), pp. 119-123, 2011.

[6] Schwenke, F. And Weideman, M., The influence that JavaScript has on the visibility of a website to search engines – a pilot study. Informatics & Design Papers and Reports, 11(4), pp. 1-10, 2006.

[7] Thelwall, M., A Web Crawler Design for Data Mining, Journal of Information Science, 27(5), pp. 319-325, 2001.

[8] Drott, M, Indexing aids at corporate websites: The use of robots.txt and meta tags. Information Processing and Management, 38(2), pp. 209-219, 2002.

[9] Lee Giles, C., Sun, Y and Council, G., I., Measuring the Web Crawler Ethics. In: Proceedings of WWW 2010, ACM, pp. 1101-1102, 2010.

[10] Sun, Y. Zhuang, Z. .and Lee Giles, C., A Large-Scale Study of Robots.txt. In: Proceedings of WWW2007, ACM, pp. 1123-1124, 2007.

[11] Wahab, M.H,A, Mohd, M.N.H, Hanafi, H. F. Mohsin, M. F.M., Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm. In: Proceedings of World Academy of Science Engineering and Technology, pp.190-196, 2008.

[12] Spiliopoulou, M., Web Usage Mining for Web Site Evaluation. *Communications of the ACM*, 43(8), pp. 127-134, 2000.

[13] Kruskal,W. H. And Wallis, W. A., Use of Ranks in one-criterion Variance analysis. Journal of the American Statistical Association, 47(260), pp. 583-621, 1952.

[14] Paneerselvam, R., Research Methodology. New Delhi, Prentice Hall of India Private Limited, 2005.
[15] Ortega, J., L. And Aguillo, I., Differences between web sessions according to the origin of their visits, Journal of Infometrics, 4, pp. 331-337, 2010.

An Ultra-fast Approach to Align Longer Short Reads onto Human Genome

Arup Ghosh and Gi-Nam Wang
Unified Digital Manufacturing Lab
Department of Industrial Engineering, Ajou University
San 5, Woncheon-dong, Yeongtong-gu, Suwon 443-749, South Korea
E-mail: {arupghosh, gnwang}@ajou.ac.kr

Satchidananda Dehuri,
Department of Systems Engineering,
Ajou University, San 5, Woncheon-dong, Yeongtong-gu, Suwon 443-749, South Korea
E-mail: satchi@ajou.ac.kr

Keywords: DNA, sequence alignment, second-generation sequencing (SGS), substring matching, BWT

Received: April 13, 2013

With the advent of second-generation sequencing (SGS) technologies, deoxyribonucleic acid (DNA) sequencing machines have started to produce reads, named as "longer short reads", which are much longer than previous generation reads, the so called "short reads". Unfortunately, most of the existing read aligners do not scale well for those second-generation longer short reads. Moreover, many of the existing aligners are limited only to the short reads of previous generation. In this paper, we have proposed a new approach to solve this essential read alignment problem for current generation longer short reads. Our ultra-fast approach uses a hash-based indexing and searching scheme to find exact matching for second-generation longer short reads within reference genome. The experimental study shows that the proposed ultra-fast approach can accurately find matching of millions of reads against human genome within few seconds and it is an order of magnitude faster than Burrows-Wheeler Transform (BWT) based methods such as BowTie and Burrows-Wheeler Aligner (BWA) for a wide range of read length.

Povzetek: Metoda omogoča izredno pohitritev iskanja daljših vzorcev v človeškem genomu.

1 Introduction

The rapid advances in DNA sequencing technology have dramatically accelerated the biomedical and biotechnology research [2, 6, 28]. Thereby opportunities have been created for data mining researchers to analyze a gamut of data. With the advent of second-generation sequencing (SGS) technologies, there is an increasing pressing need of an approach that can align large collections of reads (possibly millions) onto the reference genome rapidly. The main motivation behind this read alignment is to discover commonalities and connections between newly sequenced molecules with respect to existing reference genomes [16].

Currently, DNA sequencing machines are capable of generating millions of reads in a single run when a DNA sample is given as an input [9, 16, 27]. The DNA sequencing machines take the DNA sample as input and break it into a number of short pieces, which then are again broken into equal-length fragments called reads [25]. The 'read alignment problem' is to find matching of those reads onto a reference genome. From the computer science point of view, a genome can be considered as a long string of characters/bases (human genome contains nearly 6 billion characters/bases), and reads can be regarded as a set of equal-length small strings of characters/bases. Now, read alignment task is to map those reads (small string of characters) onto genome (long string of characters). Simply, we can think of it as a common substring matching problem [25]. The main challenge of this read alignment problem is to efficiently build the reference genome index thus reads (usually millions) can be mapped rapidly. This read alignment task has many potential applications in biomedical and bioinformatics fields, for example: 'to detect genetic variations' [4, 21] which will indeed help to identify 'disease genome' [21], 'to map DNA-protein interactions' [18], 'to profile DNA methylation patterns' [11, 13], etc.

To deal with this read alignment problem, several read alignment tools or approaches have been proposed. However, they are primarily focused on previous generation short reads which are usually of 25-70 bases long [26, 27]. Unfortunately, with the advent of SGS technologies DNA sequencing machines have started to produce reads (named as longer short reads) which are much longer than the previous short reads. Read lengths have just increased to more than 100 bases within a few years [27]. This trend of increment in read length makes the existing aligners computationally infeasible. Hence, there is an increasing need of an approach that can

handle this current generation reads efficiently and also can handle future generation more long reads (by observing the trend). Here the particular importance of the longer short read alignment problem can be realized. It is theoretically and also practically difficult to avoid the overhead of processing the increased read length. However, it is needed to bind the growth rate of the processing cost efficiently. Currently, most of the read aligners are unable to achieve this scalability which makes them limited to the short reads. To this end, this paper proposes an ultra-fast method for aligning longer short reads onto human genome by combining the best attributes of hash based indexing and searching. Our approach is not bounded to a particular range of reads and can scale well for more long reads.

The remainder of this paper is organized as follows. Section 2 discusses the related work. Our proposed approach is described in Section 3. Experimental results are presented in Section 4. Section 5 contains our conclusive remarks of the work followed by a list of relevant and state-of-the-art references.

2 Related work

The approaches proposed so far by the several research groups for read alignment problem can be broadly classified into four categories.

1) Traditional sequence mapping tools, such as Basic Local Alignment Search Tool (BLAST) [1] and BLAST-Like Alignment Tool (BLAT) [19], are unable to cope efficiently with the massive amount of reads generated by the current generation DNA sequencing machines, which make it computationally infeasible for solving the current generation read alignment problem [9, 16, 24].

2) BWT [7] based approaches, such as BowTie [20] and BWA [22], create a BWT based index and use an iterative prefix matching technique to find an alignment. A BWT-based index takes small memory footprint for example, BowTie takes less than 2 GB [30] and BWA takes less than 6 GB [29] memory to work with complete human genome. BWT based approaches have another significant feature i.e., they can handle a wide range of read lengths. For example, BowTie can handle up to 1024 bases read length [30]. So, it can easily handle current generation reads and also able to handle future generation more long reads. However, its performance degrades rapidly as the read length increases [25].

3) Hash table based approaches have got more and more popularity nowadays. Some of them create hash table based index for reads e.g., Efficient Large-Scale Alignment of Nucleotide Databases (ELAND) [10], Mapping and Assembly with Quality (MAQ) [23], Short Read Mapping Package (SHRiMP) [26] etc. Other approaches use hash table for the reference genome indexing e.g., Wisconsin's High- throughput Alignment Method (WHAM) [25], Periodic Seed Mapping (PerM) [9], Short Oligonucleotide Alignment Program (SOAP) [24], etc. However, only Q-Pick [16] uses hash table for both read and reference genome indexing. Hash table based approaches are in general significantly faster. However, those hash table based approaches or their software implementation have some significant drawbacks. WHAM and Q-Pick create reference genome index for a specific length of the read, which cannot be used for the different length reads (means, if WHAM and Q-Pick create index to align X bases length reads then that index cannot be used for alignment of N bases length reads where $X \neq N$). This is a significant issue because we have to create index for each of the read length. This will cause a significant overhead with respect to the index building time and disk space consumption because, nowadays most of the genome sequence mining companies have large number of databases of varied read lengths. The most significant problem with the above approaches is that, they are primarily focused on short reads. Thus, these approaches or their software implementations are limited to a specific read length which does not cover the read length of the current generation (for example, currently Illumina can produce read length up to 250 base long [31]) and there is no straight forward way to extend it to handle current generation longer short reads (or future generation more long reads). For example, ELAND can handle up to 32 bases [24, 33], MAQ can handle up to 127 bases [20], Shrimp can handle up to 70 bases [26], WHAM can handle up to 128 bases [36], PerM can handle up to 64 bases [34], SOAP can handle up to 60 bases [35] which are significantly lower length than the current generation read length.

4) Sorted Index File based approach such as fetchGWI and tagger [17] index either the reference genome or the query set and perform an efficient mapping of those two set of sorted entries (one for reference genome and another for query set) to find matches. However, this approach is also limited to 30 bases read length [17].

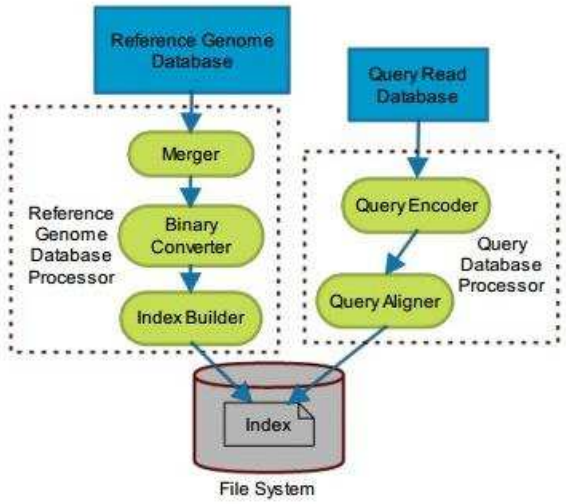


Figure 1: System Architecture.

Taking into account the limitations presented above, it can be summarized that, there is a significant lack of approaches or tools that can handle longer short reads efficiently. We propose to use BowTie and BWA approaches to solve this problem. From the trend of increment in read lengths, it does not seem that it will become infeasible in near future (because of its wide range of read length acceptance). So, we take BowTie and BWA as our base methods to experiment with longer short reads. As memory became cheap nowadays, we have no need to keep such unnecessary tight memory restriction in our approach as maintained by BowTie and BWA.

3 Proposed approach

This section is divided into two Subsections. In subsection 3.1, the statement of the problem is defined. The system architecture and working procedure of the proposed approach are described in subsection 3.2.

3.1 Problem Statement

A complete genome sequence is a set of all its chromosomal sequences. A chromosomal sequence is a series of characters. Each character (nucleic acid) is represented by the symbols A, G, C, or T (stands for adenine, guanine, cytosine and thymine respectively) or an unknown/ambiguous character, named N. The unknown character, N, represents that there is an uncertainty about the nucleotide in that position or there is a repetitive junk region in the genome and thus, all nucleotides in that region are converted into N's [25]. In the genome sequencing task, it has no biological sense to match reads onto those repetitive junk regions [25]. For simplicity, we can think N indicates error while matching [30].

The read alignment task is to efficiently build an index of the reference genome thus a fast and exhaustive mapping of a large collection of equal length query reads is possible while maintaining the accuracy in alignment. Query read database usually contains millions of reads and while mapping, read aligner has to report the matching position/s in the chromosomal sequence (if any matching occurs).

3.2 System Architecture of Proposed Approach

System architecture of our proposed approach is given in Figure 1. It has two main components: i) reference genome database processor and ii) query database processor. We will discuss them separately to present our approach in greater detail.

3.2.1 Reference Genome Database Processor

Reference genome database processor takes the complete genome sequence database as an input and creates an index for that genome into the file system (Figure 1). Complete genome sequence contains full set of chromosomal sequences. Note that though we are interested in mapping the query reads on both the forward and reverse strands of each chromosome, we will build index only for forward strand of each chromosome. We will compensate this while processing the query database (detail in subsection 3.2.2). We have selected this technique to reduce the index size because with this technique, we have to process only the half of the original genome sequence which will indeed provide us with speed gain while query read is searching.

Main idea behind our approach is to store in index all possible substrings of length L of every chromosomal sequence (only forward strand) with its position information. We set the length L value to 32. Note that as we are going to index each possible substring in a hash

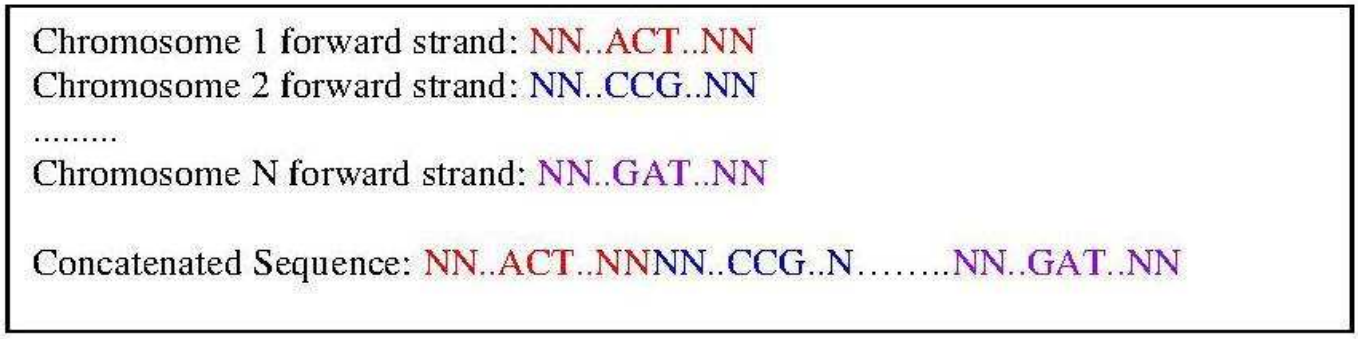


Figure 2: Concatenated Chromosomal Sequence.

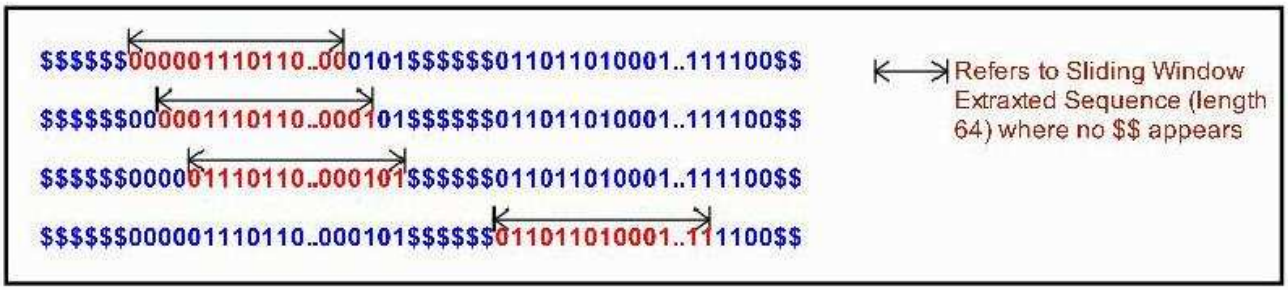


Figure 3: Sliding Window Extraction Protocol.

table, there are 4^{32} possible values (recall that, sequence can contain four bases i.e. A, C, G, or T), which will dramatically mitigate the possibility of hash collision.

Reference genome database processor starts working by concatenating (or merging) the forward strands of each chromosome (Figure 2). This is done by merger part of reference genome database processor (Figure 1). Note that with this concatenation we will lose information about "in which chromosome (or in which position of that chromosome) the subsequence of the concatenated sequence originally belongs to?" This information will be required when we find a match on a specific position on the concatenated sequence. Note that chromosome number (or position value in that chromosome) can be easily calculated by noting down the length of each chromosome. Motivation behind concatenation is to reduce the index space because with this technique, we have no need to store the chromosome number as we are going to calculate it during query read processing phase (detail in subsection 3.2.2).

Binary converter takes the concatenated sequence from merger (Figure 1) and converts each A/C/G/T character into two bit binary representation. A, C, G, T will be binary represented by 00, 01, 10, 11 respectively. Note that, as we are going to index each possible substring of L = 32, this representation will allow us to pack each of them into one computer word in the 64 bit computer architecture. Actually, we have set the L value to 32 thus our method can take advantage of current day's 64 bit computer architecture. Also note that, we are not going to index the subsequences in which N occurs (because N indicates 'error in matching'). So, if N occurs in the concatenated sequence, we will simply replace it by any two special characters (say with '$$'). By doing so, we can identify if N has occurred in the sequence. For example, if the concatenated sequence is 'NGACTN', Binary Converter will encode it as '$$10000111$$'.

Index builder takes the binary converter outputted sequence and creates an index which can be used by query database processor (Figure 1). Index builder moves a sliding window of length 64 over the input sequence and extracts the subsequence within it, and then moves two positions (Figure 3). Recall that, by doing so, it is originally extracting all possible subsequences of length L = 32 from the concatenated chromosomal sequence. Sliding window will extract the subsequences only if no $$ ($$ refers to N which means error in matching) appears in that window (Figure 3). Here, we have to keep in mind that sliding window should not extract any subsequences which do not belong to the original chromosomal sequences. This can happen while extracting subsequences from the position of concatenation of chromosomal sequence n and chromosomal sequence (n + 1) [here, n = 1, 2… up to (maximum chromosome number – 1)] (depicted in Figure 4). This can be easily avoided by keeping in mind the length of chromosomal sequences.

All extracted subsequences, which are basically 64 bit integer numbers, are hashed and their hash value provides the hash table bucket number. We have used Thomas Wang's hash function [14] to uniformly distribute values over the hash table. Thomas Wang's hash function has been widely used by many approaches for various purposes [8, 12, 15]. This is well suited for our purpose because it is fast to compute and has very high avalanche effect [3, 14]. Hash table values are the position values of the corresponding subsequences (represented by 32 bit integer numbers) in the concatenated chromosomal sequence. All those key-value pairs are inserted into the hash table, whose structures are depicted in Figure 5. Our hash table

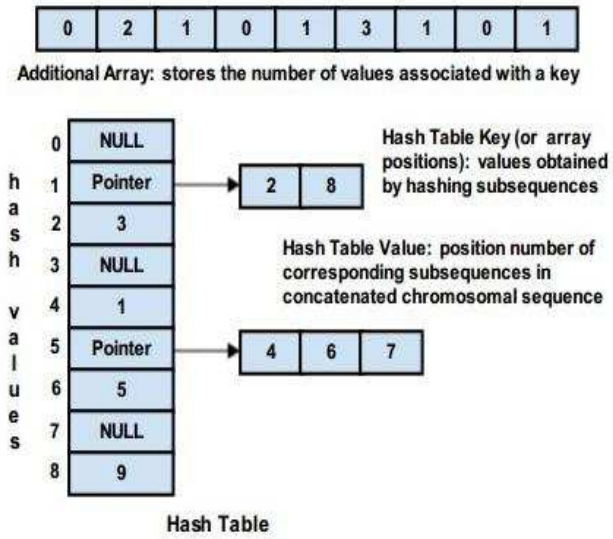


Figure 5: Hash Table Structure.

structure is basically a long array, initially filled with NULL values and when we have to insert a key-value pair, we just insert that value in the corresponding array position (array position is found by hashing the key). Note that if corresponding array position is filled, then it will be replaced by a pointer to an array and the old value (or values) and the new value will be inserted into that

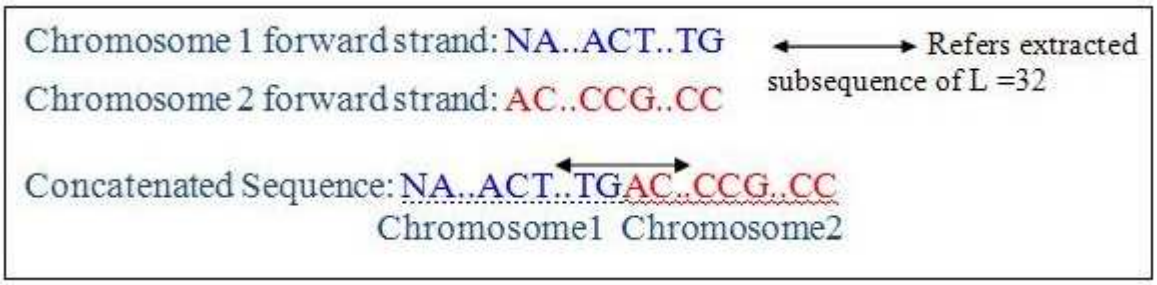


Figure 4: Error in Subsequence Extraction.

array. We have used this type of hash table structure in place of traditional hash table structure (which is usually two dimensional linked lists or an array of linked lists) to reduce the hash table space requirement. This kind of hash table structure efficiently reduces the requirement of pointers with the cost of an additional array which stores the number of values associated with the corresponding key (easily represented by 8 bit integer number - recall that we have used subsequence of length L = 32 to mitigate the possibility of collision). By doing this, we can dramatically reduce the hash table space requirement (realized through experiments also) because the size of a pointer in current day's system architecture is much longer than the 8 bit integer and many subsequences may appear only one time in the concatenated chromosomal sequence.

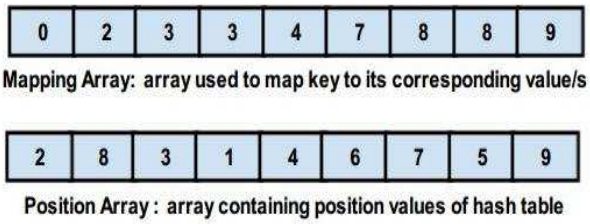


Figure 6: Converted Sequential Structure Hash Table.

However, though we have reduced the space requirement with that hash table structure (as presented in Figure 5), there are some issues with that data structure also. In such data structure, index loading will be quite time consuming as we are not able to bulk load that data structure after saving it into the file system. For a significant number of hash table keys (as many subsequences can appear more than once), every time we have to access the values (a key can have many corresponding values) associated with it through a pointer access. This will make it relatively slow in comparison with the data structure where we can access the values directly. Thus, we convert our hash table structure (Figure 5) into two sequential structures shown in Figure 6. This will make bulk loading possible and with that data structure we can directly access the value/s through its corresponding key. In Figure 6, we have depicted the conversion of the hash table structure shown in Figure 5 to the sequential structures. The conversion algorithm is simple. The position array (as in Figure 6) contains all the hash table values (or position numbers - as in Figure 5) inserted into it, one by one according to hash bucket number, started from n = 0 to (number of hash bucket -1). To calculate the mapping array value at position i, we have to just add Additional Array value of position i and mapping array value at position (i -1) (see Figure 5 and 6). Please note that, for i = 0, this is not true (as array position can't be negative). So, we have to check every time whether i = 0 or not. We just remove this checking requirement by making hash function to produce hash values greater than 0 and setting 0^{th} index of mapping array to 0 (see Figure 5 and 6). This will help us to avoid checking (thus speed gain) while performing query read mapping. Now, from mapping array, we can easily map keys to its corresponding value/s. For example, suppose the hash value of a key is i where i can be any value ranging from 1 to (number of hash bucket - 1). Now, from Figure 6, we can easily find that, (mapping array [i] - mapping array [i-1]) provides the number of value/s associated with that key (for example, if i = 1, then the key has two values associated with it, also see Figure 5). To find that value/s, we have to just run a loop, collecting value/s from position array starting from position number mapping array [i-1] (for example, if i = 1, we have to collect two values from position array starting from position number 0). The mapping process is presented in Figures 5 and 6.

3.2.2 Query Database Processor

Query database processor takes a query read database (possibly contains millions of equal length reads) and the saved index (index saved into file system by reference genome database processor (Figure 1)) as inputs and outputs 'query read matching information' into file system for each such matched reads. The 'query read matching information' contains information about query read alignment region (at what position in which chromosome the matching occurs), number of other alignments etc.

Our query database processor starts working by bulk loading the index (index refers to mapping array and position array as in Figure 6). This bulk loading (which will save significant amount of time) is possible only because we have converted our index into two sequential structures. After loading the index into memory, our query database processor takes each query read from the query database and searches into the index for matching in the following manner.

Query database processor will process each query read following the same procedure as done in section 3.2.1, except, it will not process the query reads in which N (or error in matching) occurs. Query read encoding (dividing the read into subsequence of L = 32 and converting them into binary) is performed by the Query Encoder and Query Aligner is responsible for matching task (Figure 1). Please remember that we have indexed all possible subsequence of length L = 32 of the genomic sequence. Hence, Query Encoder will first divide each query read into the subsequence of length L = 32. For example, if the query read is of 100 bases, Query Encoder will divide it into four subsequence of length L = 32. The first subsequence will be from base 1 to 32, the second subsequence will be from base 33 to 64, the third subsequence will be from base 65 to 96, and the fourth subsequence will be from base 69 to 100. Note that, the last subsequence will be taken from the end of the read and overlapping in subsequence may happen. Query Aligner searches the index for each such subsequence (after binary converted by Query Encoder) of the query read by hashing and mapping them into the hash table (following the same procedure as stated in section 3.2.1). Returned matching position/s is stored into arrays. If any of the subsequence of that read is failed to align, then we can easily conclude that the read is failed to align. The worst case time complexity to find it is $O(2n)$ where n =

number of subsequence of that read. However, reverse is not true because, if searching of all query read segments is successful, it only means that all the query read segments appear in the concatenated chromosomal sequence and not necessarily mean that the whole read appeared in the concatenated chromosomal sequence.

To check if the read is aligned or not, Query Aligner has to perform some additional task i.e., it has to check whether the returned positions are the consecutive segment positions or not. Take example of 100 bases read length. Suppose, all the four subsequence are able to align and returned position value/s are stored in the arrays named A1 = {200, 415}, A2 = {232, 327, 1215}, A3 = {264, 416, 917, 971} and A4 = {268} consecutively (values inside the curly brackets are the returned position value/s). Now, to check whether the read is aligned or not (or in which position/s), Query Aligner has to search for (A1[i] + 32) in A[2], (A1[i] + 64) in A[3], (A1[i] + 68) in A[4] means for every value of i i.e., from 0 to (size of A1 – 1). By doing so, we are only checking whether the segments are consecutive segments in the concatenated chromosomal sequence or not. If searching in all the arrays is successful, then only we can conclude that the read is aligned at position A1[i] in the concatenated chromosomal sequence (for above example, the query read matches only in position 200). Here, we should mention that all arrays that store the returned matching positions are the sorted array (easy to see). Thus, Query Encoder will perform an efficient linear search in the sorted array to find a match, instead of other searching procedures (for example binary searching). This will help us to gain speed over other searching procedures because the length of the array is typically very small due to very high indexed substring length i.e., 32 [5, 32]. With this linear search we can find all the matches by only one pass through the array (means with worst case time complexity $O(n)$ where n = very small array length). Another point to note that, with the above procedure we can only identify in which position of the concatenated chromosomal sequence the match occurs. Now, Query Aligner finds the original matching position (means chromosome number and the position value in that chromosome) by using the following procedure. First it finds the previous chromosome ending position in the concatenated chromosomal sequence (so, chromosome number is found) and then deducts that position value from the matched position value (except that matched position is not within the first chromosome ending position) to find real position in that chromosome. The previous chromosome ending position is found by performing an efficient binary search on a sorted array which contains the ending position of each chromosome in concatenated chromosomal sequence (reported by Reference Genome Database Processor).

As mentioned above, only the forward strand of each chromosome is processed by the Reference Genome Database Processor (subsection 3.2.1). This will be addressed in details in this section. Two strands of chromosome are of complementary nature i.e., A always pairs with T, and C always pairs with G (vice versa). So, for each query read in the query read database, Query Database Processor will not only search for that query read but also search for the reverse complement of it. For example, suppose, a query read is 'ACCTGGA'. Query Database Processor will first reverse it i.e., 'AGGTCCA' and then will take complement of it i.e., 'TCCAGGT' and then search into the index for matching following the same procedure as stated above.

From the above, it is easy to see that our approach has no upper limit restriction on the read length like many other approaches. In the next section, we will provide empirical evaluation of our approach for a wide length of reads.

4 Experimental study, results, and discussion

We ran our experiments on a desktop computer with 3.70 GHz Intel Xeon dual-cores CPU and 32 GB of DDR3 main memory, running 64 bit Ubuntu (kernel 3.5.0) as operating system. All our algorithms were implemented in C++, and compiled using g++ 4.7.2. We had followed the similar comparison strategy as performed in [25]. We have taken repeat-masked NCBI build 36 human genome as our reference genome and all the approaches obliged to report all the valid matches (as done in [25]). Our

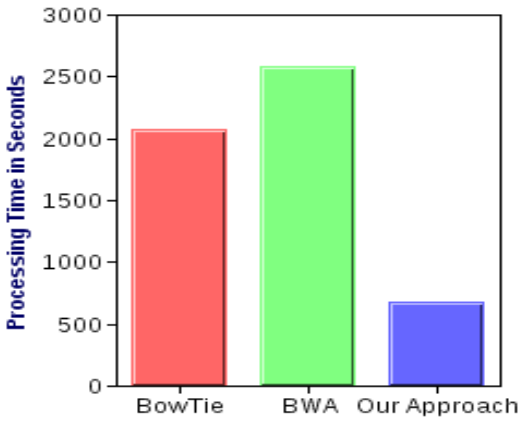


Figure 7: Comparison of Index Building Time.

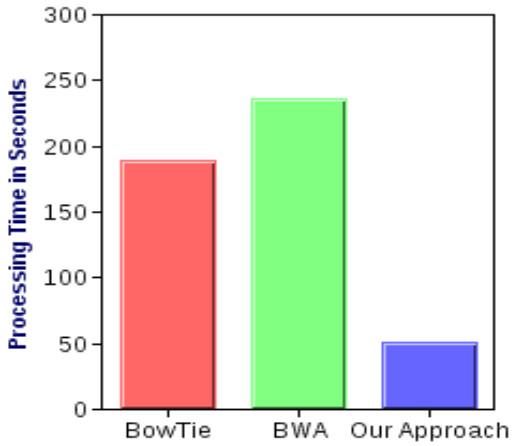


Figure 8: Comparison of Query Read Database Aligning Time for Read Length of 100 bases.

approach followed the same default output format of BowTie and all the approaches ran on single thread.

We have performed experiments with various length reads i.e. with 100, 150, 200, 250 bases read length (note that 250 bases read length is the currently maximum read length which Illumina can produce [31]). We selected four sets of query read database for experimenting with each of the read lengths. Our experimental results show comparison of our approach with BowTie [20, 30] and BWA [22, 29] by averaging results over those four sets of query read database. Each of our read databases contains 10 million of query reads. We have set the hash table bucket number to 1.5 billion throughout our experiments. We have also created four synthetic query read databases, one for each 100, 150, 200, 250 bases read lengths (each contains 10 million query reads with randomly inserted errors) to measure accuracy of our approach.

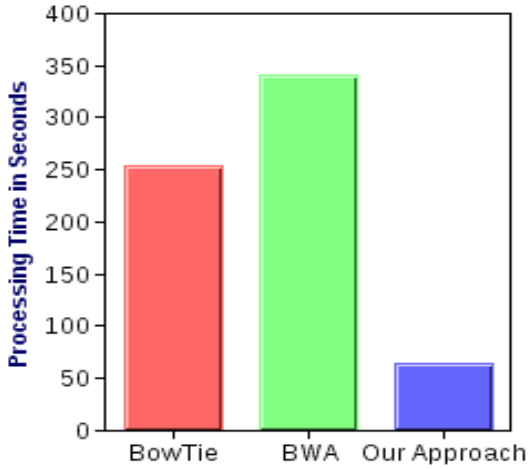


Figure 9: Comparison of Query Read Database Aligning Time for Read Length of 150 bases

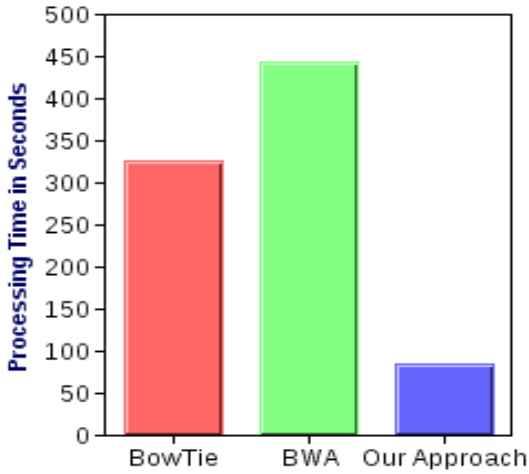


Figure 10: Comparison of Query Read Database Aligning Time for Read Length of 200 bases

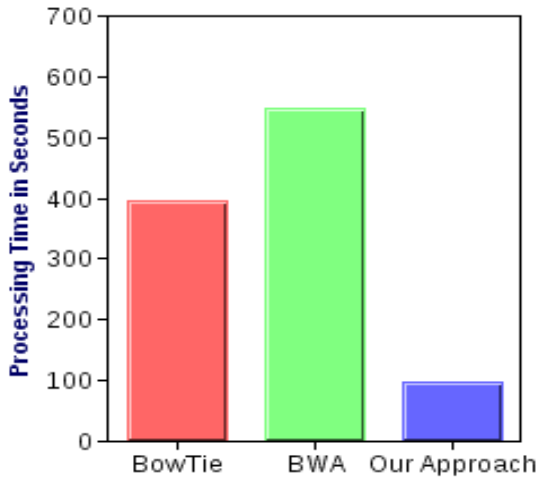


Figure 11: Comparison of Query Read Database Aligning Time for Read Length of 250 bases

Comparison of index building time of our approach with BowTie and BWA is given in Figure 7. Our experimental result show that our approach is significantly faster than BowTie (3X faster) and BWA (3.8X faster) as presented in Figure 7. Please note we have to build our index just only one time for a genome and we can use it repeatedly for searching various length reads (easy to see from section 3) unlike many other approaches. Experimental results for comparison of our approach with BowTie and BWA for various length reads i.e. of 100, 150, 200, 250 bases read length are given in Figures 8-11 respectively. From those experimental results it can be easily seen that, our approach is significantly faster for query read alignment than BowTie (3.7X, 3.9X, 3.7X, 4X faster for 100, 150, 200, 250 bases read length respectively) and BWA (4.6X, 5.2X, 5.1X, 5.6X faster for 100, 150, 200, 250 bases read length respectively). By significantly reducing the index building and query read searching time over BowTie and BWA, our approach is able to fulfill its primary motivation. To measure how much accurate our approach is, we ran it on four synthetic databases one for each 100, 150, 200, 250 bases read length, where it was previously known a number of query reads that provide an alignment. Our approach is able to align exactly the same number of query reads within these databases. In addition, during the previous experiments with BowTie for various length read databases (i.e. four sets of databases for each of 100, 150, 200, 250 bases read length, as stated early of this section), we have found that for all the databases of all the read length, our approach is able to align exactly the same number of query reads as aligned by BowTie (which is one of the most accurate read aligner as can be found from the experimental results of [16]). Actually, the accuracy of our approach can be outlined as follows:

- We have indexed subsequence of length L = 32 and used Thomas Wang's hash function (which uniformly distributes the key values) to mitigate the possibility of collision.
- We have used large number of hash table buckets i.e. 1.5 billion during our experiments which will also dramatically mitigate the possibility of collision. During index building time, we have found that our approach has extracted around 1.25 billion subsequences of

length L = 32 from NCBI human genome (build 36) which is quite lower value than 1.5 billion.

- Our approach is primarily targeted for current generation reads (or future generation more long reads) which is > 100 bases. So, query reads will be divided into >= 4 fragments and our approach will provide false positive match only if all the fragments gives collision (easy to see) which is quite unlikely to occur.

From the above discussion, we can conclude that, our approach is significantly faster than other methods presented and discussed above for comparison in all their aspects without compromising the accuracy. Moreover, from Figures 8 and 11, we can see that our approach becomes 1.91X slower (for BowTie, it is 2.09X and for BWA, it is 2.32X) by increasing the read length from 100 bases to 250 bases (note that the read length is increased 2.5X). This performance degradation rate is not completely accurate because of the difference in query reads in the databases (thus will give different processing execution). However, we can use this to get a rough idea about the growth rate of the performance degradation (as the database contains same number of reads and have to perform same kind of task). As we have argued previously, it is practically impossible to avoid the processing cost of the increased read length. However, we can summarise that our approach is able to bind it efficiently. By observing this bounded growth rate of performance degradation over BowTie and BWA, we can draw a conclusion that our approach will scale well for more long reads of future generation as well.

5 Conclusion

With the advent of second-generation sequencing technology, there is an increasing need of a fast and accurate read alignment method that can deal with longer short reads. In this paper, we address that need. Our experimental section shows that, for the longer short read of the current generation, our approach is an order of magnitude faster than BowTie and BWA in all aspects and this is done by keeping the accuracy intact. It can also be seen from the results that our approach can handle current generation's longer short read efficiently and also scale well for future generation's more longer short reads (by observing the bounded growth rate of performance degradation) and hence, will not become infeasible in near future (by observing the trend of increment in read length). Moreover, our approach has no upper bound in the read length like many other approaches.

References

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) "Basic local alignment search tool", *Journal of Molecular Biology*, Vol. 215, No. 3, pp. 403-410.

[2] Ansorge, W.J. (2009) "Next-generation DNA sequencing techniques", *New Biotechnology*, Vol. 25, No. 4, pp. 195–203.

[3] Aydin, F. and Dogan, G. (2013) "Development of a new integer hash function with variable length using prime number set", *Balkan Journal of Electrical & Computer Engineering*, Vol. 1, No. 1, pp. 10-14.

[4] Bentley, D.R., Balasubramanian, S., et al. (2008) "Accurate whole human genome sequencing using reversible terminator chemistry", *Nature*, Vol. 456, No. 7218, pp 53-59.

[5] Bentley, J.L. and McGeoch, C.C. (1985) "Amortized analyses of self-organizing sequential search heuristics", *Communications of the ACM*, Vol. 28, pp 404-411.

[6] Berglund, E.C., Kiialainen, A., and Syvanen, A.C. (2011) "Next-generation sequencing technologies and applications for human genetic history and forensics", *Investigative Genetics*, Vol. 2, No. 1, pp. 23.

[7] Burrows, M. and Wheeler, D. (1994) *A block sorting lossless data compression algorithm*, Technical Report 124, Digital Equipment Corporation.

[8] Chavarria-Miranda, D., Márquez, A., Nieplocha, J., Maschhoff, K., and Scherrer, C. (2008) "Early experience with out-of-core applications on the cray XMT", *IEEE International Symposium on parallel and Distributed Processing (IPDPS 2008)*, pp. 1-8.

[9] Chen, Y., Souaiaia, T., and Chen, T. (2009) "PerM: Efficient mapping of short sequencing reads with periodic full sensitive spaced seeds", *Bioinformatics*, Vol. 25, No. 19, pp. 2514-2521.

[10] Cox, A. J. (2007) *ELAND: efficient large-scale alignment of nucleotide databases*, Illumina, San Diego, USA.

[11] Deng, J., Shoemaker, R., et al. (2009) "Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming", *Nature Biotechnology*, Vol. 27, No. 4, pp 353–360.

[12] Devarakonda, K., Ziavras, S.G., and Rojas-Cessa, R. (2007) "Measuring Network Parameters with Hardware Support", *Third International Conference on Networking and Services (ICNS'07)*, pp. 2-2.

[13] Down, T.A., Rakyan, V.K. et al. (2008) "A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis", *Nature Biotechnology*, Vol. 26, No. 7, pp 779-785.

[14] Golubitsky, O. and Maslov, D. (2012) "A study of optimal 4-bit reversible Toffoli circuits and their synthesis", *IEEE Transactions on Computers*, Vol. 61, No. 9, pp. 1341–1353.

[15] Greuter, S., Parker, J., Stewart, N. and Leach, G. (2003) "Real-time procedural generation of 'pseudo infinite' cities", *Proceedings of the 1st International Conference on Computer Graphics and Interactive Techniques in Australasia and South East Asia (GRAPHITE '03)*, pp. 87–94.

[16] Huynh, T., Vlachos, M. and Rigoutsos, I. (2010) "Anchoring millions of distinct reads on the human genome within seconds", *Proceedings of the 13th*

International Conference on Extending Database Technology, pp. 252-262.

[17] Iseli, C., Ambrosini, G., Bucher, P. and Jongeneel, C. (2007) "Indexing Strategies for Rapid Searches of Short Words in Genome Sequences", *PLoS ONE*, Vol. 2, No. 6, Article e579.

[18] Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. (2007) "Genome-wide mapping of in vivo protein-DNA interactions", *Science*, Vol. 316, No. 5830, pp. 1497-1502.

[19] Kent, W. J. (2002) "BLAT–the BLAST-like alignment tool", *Genome Research*, Vol. 12, No. 4, pp. 656–664.

[20] Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009), "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome", *Genome Biology*, Vol. 10, No. 3, Article R25.

[21] Ley, T.J., Mardis, E.R. et al. (2008) "DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome", *Nature*, Vol. 456, No. 7218, pp. 66–72.

[22] Li, H. and Durbin, R. (2009) "Fast and accurate short read alignment with Burrows-Wheeler transform", *Bioinformatics*, Vol. 25, No. 14, pp. 1754–1760.

[23] Li, H., Ruan, J. and Durbin, R. (2008) "Mapping short DNA sequencing reads and calling variants using mapping quality scores", *Genome Research*, Vol. 18, No. 11, pp. 1851–1858.

[24] Li, R., Li, Y., Kristiansen, K. and Wang, J. (2008) "SOAP: short oligonucleotide alignment program", *Bioinformatics*, Vol. 24, No. 5, pp. 713–714.

[25] Li, Y., Terrell, A. and Patel, J.M. (2011) "WHAM: A High-throughput Sequence Alignment Method", *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 445–456.

[26] Rumble, S.M., Lacroute, P., Dalca, A.V., Fiume, M., Sidow, A. and Brudno, M. (2009) "SHRiMP Accurate Mapping of Short Color-space Reads", *PLoS Computational Biology*, Vol. 5, No. 5, Article e1000386.

[27] Schatz, M.C., Delcher, A.L. and Salzberg, S.L. (2010) "Assembly of large genomes using second-generation sequencing", *Genome Research*, Vol. 20, No. 9, pp. 1165-1173.

[28] Shendure, J. and Ji, H. (2008) "Next-generation DNA sequencing", *Nature Biotechnology*, Vol. 26, No. 10, pp. 1135–1145.

[29] BWA Software, available from: http://bio-bwa.sourceforge.net/ (last visited: 15 December 2012).

[30] BowTie Software, available from: http://BowTie-bio.sourceforge.net/index.shtml (last visited: 15 December 2012).

[31] Illumina Sequencing Systems, available from: http://www.illumina.com/systems/sequencing.ilmn (last visited: 17March 2013).

[32] Wikipedia - Linear Search, available from: https://en.wikipedia.org/wiki/Linear_search (last visited: 17 March 2013).

[33] NGS Alignment Programs, available from: http://lh3lh3.users.sourceforge.net/NGSalign.shtml (last visited: 15 February 2013).

[34] PerM Software, available from: http://code.google.com/p/perm (last visited: 1 February 2013).

[35] SOAP Software, available from: http://soap.genomics.org.cn/soap1/ (last visited: 1 February 2013).

[36] WHAM Software, available from: http://research.cs.wisc.edu/wham/ (last visited: 15 December 2012).

Intuitionistic Fuzzy Jensen-Rényi Divergence: Applications to Multiple-Attribute Decision Making

Rajkumar Verma and Bhu Dev Sharma
Department of Mathematics
Jaypee Institute of Information Technology (Deemed University)
Noida-201307, U.P., India
E-mail: rkver83@gmail.com, bhudev.sharma@jiit.ac.in

Keywords: intuitionistic fuzzy set, Rényi entropy, Jensen-Shannon divergence, Jensen- Rényi divergence, MADM

Received: May 17, 2013

Vagueness in the scientific studies presents a challenging dimension. Intuitionistic fuzzy set theory has emerged as a tool for its characterization. There is need to associate measures which can measure vagueness and differences in the underlying characterizing IFSs. In the present paper we introduce an information theoretic divergence measure, called intuitionistic fuzzy Jensen-Rényi divergence. It is a difference measure in the setting of intuitionistic fuzzy set theory, involving parameters that provide flexibility and choice. The strength of the new measure lies in its properties and applications. An approach to multiple-attribute decision making based on intuitionistic fuzzy Jensen-Rényi divergence is proposed. A numerical example illustrates the application of the new measure and the role of various parameters therein to multipleattribute decision making problem formulated in terms of intuitionistic fuzzy sets.

Povzetek: Razvita je nova verzija intuitivne mehke logike za uporabo v procesu odločanja.

1 Introduction

In probability theory and statistics, divergence measures are commonly used for measuring the differences between two probability distributions [13 and 22]. Kullback-Leibler [13] divergence is the well known such information theoretic divergence. Another important information theoretic divergence measure is the Jensen-Shannon divergence (JSD) [22] which has attracted quite some attention. It has been shown that the square root of JSD turns out to be a metric [9], satisfying (i) non-negativity (ii) (minimal) zero value only for identical distributions (iii) symmetric and (iv) satisfying triangular inequality, i.e. it is bounded from below and from above in terms of the norms of the distributions. However it may be mentioned that JSD itself is not a metric. It satisfies the first three axioms, and not the triangular inequality. These divergence measures have been applied in several disciplines like signal processing, pattern recognition, finance, economics etc.

Some generalizations of Jensen-Shannon divergence measure have been studied in the last couple of years. For instance, He et al. [10] proposed a one parametric generalization of JSD based on Rényi's entropy function [21], called Jensen-Rényi divergence and used it in image registration.

Other than probabilistic, there are vague/fuzzy phenomena. These are best characterized in terms of 'fuzzy sets', and their generalizations. The theory of fuzzy sets proposed by Zadeh [32] in 1965 addresses these situations and has found applications in various fields. In fuzzy set theory, the membership of an element is a single value lying between zero and one, where the degree of non-membership is just automatically equal to one minus the degree of membership.

As a generalization of Zadeh's fuzzy sets, Atanassov [1, 2], introduced intuitionistic fuzzy sets. In their general setting, these involve three non-negative functions expressing the degree of membership, the degree of non-membership, and hesitancy, their sum being one. These considerations imbue IFSs with inbuilt structure to consider varieties of factors responsible of vagueness in the phenomena. IFSs have been applied in many practically uncertain/vague situations, such as decision making [3, 4, 8, 14, 16-18, 20, 25, 27-30 and 33] medical diagnosis [5, 24] and pattern recognition [6, 11, 12, 19 and 24] etc. Atanassov [2] and Szmidt and Kacprzyk [26] suggested some methods for measuring distance/difference between two intuitionistic fuzzy sets. Their measures are generalizations of the well known Hamming and Euclidean distances. Dengfeng and Chutian [6] and Dengfeng [7] proposed some other similarity and dissimilarity measures for measuring differences between pairs of intuitionistic fuzzy sets. In addition, Yanhong et al. [31] undertook a comparative analysis of these similarity measures. Recently, Verma and Sharma [25] proposed a generalized intuitionistic fuzzy divergence and studied its applications to multi criteria decision making.

In this paper, we extend the idea of Jensen-Rényi divergence to intuitionistic fuzzy sets and propose a new divergence measure, called *intuitionistic fuzzy Jensen-Rényi divergence* (IFJRD) to measure the difference between two IFSs. After studying its properties, we give

an example of its applications in multiple-attribute decision making based on intuitionistic fuzzy information. The paper is organized as follows: In Section 2 some basic definitions related to probability theory, fuzzy set theory and intuitionistic fuzzy set theory are briefly given. In Section 3, the intuitionistic fuzzy Jensen-Rényi divergence (IFJRD) between two intuitionistic fuzzy sets is proposed. Some of its basic properties are analysed there, along with the limiting case. In Section 4 some more properties of the proposed measure are studied. In Section 5 application of proposed *intuitionistic fuzzy Jensen-Rényi divergence* measure to multiple-attribute decision making are illustrated and our conclusions are also presented here.

2 Preliminaries

We start with probabilistic background. We denote the set of *n*-complete $(n \geq 2)$ probability distributions by

$$\Gamma_n = \left\{ P = (p_1, p_2, ..., p_n) : p_i \geq 0, \sum_{i=1}^{n} p_i = 1 \right\}. \quad (1)$$

For a probability distribution

$$P = (p_1, p_2, ..., p_n) \in \Gamma_n,$$

the well known Shannon's entropy [23], is defined as

$$H(P) = -\sum_{i=1}^{n} p_i \log p_i . \quad (2)$$

Various generalized entropies have been introduced in the literature taking the Shannon entropy as basic and have found applications in various disciplines such as economics, statistics, information processing and computing etc.

A generalizations of Shannon's entropy introduced by Rényi's [21], Rényi's entropy of order α, is given by

$$H_\alpha(P) = \frac{1}{1-\alpha} \log\left(\sum_{i=1}^{n} p_i^\alpha\right), \quad \alpha \neq 1, \alpha > 0 . \quad (3)$$

For $\alpha \in (0,1)$, it is easy to see that $H_\alpha(P)$ is a concave function of *P*, and in the limiting case $\alpha \to 1$, it tends to Shannon's entropy. It can also be easily verified that $H_\alpha(P)$ is a non-increasing function of $\alpha \in (0,1)$ and thus

$$H_\alpha(P) \geq H(P) \quad \forall\ \alpha \in (0,1) \quad (4)$$

In sequel, we will restrict $\alpha \in (0,1)$, unless otherwise specified and will use base 2 for the logarithm.

Next, we mention *Jensen-Shannon divergence* [15]. Let $\lambda_1, \lambda_2 \geq 0$, $\lambda_1 + \lambda_2 = 1$ be the weights of two probability distributions $P, Q \in \Gamma_n$, respectively. Then the Jensen-Shannon divergence, is defined as

$$JS_\lambda(P,Q) = H(\lambda_1 P + \lambda_2 Q) - \lambda_1 H(P) - \lambda_2 H(Q) . \quad (5)$$

Since $H(P)$ is a concave function, according to Jensen's inequality, $JS_\lambda(P,Q)$ is nonnegative and vanishes when $P = Q$. One of the major features of the Jensen-Shannon divergence is that we can assign different weights to the probability distributions involved according to their importance. This is particularly useful in the study of decision problems.

A generalization of the above concept is the *Jensen-Rényi divergence* proposed by He [10], given by

$$JR_{\lambda,\alpha}(P,Q) = H_\alpha(\lambda_1 P + \lambda_2 Q) - \lambda_1 H_\alpha(P) - \lambda_2 H_\alpha(Q), \quad \alpha \in (0,1) \quad (6)$$

where $H_\alpha(P)$ is Rényi's entropy, and $\lambda = (\lambda_1, \lambda_2)$ is the weight vector, with $\lambda_1, \lambda_2 \geq 0$, $\lambda_1 + \lambda_2 = 1$, as before.

***Properties of Jensen-Rényi Divergence*:** Briefly we note some simple properties:

i. $JR_{\lambda,\alpha}(P,Q)$ is nonnegative and is equal to zero when $P = Q$.

ii. For $\alpha \in (0,1)$, $JR_{\lambda,\alpha}(P,Q)$ is a convex function of P and Q.

iii. $JR_{\lambda,\alpha}(P,Q)$, achieves its maximum value when P and Q are degenerate distributions.

The Jensen-Shannon divergence (5) is a limiting case of $JR_{\lambda,\alpha}(P,Q)$ when $\alpha \to 1$.

Definition 1. ***Fuzzy Set*** [32]**:** A fuzzy set $\tilde{A}$ in a finite universe of discourse $X = \{x_1, x_2, ..., x_n\}$ is defined as

$$\tilde{A} = \left\{ \langle x, \mu_{\tilde{A}}(x) \rangle \,\middle|\, x \in X \right\}, \quad (7)$$

where $\mu_{\tilde{A}}(x): X \to [0,1]$ is measure of belongingness or degree of membership of an element $x \in X$ to $\tilde{A}$.

Thus, automatically the measure of non-belongingness of $x \in X$ to $\tilde{A}$ is $(1 - \mu_{\tilde{A}}(x))$.

Atanassov [1, 2] introduced following generalization of fuzzy sets, called intuitionistic fuzzy sets.

Definition 2. ***Intuitionistic Fuzzy Set*** [1, 2]**:** An intuitionistic fuzzy set A in a finite universe of discourse $X = \{x_1, x_2, ..., x_n\}$ is defined as

$$A = \left\{ \langle x, \mu_A(x), \nu_A(x) \rangle \,\middle|\, x \in X \right\}, \quad (8)$$

where $\mu_A: X \to [0,1]$ and $\nu_A: X \to [0,1]$ with the condition $0 \leq \mu_A(x) + \nu_A(x) \leq 1$. For each $x \in X$, the numbers $\mu_A(x)$ and $\nu_A(x)$ denote the degree of membership and degree of non-membership of x to A respectively.

Further, we call $\pi_A(x) = 1 - \mu_A(x) - \nu_A(x)$, the degree of hesitance or the intuitionistic index of $x \in X$ to A.

Obviously, when $\pi_A(x)=0$, i.e., $\nu_A(x)=1-\mu_A(x)$ for every $x\in X$, then the IFS A becomes a fuzzy set. Thus, FSs are the special cases of IFSs.

Definition 3: Let $IFS(X)$ denote the family of all IFSs defined in the universe X, and let $A,B\in IFS(X)$ be given by

$$A=\{\langle x,\mu_A(x),\nu_A(x)\rangle | x\in X\},$$
$$B=\{\langle x,\mu_B(x),\nu_B(x)\rangle | x\in X\}.$$

These being sets, Atanassov further defined set operations on $IFS(X)$ as follows:

(i) $A\subseteq B$ iff $\mu_A(x)\le\mu_B(x)$ and $\nu_A(x)\ge\nu_B(x)\ \forall\ x\in X$;

(ii) $A=B$ iff $A\subseteq B$ and $B\subseteq A$;

(iii) $A^C=\{\langle x,\nu_A(x),\mu_A(x)\rangle \mid x\in X\}$;

(iv) $A\cup B=\left\{\left\langle x,\begin{matrix}\max(\mu_A(x),\mu_B(x)),\\ \min(\nu_A(x),\nu_B(x))\end{matrix}\right\rangle | x\in X\right\}$;

(v) $A\cap B=\left\{\left\langle x,\begin{matrix}\min(\mu_A(x),\mu_B(x)),\\ \max(\nu_A(x),\nu_B(x))\end{matrix}\right\rangle | x\in X\right\}$.

Extending the idea from probabilistic to intuitionistic phenomena, in the next section, we propose a divergence measure called *'Intuitionistic Fuzzy Jensen-Rényi Divergence'* (IFJRD) on intuitionistic fuzzy sets to quantify the difference between two intuitionistic fuzzy sets and discuss its limiting case.

3 Intuitionistic Fuzzy Jensen-Rényi Divergence (IFJRD)

Single element universe: First, let A and B be two intuitionistic fuzzy sets defined on a single element universal set $X=\{x\}$.

Precisely speaking, we have:

$$A=(\mu_A(x),\nu_A(x),\pi_A(x)),$$

and

$$B=(\mu_B(x),\nu_B(x),\pi_B(x)),$$

where

$$\mu_A(x)+\nu_A(x)+\pi_A(x)=1,$$

and

$$\mu_B(x)+\nu_B(x)+\pi_B(x)=1,$$

with

$$0\le\mu_A(x),\nu_A(x),\pi_A(x),\ \mu_B(x),\nu_B(x),\pi_B(x)\le 1.$$

Regarding (μ_A,ν_A,π_A) and (μ_B,ν_B,π_B) as two probability distributions, in analogy of (6), we define the intuitionistic fuzzy Jensen-Rényi divergence measure between IFSs A and B, as

$$JR^*_{\lambda,\alpha}(A,B)=H_\alpha(\lambda_1A+\lambda_2B)-\lambda_1H_\alpha(A)-\lambda_2H_\alpha(B)', \tag{9}$$

where $H_\alpha(\bullet)$ is Rényi's entropy for intuitionistic fuzzy set $(\bullet)$, $\alpha\in(0,1)$, $\lambda_1+\lambda_2=1$, $\lambda_1,\lambda_2\ge 0$, and

$$\lambda_1A+\lambda_2B=\begin{pmatrix}\lambda_1\mu_A(x)+\lambda_2\mu_B(x),\\ \lambda_1\nu_A(x)+\lambda_2\nu_B(x),\\ \lambda_1\pi_A(x)+\lambda_2\pi_B(x)\end{pmatrix}.$$

That is

$$JR^*_{\lambda,\alpha}(A,B)=\frac{1}{(1-\alpha)}\begin{bmatrix}\log\left\{\begin{matrix}(\lambda_1\mu_A(x)+\lambda_2\mu_B(x))^\alpha\\ +(\lambda_1\nu_A(x)+\lambda_2\nu_B(x))^\alpha\\ +(\lambda_1\pi_A(x)+\lambda_2\pi_B(x))^\alpha\end{matrix}\right\}\\ -\lambda_1\log\left\{\begin{matrix}(\mu_A(x))^\alpha+(\nu_A(x))^\alpha\\ +(\pi_A(x))^\alpha\end{matrix}\right\}\\ -\lambda_2\log\left\{\begin{matrix}(\mu_B(x))^\alpha+(\nu_B(x))^\alpha\\ +(\pi_B(x))^\alpha\end{matrix}\right\}\end{bmatrix}, \tag{10}$$

where $\alpha\in(0,1)$.

Next, in theorem below we study properties of $JR^*_{\lambda,\alpha}(A,B)$ defined in (10).

Theorem1: For $A,B\in IFS(X)$, $JR^*_{\lambda,\alpha}(A,B)$ satisfies the following properties:

i. $JR^*_{\lambda,\alpha}(A,B)\ge 0$, with equality if and only if $A=B$.

ii. $0\le JR^*_{\lambda,\alpha}(A,B)\le 1$.

iii. For three *IFSs* A,B,C in X and $A\subseteq B\subseteq C$,

$$JR^*_{\lambda,\alpha}(A,B)\le JR^*_{\lambda,\alpha}(A,C),$$

and

$$JR^*_{\lambda,\alpha}(B,C)\le JR^*_{\lambda,\alpha}(A,C).$$

Proof: (i) The result directly follows from Jensen's inequality.

(ii) Since $JR^*_{\lambda,\alpha}(A,B)$ is convex for $\alpha\in(0,1)$, refer Proposition 1 of He et al. [10], therefore, for $\alpha\in(0,1)$, $JR^*_{\lambda,\alpha}(A,B)$ increases as $\|A-B\|_1$ increases, where

$$\|A-B\|_1=|\mu_A(x)-\mu_B(x)|+|\nu_A(x)-\nu_B(x)|+|\pi_A(x)-\pi_B(x)|. \tag{11}$$

Thus, $JR^*_{\lambda,\alpha}(A,B)\ \forall\ \alpha\in(0,1)$, attains its maximum for following degenerate cases:

$$A=(1,0,0),B=(0,1,0)\text{ or }A=(0,1,0),B=(1,0,0)\text{ or }A=(0,0,1),B=(0,1,0).$$

This gives

$$0\le JR^*_{\lambda,\alpha}(A,B)\le 1.$$

(iii) For $A, B, C \in IFS(X)$,

$$\| A-B \|_1 \le \| A-C \|_1$$

and $\| B-C \|_1 \le \| A-C \|_1$, if $A \subseteq B \subseteq C$.

Thus,

$$JR^*_{\lambda,\alpha}(A,B) \le JR^*_{\lambda,\alpha}(A,C)$$

and $JR^*_{\lambda,\alpha}(B,C) \le JR^*_{\lambda,\alpha}(A,C) \quad \forall \ \alpha \in (0,1)$. (12)

This proves the theorem.

Limiting case: When $\alpha \to 1$ and $\lambda_1 = \lambda_2 = \frac{1}{2}$, then measure (10) reduces to *J*-divergence on intuitionistic fuzzy sets proposed by Hung and Yang [11] as

$$J(A,B) = \left[\begin{array}{l} -\left(\begin{array}{l} \left(\frac{\mu_A(x)+\mu_B(x)}{2}\right)\log\left(\frac{\mu_A(x)+\mu_B(x)}{2}\right) \\ +\left(\frac{\nu_A(x)+\nu_B(x)}{2}\right)\log\left(\frac{\nu_A(x)+\nu_B(x)}{2}\right) \\ +\left(\frac{\pi_A(x)+\pi_B(x)}{2}\right)\log\left(\frac{\pi_A(x)+\pi_B(x)}{2}\right) \end{array} \right) \\ +\left(\frac{\left(\begin{array}{l}\mu_A(x)\log\mu_A(x)\\ +\nu_A(x)\log\nu_A(x)\\ +\pi_A(x)\log\pi_A(x)\end{array}\right)+\left(\begin{array}{l}\mu_B(x)\log\mu_B(x)\\ +\nu_B(x)\log\nu_B(x)\\ +\pi_B(x)\log\pi_B(x)\end{array}\right)}{2} \right) \end{array} \right]. \quad (13)$$

Definition 4: $JR_{\lambda,\alpha}(A,B)$ **on Finite Universe:** Previously, we considered single element universe set. The idea can be extended to any finite universe set. If *A* and *B* are two IFSs defined in finite universe of discourse $X = \{x_1, x_2, ..., x_n\}$, then, we define, the associated intuitionistic fuzzy Jensen-Rényi divergence by

$$JR_{\lambda,\alpha}(A,B) = \frac{1}{n}\sum_{i=1}^{n} JR^*_{\lambda,\alpha}(A(x_i), B(x_i)) \quad (14)$$

where $A(x_i) = \{(x_i, \mu_A(x_i), \nu_A(x_i), \pi_A(x_i))\}$,

and $B(x_i) = \{(x_i, \mu_B(x_i), \nu_B(x_i), \pi_B(x_i))\}$.

In the next section, we study several properties of $JR_{\lambda,\alpha}(A,B)$. While proving these properties, we consider separation of X into two parts X_1 and X_2, such that

$$X_1 = \{x_i \mid x_i \in X,\ A(x_i) \subseteq B(x_i)\}, \quad (15)$$

$$X_2 = \{x_i \mid x_i \in X,\ A(x_i) \supseteq B(x_i)\}. \quad (16)$$

Further it may be noted that for all $x_i \in X_1$,

$$\mu_A(x_i) \le \mu_B(x_i) \text{ and } \nu_A(x_i) \ge \nu_B(x_i),$$

as also for $\forall \ x_i \in X_2$,

$$\mu_A(x_i) \ge \mu_B(x_i) \text{ and } \nu_A(x_i) \le \nu_B(x_i).$$

4 Properties of intuitionistic fuzzy Jensen- Rényi divergence measure

The measure $JR_{\lambda,\alpha}(A,B)$ defined in (10) has the following properties:

Theorem 2: For $A, B \in IFS(X)$,

(i) $JR_{\lambda,\alpha}(A \cup B, A \cap B) = JR_{\lambda,\alpha}(A,B)$,

(ii) $JR_{\lambda,\alpha}(A \cap B, A \cup B) = JR_{\lambda,\alpha}(B,A)$.

Proof: We prove (i) only, (ii) can be proved analogously.

(i) From definition in (10), we have:

$$JR_{\lambda,\alpha}(A \cup B, A \cap B)$$

$$= \frac{1}{n(1-\alpha)}\sum_{i=1}^{n}\left[\begin{array}{l} \log\left\{ \begin{array}{l} (\lambda_1\mu_{A\cup B}(x_i)+\lambda_2\mu_{A\cap B}(x_i))^\alpha \\ +(\lambda_1\nu_{A\cap B}(x_i)+\lambda_2\nu_{A\cup B}(x_i))^\alpha \\ +\left(\begin{array}{l}\lambda_1(1-\mu_{A\cup B}(x_i)-\nu_{A\cap B}(x_i))\\ +\lambda_2(1-\mu_{A\cap B}(x_i)-\nu_{A\cup B}(x_i))\end{array}\right)^\alpha \end{array} \right\} \\ -\lambda_1\log\left\{ \begin{array}{l} (\mu_{A\cup B}(x_i))^\alpha+(\nu_{A\cap B}(x_i))^\alpha \\ +(1-\mu_{A\cup B}(x_i)-\nu_{A\cap B}(x_i))^\alpha \end{array} \right\} \\ -\lambda_2\log\left\{ \begin{array}{l} (\mu_{A\cap B}(x_i))^\alpha+(\nu_{A\cup B}(x_i))^\alpha \\ +(1-\mu_{A\cap B}(x_i)-\nu_{A\cup B}(x_i))^\alpha \end{array} \right\} \end{array} \right]$$

$$= \frac{1}{n(1-\alpha)}\left[\sum_{x_i \in X_1}\left\{ \begin{array}{l} \log\left\{ \begin{array}{l} (\lambda_1\mu_B(x_i)+\lambda_2\mu_A(x_i))^\alpha \\ +(\lambda_1\nu_B(x_i)+\lambda_2\nu_A(x_i))^\alpha \\ +\left(\begin{array}{l}\lambda_1(1-\mu_B(x_i)-\nu_B(x_i))\\ +\lambda_2(1-\mu_A(x_i)-\nu_A(x_i))\end{array}\right)^\alpha \end{array} \right\} \\ -\lambda_1\log\left\{ \begin{array}{l} (\mu_B(x_i))^\alpha+(\nu_B(x_i))^\alpha \\ +(1-\mu_B(x_i)-\nu_B(x_i))^\alpha \end{array} \right\} \\ -\lambda_2\log\left\{ \begin{array}{l} (\mu_A(x_i))^\alpha+(\nu_A(x_i))^\alpha \\ +(1-\mu_A(x_i)-\nu_A(x_i))^\alpha \end{array} \right\} \end{array} \right\} \right.$$

$$\left. + \sum_{x_i \in X_2}\left\{ \begin{array}{l} \log\left\{ \begin{array}{l} (\lambda_1\mu_A(x_i)+\lambda_2\mu_B(x_i))^\alpha \\ +(\lambda_1\nu_A(x_i)+\lambda_2\nu_B(x_i))^\alpha \\ +\left(\begin{array}{l}\lambda_1(1-\mu_A(x_i)-\nu_A(x_i))\\ +\lambda_2(1-\mu_B(x_i)-\nu_B(x_i))\end{array}\right)^\alpha \end{array} \right\} \\ -\lambda_1\log\left\{ \begin{array}{l} (\mu_A(x_i))^\alpha+(\nu_A(x_i))^\alpha \\ +(1-\mu_A(x_i)-\nu_A(x_i))^\alpha \end{array} \right\} \\ -\lambda_2\log\left\{ \begin{array}{l} (\mu_B(x_i))^\alpha+(\nu_B(x_i))^\alpha \\ +(1-\mu_B(x_i)-\nu_B(x_i))^\alpha \end{array} \right\} \end{array} \right\} \right]$$

$$= JR_{\lambda,\alpha}(A,B).$$

This proves the theorem.

Theorem 3: For $A, B \in IFS(X)$,

(i) $JR_{\lambda,\alpha}(A, A\cup B)+JR_{\lambda,\alpha}(A, A\cap B)=JR_{\lambda,\alpha}(A, B)$,

(ii) $JR_{\lambda,\alpha}(B, A\cup B)+JR_{\lambda,\alpha}(B, A\cap B)=JR_{\lambda,\alpha}(B, A)$.

Proof: In the following, we prove only (i), (ii) can be proved analogously.

(i) Using definition in (10), we first have

$$JR_{\lambda,\alpha}(A, A\cup B)$$

$$=\frac{1}{n(1-\alpha)}\sum_{i=1}^{n}\left[\begin{array}{l}\log\left\{\begin{array}{l}(\lambda_1\mu_A(x_i)+\lambda_2\mu_{A\cup B}(x_i))^{\alpha}\\+(\lambda_1\nu_A(x_i)+\lambda_2\nu_{A\cap B}(x_i))^{\alpha}\\+\left(\begin{array}{l}\lambda_1(1-\mu_A(x_i)-\nu_A(x_i))\\+\lambda_2(1-\mu_{A\cup B}(x_i)-\nu_{A\cap B}(x_i))\end{array}\right)^{\alpha}\end{array}\right\}\\-\lambda_1\log\left\{\begin{array}{l}(\mu_A(x_i))^{\alpha}+(\nu_A(x_i))^{\alpha}\\+(1-\mu_A(x_i)-\nu_A(x_i))^{\alpha}\end{array}\right\}\\-\lambda_2\log\left\{\begin{array}{l}(\mu_{A\cup B}(x_i))^{\alpha}+(\nu_{A\cap B}(x_i))^{\alpha}\\+(1-\mu_{A\cup B}(x_i)-\nu_{A\cap B}(x_i))^{\alpha}\end{array}\right\}\end{array}\right]$$

$$=\frac{1}{n(1-\alpha)}\left[\sum_{x_i\in X_1}\left\{\begin{array}{l}\log\left\{\begin{array}{l}(\lambda_1\mu_A(x_i)+\lambda_2\mu_B(x_i))^{\alpha}\\+(\lambda_1\nu_A(x_i)+\lambda_2\nu_B(x_i))^{\alpha}\\+\left(\begin{array}{l}\lambda_1(1-\mu_A(x_i)-\nu_A(x_i))\\+\lambda_2(1-\mu_B(x_i)-\nu_B(x_i))\end{array}\right)^{\alpha}\end{array}\right\}\\-\lambda_1\log\left\{\begin{array}{l}(\mu_A(x_i))^{\alpha}+(\nu_A(x_i))^{\alpha}\\+(1-\mu_A(x_i)-\nu_A(x_i))^{\alpha}\end{array}\right\}\\-\lambda_2\log\left\{\begin{array}{l}(\mu_B(x_i))^{\alpha}+(\nu_B(x_i))^{\alpha}\\+(1-\mu_B(x_i)-\nu_B(x_i))^{\alpha}\end{array}\right\}\end{array}\right\}\right.$$

$$\left.+\sum_{x_i\in X_2}\left\{\begin{array}{l}\log\left\{\begin{array}{l}(\lambda_1\mu_A(x_i)+\lambda_2\mu_A(x_i))^{\alpha}\\+(\lambda_1\nu_A(x_i)+\lambda_2\nu_A(x_i))^{\alpha}\\+\left(\begin{array}{l}\lambda_1(1-\mu_A(x_i)-\nu_A(x_i))\\+\lambda_2(1-\mu_A(x_i)-\nu_A(x_i))\end{array}\right)^{\alpha}\end{array}\right\}\\-\lambda_1\log\left\{\begin{array}{l}(\mu_A(x_i))^{\alpha}+(\nu_A(x_i))^{\alpha}\\+(1-\mu_A(x_i)-\nu_A(x_i))^{\alpha}\end{array}\right\}\\-\lambda_2\log\left\{\begin{array}{l}(\mu_A(x_i))^{\alpha}+(\nu_A(x_i))^{\alpha}\\+(1-\mu_A(x_i)-\nu_A(x_i))^{\alpha}\end{array}\right\}\end{array}\right\}\right]$$

$$=\frac{1}{n(1-\alpha)}\left[\sum_{x_i\in X_1}\left\{\begin{array}{l}\log\left\{\begin{array}{l}(\lambda_1\mu_A(x_i)+\lambda_2\mu_B(x_i))^{\alpha}\\+(\lambda_1\nu_A(x_i)+\lambda_2\nu_B(x_i))^{\alpha}\\+\left(\begin{array}{l}\lambda_1(1-\mu_A(x_i)-\nu_A(x_i))\\+\lambda_2(1-\mu_B(x_i)-\nu_B(x_i))\end{array}\right)^{\alpha}\end{array}\right\}\\-\lambda_1\log\left\{\begin{array}{l}(\mu_A(x_i))^{\alpha}+(\nu_A(x_i))^{\alpha}\\+(1-\mu_A(x_i)-\nu_A(x_i))^{\alpha}\end{array}\right\}\\-\lambda_2\log\left\{\begin{array}{l}(\mu_B(x_i))^{\alpha}+(\nu_B(x_i))^{\alpha}\\+(1-\mu_B(x_i)-\nu_B(x_i))^{\alpha}\end{array}\right\}\end{array}\right\}\right] \quad (17)$$

Next, again from definition in (10), we have

$$JR_{\lambda,\alpha}(A, A\cap B)$$

$$=\frac{1}{n(1-\alpha)}\sum_{i=1}^{n}\left[\begin{array}{l}\log\left\{\begin{array}{l}(\lambda_1\mu_A(x_i)+\lambda_2\mu_{A\cap B}(x_i))^{\alpha}\\+(\lambda_1\nu_A(x_i)+\lambda_2\nu_{A\cup B}(x_i))^{\alpha}\\+\left(\begin{array}{l}\lambda_1(1-\mu_A(x_i)-\nu_A(x_i))\\+\lambda_2(1-\mu_{A\cap B}(x_i)-\nu_{A\cup B}(x_i))\end{array}\right)^{\alpha}\end{array}\right\}\\-\lambda_1\log\left\{\begin{array}{l}(\mu_A(x_i))^{\alpha}+(\nu_A(x_i))^{\alpha}\\+(1-\mu_A(x_i)-\nu_A(x_i))^{\alpha}\end{array}\right\}\\-\lambda_2\log\left\{\begin{array}{l}(\mu_{A\cap B}(x_i))^{\alpha}+(\nu_{A\cup B}(x_i))^{\alpha}\\+(1-\mu_{A\cap B}(x_i)-\nu_{A\cup B}(x_i))^{\alpha}\end{array}\right\}\end{array}\right]$$

$$=\frac{1}{n(1-\alpha)}\left[\sum_{x_i\in X_1}\left\{\begin{array}{l}\log\left\{\begin{array}{l}(\lambda_1\mu_A(x_i)+\lambda_2\mu_A(x_i))^{\alpha}\\+(\lambda_1\nu_A(x_i)+\lambda_2\nu_A(x_i))^{\alpha}\\+\left(\begin{array}{l}\lambda_1(1-\mu_A(x_i)-\nu_A(x_i))\\+\lambda_2(1-\mu_A(x_i)-\nu_A(x_i))\end{array}\right)^{\alpha}\end{array}\right\}\\-\lambda_1\log\left\{\begin{array}{l}(\mu_A(x_i))^{\alpha}+(\nu_A(x_i))^{\alpha}\\+(1-\mu_A(x_i)-\nu_A(x_i))^{\alpha}\end{array}\right\}\\-\lambda_2\log\left\{\begin{array}{l}(\mu_A(x_i))^{\alpha}+(\nu_A(x_i))^{\alpha}\\+(1-\mu_A(x_i)-\nu_A(x_i))^{\alpha}\end{array}\right\}\end{array}\right\}\right.$$

$$\left.+\sum_{x_i\in X_2}\left\{\begin{array}{l}\log\left\{\begin{array}{l}(\lambda_1\mu_A(x_i)+\lambda_2\mu_B(x_i))^{\alpha}\\+(\lambda_1\nu_A(x_i)+\lambda_2\nu_B(x_i))^{\alpha}\\+\left(\begin{array}{l}\lambda_1(1-\mu_A(x_i)-\nu_A(x_i))\\+\lambda_2(1-\mu_B(x_i)-\nu_B(x_i))\end{array}\right)^{\alpha}\end{array}\right\}\\-\lambda_1\log\left\{\begin{array}{l}(\mu_A(x_i))^{\alpha}+(\nu_A(x_i))^{\alpha}\\+(1-\mu_A(x_i)-\nu_A(x_i))^{\alpha}\end{array}\right\}\\-\lambda_2\log\left\{\begin{array}{l}(\mu_B(x_i))^{\alpha}+(\nu_B(x_i))^{\alpha}\\+(1-\mu_B(x_i)-\nu_B(x_i))^{\alpha}\end{array}\right\}\end{array}\right\}\right]$$

$$= \frac{1}{n(1-\alpha)} \left[\sum_{x_i \in X_2} \left\{ \begin{array}{l} \log \left\{ \begin{array}{l} (\lambda_1 \mu_A(x_i) + \lambda_2 \mu_B(x_i))^\alpha \\ + (\lambda_1 \nu_A(x_i) + \lambda_2 \nu_B(x_i))^\alpha \\ + \left(\begin{array}{l} \lambda_1 (1 - \mu_A(x_i) - \nu_A(x_i)) \\ + \lambda_2 (1 - \mu_B(x_i) - \nu_B(x_i)) \end{array} \right)^\alpha \end{array} \right\} \\ - \lambda_1 \log \left\{ \begin{array}{l} (\mu_A(x_i))^\alpha + (\nu_A(x_i))^\alpha \\ + (1 - \mu_A(x_i) - \nu_A(x_i))^\alpha \end{array} \right\} \\ - \lambda_2 \log \left\{ \begin{array}{l} (\mu_B(x_i))^\alpha + (\nu_B(x_i))^\alpha \\ + (1 - \mu_B(x_i) - \nu_B(x_i))^\alpha \end{array} \right\} \end{array} \right\} \right] \quad (18)$$

Adding (17) and (18), we get the result.

Theorem 4: For $A, B, C \in IFS(X)$,

(i) $JR_{\lambda,\alpha}(A \cup B, C) \leq JR_{\lambda,\alpha}(A, C) + JR_{\lambda,\alpha}(B, C)$;

(ii) $JR_{\lambda,\alpha}(A \cap B, C) \leq JR_{\lambda,\alpha}(A, C) + JR_{\lambda,\alpha}(B, C)$;

Proof: We prove (i) only, (ii) can be proved analogously.

(i) Let us consider the expression

$$JR_{\lambda,\alpha}(A, C) + JR_{\lambda,\alpha}(B, C) - JR_{\lambda,\alpha}(A \cup B, C) \quad (19)$$

$$= \frac{1}{n(1-\alpha)} \sum_{i=1}^{n} \left[\begin{array}{l} \log \left\{ \begin{array}{l} (\lambda_1 \mu_A(x_i) + \lambda_2 \mu_C(x_i))^\alpha \\ + (\lambda_1 \nu_A(x_i) + \lambda_2 \nu_C(x_i))^\alpha \\ + \left(\begin{array}{l} \lambda_1 (1 - \mu_A(x_i) - \nu_A(x_i)) \\ + \lambda_2 (1 - \mu_C(x_i) - \nu_C(x_i)) \end{array} \right)^\alpha \end{array} \right\} \\ - \lambda_1 \log \left\{ \begin{array}{l} (\mu_A(x_i))^\alpha + (\nu_A(x_i))^\alpha \\ + (1 - \mu_A(x_i) - \nu_A(x_i))^\alpha \end{array} \right\} \\ - \lambda_2 \log \left\{ \begin{array}{l} (\mu_C(x_i))^\alpha + (\nu_C(x_i))^\alpha \\ + ((1 - \mu_C(x_i) - \nu_C(x_i)))^\alpha \end{array} \right\} \end{array} \right]$$

$$+ \frac{1}{n(1-\alpha)} \sum_{i=1}^{n} \left[\begin{array}{l} \log \left\{ \begin{array}{l} (\lambda_1 \mu_B(x_i) + \lambda_2 \mu_C(x_i))^\alpha \\ + (\lambda_1 \nu_B(x_i) + \lambda_2 \nu_C(x_i))^\alpha \\ + \left(\begin{array}{l} \lambda_1 (1 - \mu_B(x_i) - \nu_B(x_i)) \\ + \lambda_2 (1 - \mu_C(x_i) - \nu_C(x_i)) \end{array} \right)^\alpha \end{array} \right\} \\ - \lambda_1 \log \left\{ \begin{array}{l} (\mu_B(x_i))^\alpha + (\nu_B(x_i))^\alpha \\ + (1 - \mu_B(x_i) - \nu_B(x_i))^\alpha \end{array} \right\} \\ - \lambda_2 \log \left\{ \begin{array}{l} (\mu_C(x_i))^\alpha + (\nu_C(x_i))^\alpha \\ + ((1 - \mu_C(x_i) - \nu_C(x_i)))^\alpha \end{array} \right\} \end{array} \right]$$

$$- \frac{1}{n(1-\alpha)} \sum_{i=1}^{n} \left[\begin{array}{l} \log \left\{ \begin{array}{l} (\lambda_1 \mu_{A \cup B}(x_i) + \lambda_2 \mu_C(x_i))^\alpha \\ + (\lambda_1 \nu_{A \cap B}(x_i) + \lambda_2 \nu_C(x_i))^\alpha \\ + \left(\begin{array}{l} \lambda_1 (1 - \mu_{A \cup B}(x_i) - \nu_{A \cap B}(x_i)) \\ + \lambda_2 (1 - \mu_C(x_i) - \nu_C(x_i)) \end{array} \right)^\alpha \end{array} \right\} \\ - \lambda_1 \log \left\{ \begin{array}{l} (\mu_{A \cup B}(x_i))^\alpha + (\nu_{A \cap B}(x_i))^\alpha \\ + (1 - \mu_{A \cup B}(x_i) - \nu_{A \cap B}(x_i))^\alpha \end{array} \right\} \\ - \lambda_2 \log \left\{ \begin{array}{l} (\mu_C(x_i))^\alpha + (\nu_C(x_i))^\alpha \\ + ((1 - \mu_C(x_i) - \nu_C(x_i)))^\alpha \end{array} \right\} \end{array} \right]$$

$$= \frac{1}{n(1-\alpha)} \left[\sum_{x_i \in X_2} \left\{ \begin{array}{l} \log \left\{ \begin{array}{l} (\lambda_1 \mu_B(x_i) + \lambda_2 \mu_C(x_i))^\alpha \\ + (\lambda_1 \nu_B(x_i) + \lambda_2 \nu_C(x_i))^\alpha \\ + \left(\begin{array}{l} \lambda_1 (1 - \mu_B(x_i) - \nu_B(x_i)) \\ + \lambda_2 (1 - \mu_C(x_i) - \nu_C(x_i)) \end{array} \right)^\alpha \end{array} \right\} \\ - \lambda_1 \log \left\{ \begin{array}{l} (\mu_B(x_i))^\alpha + (\nu_B(x_i))^\alpha \\ + (1 - \mu_B(x_i) - \nu_B(x_i))^\alpha \end{array} \right\} \\ - \lambda_2 \log \left\{ \begin{array}{l} (\mu_C(x_i))^\alpha + (\nu_C(x_i))^\alpha \\ + (1 - \mu_C(x_i) - \nu_C(x_i))^\alpha \end{array} \right\} \end{array} \right\} \right]$$

$$+ \frac{1}{n(1-\alpha)} \left[\sum_{x_i \in X_1} \left\{ \begin{array}{l} \log \left\{ \begin{array}{l} (\lambda_1 \mu_A(x_i) + \lambda_2 \mu_C(x_i))^\alpha \\ + (\lambda_1 \nu_A(x_i) + \lambda_2 \nu_C(x_i))^\alpha \\ + \left(\begin{array}{l} \lambda_1 (1 - \mu_A(x_i) - \nu_A(x_i)) \\ + \lambda_2 (1 - \mu_C(x_i) - \nu_C(x_i)) \end{array} \right)^\alpha \end{array} \right\} \\ - \lambda_1 \log \left\{ \begin{array}{l} (\mu_A(x_i))^\alpha + (\nu_A(x_i))^\alpha \\ + (1 - \mu_A(x_i) - \nu_A(x_i))^\alpha \end{array} \right\} \\ - \lambda_2 \log \left\{ \begin{array}{l} (\mu_C(x_i))^\alpha + (\nu_C(x_i))^\alpha \\ + (1 - \mu_C(x_i) - \nu_C(x_i))^\alpha \end{array} \right\} \end{array} \right\} \right]$$

≥ 0

This proves the theorem.

Theorem 5: For $A, B, C \in IFS(X)$,

$$JR_{\lambda,\alpha}(A \cup B, C) + JR_{\lambda,\alpha}(A \cap B, C) = JR_{\lambda,\alpha}(A, C) + JR_{\lambda,\alpha}(B, C).$$

Proof: Using definition in (10), we first have:

$$JR_{\lambda,\alpha}(A \cup B, C)$$

$$= \frac{1}{n(1-\alpha)} \sum_{i=1}^{n} \left[\begin{array}{l} \log \left\{ \begin{array}{l} \left(\lambda_1 \mu_{A\cup B}(x_i) + \lambda_2 \mu_C(x_i)\right)^{\alpha} \\ + \left(\lambda_1 \nu_{A\cap B}(x_i) + \lambda_2 \nu_C(x_i)\right)^{\alpha} \\ + \left(\begin{array}{l} \lambda_1 \left(1 - \mu_{A\cup B}(x_i) - \nu_{A\cap B}(x_i)\right) \\ + \lambda_2 \left(1 - \mu_C(x_i) - \nu_C(x_i)\right) \end{array} \right)^{\alpha} \end{array} \right\} \\ - \lambda_1 \log \left\{ \begin{array}{l} \left(\mu_{A\cup B}(x_i)\right)^{\alpha} + \left(\nu_{A\cap B}(x_i)\right)^{\alpha} \\ + \left(1 - \mu_{A\cup B}(x_i) - \nu_{A\cap B}(x_i)\right)^{\alpha} \end{array} \right\} \\ - \lambda_2 \log \left\{ \begin{array}{l} \left(\mu_C(x_i)\right)^{\alpha} + \left(\nu_C(x_i)\right)^{\alpha} \\ + \left(1 - \mu_C(x_i) - \nu_C(x_i)\right)^{\alpha} \end{array} \right\} \end{array} \right]$$

$$= \frac{1}{n(1-\alpha)} \left[\sum_{x_i \in X_1} \left\{ \begin{array}{l} \log \left\{ \begin{array}{l} \left(\lambda_1 \mu_B(x_i) + \lambda_2 \mu_C(x_i)\right)^{\alpha} \\ + \left(\lambda_1 \nu_B(x_i) + \lambda_2 \nu_C(x_i)\right)^{\alpha} \\ + \left(\begin{array}{l} \lambda_1 \left(1 - \mu_B(x_i) - \nu_B(x_i)\right) \\ + \lambda_2 \left(1 - \mu_C(x_i) - \nu_C(x_i)\right) \end{array} \right)^{\alpha} \end{array} \right\} \\ - \lambda_1 \log \left\{ \begin{array}{l} \left(\mu_B(x_i)\right)^{\alpha} + \left(\nu_B(x_i)\right)^{\alpha} \\ + \left(1 - \mu_B(x_i) - \nu_B(x_i)\right)^{\alpha} \end{array} \right\} \\ - \lambda_2 \log \left\{ \begin{array}{l} \left(\mu_C(x_i)\right)^{\alpha} + \left(\nu_C(x_i)\right)^{\alpha} \\ + \left(1 - \mu_C(x_i) - \nu_C(x_i)\right)^{\alpha} \end{array} \right\} \end{array} \right\} \right.$$

$$\left. + \sum_{x_i \in X_2} \left\{ \begin{array}{l} \log \left\{ \begin{array}{l} \left(\lambda_1 \mu_A(x_i) + \lambda_2 \mu_C(x_i)\right)^{\alpha} \\ + \left(\lambda_1 \nu_A(x_i) + \lambda_2 \nu_C(x_i)\right)^{\alpha} \\ + \left(\begin{array}{l} \lambda_1 \left(1 - \mu_A(x_i) - \nu_A(x_i)\right) \\ + \lambda_2 \left(1 - \mu_C(x_i) - \nu_C(x_i)\right) \end{array} \right)^{\alpha} \end{array} \right\} \\ - \lambda_1 \log \left\{ \begin{array}{l} \left(\mu_A(x_i)\right)^{\alpha} + \left(\nu_A(x_i)\right)^{\alpha} \\ + \left(1 - \mu_A(x_i) - \nu_A(x_i)\right)^{\alpha} \end{array} \right\} \\ - \lambda_2 \log \left\{ \begin{array}{l} \left(\mu_C(x_i)\right)^{\alpha} + \left(\nu_C(x_i)\right)^{\alpha} \\ + \left(1 - \mu_C(x_i) - \nu_C(x_i)\right)^{\alpha} \end{array} \right\} \end{array} \right\} \right]. \quad (20)$$

Next, again using definition in (10), we have

$JR_{\lambda,\alpha}(A \cap B, C)$

$$= \frac{1}{n(1-\alpha)} \sum_{i=1}^{n} \left[\begin{array}{l} \log \left\{ \begin{array}{l} \left(\lambda_1 \mu_{A\cap B}(x_i) + \lambda_2 \mu_C(x_i)\right)^{\alpha} \\ + \left(\lambda_1 \nu_{A\cup B}(x_i) + \lambda_2 \nu_C(x_i)\right)^{\alpha} \\ + \left(\begin{array}{l} \lambda_1 \left(1 - \mu_{A\cap B}(x_i) - \nu_{A\cup B}(x_i)\right) \\ + \lambda_2 \left(1 - \mu_C(x_i) - \nu_C(x_i)\right) \end{array} \right)^{\alpha} \end{array} \right\} \\ - \lambda_1 \log \left\{ \begin{array}{l} \left(\mu_{A\cap B}(x_i)\right)^{\alpha} + \left(\nu_{A\cup B}(x_i)\right)^{\alpha} \\ + \left(1 - \mu_{A\cap B}(x_i) - \nu_{A\cup B}(x_i)\right)^{\alpha} \end{array} \right\} \\ - \lambda_2 \log \left\{ \begin{array}{l} \left(\mu_C(x_i)\right)^{\alpha} + \left(\nu_C(x_i)\right)^{\alpha} \\ + \left(1 - \mu_C(x_i) - \nu_C(x_i)\right)^{\alpha} \end{array} \right\} \end{array} \right]$$

$$= \frac{1}{n(1-\alpha)} \left[\sum_{x_i \in X_1} \left\{ \begin{array}{l} \log \left\{ \begin{array}{l} \left(\lambda_1 \mu_A(x_i) + \lambda_2 \mu_C(x_i)\right)^{\alpha} \\ + \left(\lambda_1 \nu_A(x_i) + \lambda_2 \nu_C(x_i)\right)^{\alpha} \\ + \left(\begin{array}{l} \lambda_1 \left(1 - \mu_A(x_i) - \nu_A(x_i)\right) \\ + \lambda_2 \left(1 - \mu_C(x_i) - \nu_C(x_i)\right) \end{array} \right)^{\alpha} \end{array} \right\} \\ - \lambda_1 \log \left\{ \begin{array}{l} \left(\mu_A(x_i)\right)^{\alpha} + \left(\nu_A(x_i)\right)^{\alpha} \\ + \left(1 - \mu_A(x_i) - \nu_A(x_i)\right)^{\alpha} \end{array} \right\} \\ - \lambda_2 \log \left\{ \begin{array}{l} \left(\mu_C(x_i)\right)^{\alpha} + \left(\nu_C(x_i)\right)^{\alpha} \\ + \left(1 - \mu_C(x_i) - \nu_C(x_i)\right)^{\alpha} \end{array} \right\} \end{array} \right\} \right.$$

$$\left. + \sum_{x_i \in X_2} \left\{ \begin{array}{l} \log \left\{ \begin{array}{l} \left(\lambda_1 \mu_B(x_i) + \lambda_2 \mu_C(x_i)\right)^{\alpha} \\ + \left(\lambda_1 \nu_B(x_i) + \lambda_2 \nu_C(x_i)\right)^{\alpha} \\ + \left(\begin{array}{l} \lambda_1 \left(1 - \mu_B(x_i) - \nu_B(x_i)\right) \\ + \lambda_2 \left(1 - \mu_C(x_i) - \nu_C(x_i)\right) \end{array} \right)^{\alpha} \end{array} \right\} \\ - \lambda_1 \log \left\{ \begin{array}{l} \left(\mu_B(x_i)\right)^{\alpha} + \left(\nu_B(x_i)\right)^{\alpha} \\ + \left(1 - \mu_B(x_i) - \nu_B(x_i)\right)^{\alpha} \end{array} \right\} \\ - \lambda_2 \log \left\{ \begin{array}{l} \left(\mu_C(x_i)\right)^{\alpha} + \left(\nu_C(x_i)\right)^{\alpha} \\ + \left(1 - \mu_C(x_i) - \nu_C(x_i)\right)^{\alpha} \end{array} \right\} \end{array} \right\} \right]. \quad (21)$$

Adding (20) and (21), we get the result.

Theorem 6: For $A, B \in IFS(X)$,

- **(a)** $JR_{\lambda,\alpha}(A,B) = JR_{\lambda,\alpha}(A^C, B^C)$
- **(b)** $JR_{\lambda,\alpha}(A,B^C) = JR_{\lambda,\alpha}(A^C,B)$;
- **(c)** $JR_{\lambda,\alpha}(A,B) + JR_{\lambda,\alpha}(A^C,B) = JR_{\lambda,\alpha}(A^C,B^C) + JR_{\lambda,\alpha}(A,B^C)$.

where A^C and B^C represents the complement of intuitionistic fuzzy sets A and B respectively.

Proof: (a) The proof simply follows from the relation of membership and non-membership functions of an element in a set and its complement.

(b) Let us consider the expression

$$JR_{\lambda,\alpha}(A,B^C) - JR_{\lambda,\alpha}(A^C,B) \quad (22)$$

$$= \frac{1}{n(1-\alpha)} \left[\begin{array}{l} \log \left\{ \begin{array}{l} \left(\lambda_1 \mu_A(x_i) + \lambda_2 \nu_B(x_i)\right)^{\alpha} \\ + \left(\lambda_1 \nu_A(x_i) + \lambda_2 \mu_B(x_i)\right)^{\alpha} \\ + \left(\begin{array}{l} \lambda_1 \left(1 - \mu_A(x_i) - \nu_A(x_i)\right) \\ + \lambda_2 \left(1 - \mu_B(x_i) - \nu_B(x_i)\right) \end{array} \right)^{\alpha} \end{array} \right\} \\ - \lambda_1 \log \left\{ \begin{array}{l} \left(\mu_A(x_i)\right)^{\alpha} + \left(\nu_A(x_i)\right)^{\alpha} \\ + \left(1 - \mu_A(x_i) - \nu_A(x_i)\right)^{\alpha} \end{array} \right\} \\ - \lambda_2 \log \left\{ \begin{array}{l} \left(\nu_B(x_i)\right)^{\alpha} + \left(\mu_B(x_i)\right)^{\alpha} \\ + \left(1 - \nu_B(x_i) - \mu_B(x_i)\right)^{\alpha} \end{array} \right\} \end{array} \right]$$

$$-\frac{1}{n(1-\alpha)}\left[\begin{array}{l}\log\left\{\begin{array}{l}(\lambda_1\nu_A(x_i)+\lambda_2\mu_B(x_i))^\alpha\\+(\lambda_1\mu_A(x_i)+\lambda_2\nu_B(x_i))^\alpha\\+\left(\begin{array}{l}\lambda_1(1-\nu_A(x_i)-\mu_A(x_i))\\+\lambda_2(1-\mu_B(x_i)-\nu_B(x_i))\end{array}\right)^\alpha\end{array}\right\}\\-\lambda_1\log\left\{\begin{array}{l}(\nu_A(x_i))^\alpha+(\mu_A(x_i))^\alpha\\+(1-\nu_A(x_i)-\mu_A(x_i))^\alpha\end{array}\right\}\\-\lambda_2\log\left\{\begin{array}{l}(\mu_B(x_i))^\alpha+(\nu_B(x_i))^\alpha\\+(1-\mu_B(x_i)-\nu_B(x_i))^\alpha\end{array}\right\}\end{array}\right]$$

= 0.

(c) It immediately follows **(a)** and **(b)**.

This completes proof the theorem.

In the next section, we suggest an application of the measure proposed to multiple-attribute decision making problem and give an illustrative example.

5 Applications of intuitionistic fuzzy Jensen-Rényi divergence to multiple-attribute decision making

Vagueness is a fact of life and needs attention in matters of management. It can have several forms, for example, imperfectly defined facts, indirect data, or imprecise knowledge. For mathematical study, vague phenomena have got to be first suitably represented. IFSs are found to be suitable tools for this purpose. In this section, we present a method based on our proposed *intuitionistic fuzzy Jensen-Rényi divergence* defined over IFSs, to solve multiple-attribute decision making problems. It may be remarked that for a deterministic or probabilistic phenomenon where patterns show *stability* of the form, parameters have perhaps limited rule, but in vague phenomena, parameters provide a class of measures and choice for making appropriate selection by testing further. Intuitionistic fuzzy Jensen-Rényi divergence defined has parameters of two categories- the averaging parameters, $\lambda's$, and an extraneous parameter α, each serving a different purpose. In the example below, we bring out their role in multiple-attribute decision making.

Multiple-attribute decision making problems are defined on a set of alternatives, from which the decision maker has to select the best alternative according to some attributes. Suppose that there exists an alternative set $A=\{A_1,A_2,...,A_m\}$ which consists of m alternatives, the decision maker will choose the best alternative from the set A according to a set of n attributes $G=\{G_1,G_2,...,G_n\}$. Further let $D=(d_{ij})_{n\times m}$ be the intuitionistic fuzzy decision matrix, where $d_{ij}=(\mu_{ij},\nu_{ij},\pi_{ij})$ is an attribute value provided by the decision maker, such that μ_{ij} indicates the degree with which the alternative A_j satisfies the attribute G_i, ν_{ij} indicates the degree with which the alternative A_j does not satisfies the attribute G_i, and π_{ij} indicates the indeterminacy degree of alternative A_j to the attribute G_i, such that:

$\mu_{ij}\in[0,1]$, $\nu_{ij}\in[0,1]$, $\mu_{ij}+\nu_{ij}\le\pi_{ij}=1$, $\pi_{ij}=1-\mu_{ij}-\nu_{ij}$ $i=1,2,...,n$ and $j=1,2,...,m$.

To harmonize the data, first step is to look at the attributes. These, in general, can be of different types. If all the attributes $G=\{G_1,G_2,...,G_n\}$ are of the same type, then the attribute values do not need harmonization. However if these involve different scales and/or units, there is need to convert them all to the same scale and/or unit. Just to make this point clear, let us consider two types of attributes, namely, (i) cost type and the (ii) benefit type. Considering their natures, a benefit attribute (the bigger the values better is it) and cost attribute (the smaller the values the better) are of rather opposite type. In such cases, we need to first transform the attribute values of cost type into the attribute values of benefit type. So, we transform the intuitionistic fuzzy decision matrix $D=(d_{ij})_{n\times m}$ into the normalized intuitionistic fuzzy decision matrix $R=(r_{ij})_{m\times n}$ by the method given by Xu and Hu [30], where

$$r_{ij}=(\mu_{ij},\nu_{ij},\pi_{ij})=\begin{cases}d_{ij}, & \text{for benefit attribute } G_i\\ (d_{ij})^C, & \text{for cost attribute } G_i\end{cases}, \quad (23)$$

$$i=1,2,...,n;\ \ j=1,2,...,m$$

where $(d_{ij})^C$ is the complement of d_{ij}, such that $(d_{ij})^C=(\nu_{ij},\mu_{ij},\pi_{ij})$.

With attributes harmonized, using the measure defined in (10), we now stipulate following steps to solve our multiple-attribute intuitionistic fuzzy decision making problem:

Step 1: Based on the matrix $R=(r_{ij})_{m\times n}$, specify the options A_j $(j=1,2,...,m)$ by the characteristic sets:

$$A_j=\left\{\left\langle G_i,\mu_{ij},\nu_{ij},\pi_{ij}\right\rangle \mid G_i\in G\right\},$$
$$j=1,2,...,m \text{ and } i=1,2,...,n.$$

Step 2: Find the ideal solution A^*, given by:

$$A^*=\left\{\begin{array}{l}\left\langle\mu_{1*},\nu_{1*},\pi_{1*}\right\rangle,\left\langle\mu_{2*},\nu_{2*},\pi_{2*}\right\rangle,\\ \qquad ...,\left\langle\mu_{n*},\nu_{n*},\pi_{n*}\right\rangle\end{array}\right\}, \quad (24)$$

where, for each $i=1,2,...,n$,

$$(\mu_{i*},\nu_{i*},\pi_{i*})=\left(\begin{array}{l}\max\limits_j\mu_{ij},\min\limits_j\nu_{ij},\\ \qquad 1-\max\limits_j\mu_{ij}-\min\limits_j\nu_{ij}\end{array}\right). \quad (25)$$

Step 3: Calculate $JR_{\lambda',\alpha}(A_j,A^*)$ using the following expression for it:

$$= \frac{1}{n(1-\alpha)} \sum_{i=1}^{n} \left[\begin{array}{l} \log \left\{ \begin{array}{l} \left(\lambda_1^j \mu_{A_j}(x_i) + \lambda_2^j \mu_{A^*}(x_i)\right)^\alpha \\ + \left(\lambda_1^j \nu_{A_j}(x_i) + \lambda_2^j \nu_{A^*}(x_i)\right)^\alpha \\ + \left(\lambda_1^j \pi_{A_j}(x_i) + \lambda_2^j \pi_{A^*}(x_i)\right)^\alpha \end{array} \right\} \\ - \lambda_1^j \log \left\{ \begin{array}{l} \left(\mu_{A_j}(x_i)\right)^\alpha + \left(\nu_{A_j}(x_i)\right)^\alpha \\ + \left(\pi_{A_j}(x_i)\right)^\alpha \end{array} \right\} \\ - \lambda_2^j \log \left\{ \begin{array}{l} \left(\mu_{A^*}(x_i)\right)^\alpha + \left(\nu_{A^*}(x_i)\right)^\alpha \\ + \left(\pi_{A^*}(x_i)\right)^\alpha \end{array} \right\} \end{array} \right] \quad (26)$$

where $\lambda_1^j, \lambda_2^j \in [0,1]$, and $\lambda_1^j + \lambda_2^j = 1 \ \ \forall \ j = 1,2,...,m$.

Step 4: Rank the alternatives A_j, $j = 1,2,...,m$, in accordance with the values $JR_{\lambda^j,\alpha}(A_j, A^*)$, $j = 1,2,...,m$, and select the best one alternative, denoted by A_k with smallest $JR_{\lambda^j,\alpha}(A_j, A^*)$. Then A_k is the best choice.

In order to demonstrate the application of the above proposed method to a real multiple attribute decision making, we consider below a numerical example.

Example: Consider a customer who wants to buy a car. Let five types of cars (alternatives) A_j $(j = 1,2,3,4,5)$ be available. The customer takes into account six attributes to decide which car to buy: (1) G_1: fuel economy, (2) G_2: aerodynamic degree, (3) G_3: price, (4) G_4: comfort, (5) G_5: design and (6) G_6: safety. We note that G_3 is a cost attribute while other five are benefit attributes. Next let us assume that the characteristics of the alternatives A_j $(j = 1,2,3,4,5)$ are represented by the intuitionistic fuzzy decision matrix $D = (d_{ij})_{6\times5}$ shown in the following table:

Table I: Intuitionistic fuzzy decision matrix D

	A_1	A_2	A_3	A_4	A_5
G_1	(0.5,0.4, 0.1)	(0.4,0.3, 0.3)	(0.5,0.2, 0.3)	(0.4,0.2, 0.4)	(0.6,0.4, 0.0)
G_2	(0.7,0.2, 0.1)	(0.8,0.2, 0.0)	(0.9,0.1, 0.0)	(0.8,0.0, 0.2)	(0.5,0.2, 0.3)
G_3	(0.4,0.3, 0.3)	(0.5,0.2, 0.3)	(0.6,0.1, 0.3)	(0.7,0.3, 0.0)	(0.8,0.1, 0.1)
G_4	(0.6,0.2, 0.2)	(0.6,0.3, 0.1)	(0.8,0.1, 0.1)	(0.9,0.1, 0.0)	(0.4,0.2, 0.4)
G_5	(0.4,0.5, 0.1)	(0.6,0.4, 0.0)	(0.3,0.5, 0.2)	(0.5,0.3, 0.2)	(0.9,0.0, 0.1)
G_6	(0.3,0.1, 0.6)	(0.7,0.1, 0.2)	(0.6,0.2, 0.2)	(0.6,0.1, 0.3)	(0.4,0.3, 0.3)

First, we transform the attribute values of cost type (G_3) into the attribute values of benefit type (G_3') by using Eq. (23):

$$G_3' = (G_3)^C = \left\{ \begin{array}{l} (0.3,0.4,0.3), (0.2,0.5,0.3), (0.1,0.6,0.3), \\ (0.3,0.7,0.0), (0.1,0.8,0.1) \end{array} \right\},$$

and then $D = (d_{ij})_{6\times5}$ is transformed into $R = (r_{ij})_{6\times5}$, we get the following table:

Table II: Normalized intuitionistic fuzzy decision matrix R

	A_1	A_2	A_3	A_4	A_5
G_1	(0.5,0.4, 0.1)	(0.4,0.3, 0.3)	(0.5,0.2, 0.3)	(0.4,0.2, 0.4)	(0.6,0.4, 0.0)
G_2	(0.7,0.2, 0.1)	(0.8,0.2, 0.0)	(0.9,0.1, 0.0)	(0.8,0.0, 0.2)	(0.5,0.2, 0.3)
G_3'	(0.3,0.4, 0.3)	(0.2,0.5, 0.3)	(0.1,0.6, 0.3)	(0.3,0.7, 0.0)	(0.1,0.8, 0.1)
G_4	(0.6,0.2, 0.2)	(0.6,0.3, 0.1)	(0.8,0.1, 0.1)	(0.9,0.1, 0.0)	(0.4,0.2, 0.4)
G_5	(0.4,0.5, 0.1)	(0.6,0.4, 0.0)	(0.3,0.5, 0.2)	(0.5,0.3, 0.2)	(0.9,0.0, 0.1)
G_6	(0.3,0.1, 0.6)	(0.7,0.1, 0.2)	(0.6,0.2, 0.2)	(0.6,0.1, 0.3)	(0.4,0.3, 0.3)

The step-wise procedure now goes as follows.

Step 1: Based on $R = (r_{ij})_{6\times5}$, we have characteristic sets of the alternatives A_j $(j = 1,2,...,5)$ by

$$A_1 = \left\{ \begin{array}{l} (0.5,0.4,0.1), (0.7,0.2,0.1), (0.3,0.4,0.3), \\ (0.6,0.2,0.2), (0.4,0.5,0.1), (0.3,0.1,0.6) \end{array} \right\},$$

$$A_2 = \left\{ \begin{array}{l} (0.4,0.3,0.3), (0.8,0.2,0.0), (0.2,0.5,0.3), \\ (0.6,0.3,0.1), (0.6,0.4,0.0), (0.7,0.1,0.2) \end{array} \right\},$$

$$A_3 = \left\{ \begin{array}{l} (0.5,0.2,0.3), (0.9,0.1,0.0), (0.1,0.6,0.3), \\ (0.8,0.1,0.1), (0.3,0.5,0.2), (0.6,0.2,0.2) \end{array} \right\},$$

$$A_4 = \left\{ \begin{array}{l} (0.4,0.2,0.4), (0.8,0.0,0.2), (0.3,0.7,0.0), \\ (0.9,0.1,0.0), (0.5,0.3,0.2), (0.6,0.1,0.3) \end{array} \right\},$$

$$A_5 = \left\{ \begin{array}{l} (0.6,0.4,0.0), (0.5,0.2,0.3), (0.1,0.8,0.1), \\ (0.4,0.2,0.4), (0.9,0.0,0.1), (0.4,0.3,0.3) \end{array} \right\}.$$

Step 2: Using (24) and (25), we obtain A^*:

$$A^* = \left\{ \begin{array}{l} (0.6,0.2,0.2), (0.9,0.0,0.1), (0.3,0.4,0.3), \\ (0.9,0.1,0.0), (0.9,0.0,0.1), (0.7,0.1,0.2) \end{array} \right\}.$$

Step3: We use formula (26) to measure $JR_{\lambda^j,\alpha}(A_j, A^*)$, choosing the various values of parameter. First we take $\lambda_1^j = \lambda_2^j = 0.5 \ \ \forall \ j = 1,2,...,5$; and $\alpha = 0.3$, $\alpha = 0.5$ and $\alpha = 0.7$ respectively, we get the following table:

Table III: Values of $JR_{\lambda^j,\alpha}(A_j, A^*)$ for $\alpha = 0.3, 0.5, 0.7$

	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$
$JR_{\lambda^j,\alpha}(A_1, A^*)$	0.1453	0.1409	0.1345
$JR_{\lambda^j,\alpha}(A_2, A^*)$	0.1908	0.1584	0.1299
$JR_{\lambda^j,\alpha}(A_3, A^*)$	0.1617	0.1400	0.1214
$JR_{\lambda^j,\alpha}(A_4, A^*)$	0.0946	0.0905	0.0849
$JR_{\lambda^j,\alpha}(A_5, A^*)$	0.1483	0.1467	0.1424

Based on the calculated values of $JR_{\lambda^j,\alpha}(A_j, A^*)$ in table III, we get the following orderings of ranks of the alternatives A_j $(j = 1,2,3,4,5)$:

For $\alpha = 0.3$, $A_4 \succ A_1 \succ A_5 \succ A_3 \succ A_2$.

For $\alpha = 0.5$, $A_4 \succ A_3 \succ A_1 \succ A_5 \succ A_2$.

For $\alpha = 0.7$, $A_4 \succ A_3 \succ A_2 \succ A_1 \succ A_5$.

Since $JR_{\lambda^j,\alpha}(A_4, A^*)$ is smallest among the values of $JR_{\lambda^j,\alpha}(A_j, A^*)$ $\{j = 1,2,...,5\}$ for $\alpha = 0.3$, $\alpha = 0.5$ and $\alpha = 0.7$, so A_4 is the most preferable alternative. Thus here we find that variation in values of α brings about change in ranking, but leaves the best choice unchanged.

Change in Consideration: In the above consideration, same values of λ_i^j were taken. But in a realistic situation these can also be different for different alternatives. The value of λ_i^j may then depend on an un-explicit (like past experience or pressures) on the decision maker.

Let us next consider intuitionistic fuzzy Jensen-Rényi divergence measures $JR_{\lambda^j,\alpha}(A_j, A^*)$, taking different values of λ_i^j:

We take $\lambda_1^1 = 0.5$, $\lambda_2^1 = 0.5$; $\lambda_1^2 = 0.4$, $\lambda_2^2 = 0.6$; $\lambda_1^3 = 0.8$, $\lambda_2^3 = 0.2$; $\lambda_1^4 = 0.5$, $\lambda_2^4 = 0.5$; $\lambda_1^5 = 0.3$, $\lambda_2^5 = 0.7$ and $\alpha = 0.5$.

Calculating $JR_{\lambda^j,\alpha}(A_j, A^*)$, we get the following table:

Table IV: Values of $JR_{\lambda^j,\alpha}(A_j, A^*)$ for $\alpha = 0.5$

$JR_{\lambda^1,\alpha}(A_1, A^*)$	0.0965
$JR_{\lambda^2,\alpha}(A_2, A^*)$	0.1644
$JR_{\lambda^3,\alpha}(A_3, A^*)$	0.0856
$JR_{\lambda^4,\alpha}(A_4, A^*)$	0.1178
$JR_{\lambda^5,\alpha}(A_5, A^*)$	0.1479

The resulting order of rankings then is

$$A_3 \succ A_1 \succ A_4 \succ A_5 \succ A_2.$$

Thus A_3 is the most preferable alternative.

If we take

$\lambda_1^1 = 0.5$, $\lambda_2^1 = 0.5$; $\lambda_1^2 = 0.7$, $\lambda_2^2 = 0.3$; $\lambda_1^3 = 0.3$, $\lambda_2^3 = 0.7$; $\lambda_1^4 = 0.4$, $\lambda_2^4 = 0.6$; $\lambda_1^5 = 0.8$, $\lambda_2^5 = 0.2$ and $\alpha = 0.5$, calculating $JR_{\lambda^j,\alpha}(A_j, A^*)$, we get the following table:

Table V: Values of $JR_{\lambda^j,\alpha}(A_j, A^*)$ for $\alpha = 0.5$

$JR_{\lambda^1,\alpha}(A_1, A^*)$	0.1409
$JR_{\lambda^2,\alpha}(A_2, A^*)$	0.1296
$JR_{\lambda^3,\alpha}(A_3, A^*)$	0.1493
$JR_{\lambda^4,\alpha}(A_4, A^*)$	0.1268
$JR_{\lambda^5,\alpha}(A_5, A^*)$	0.0965

The resulting order of rankings then is

$$A_5 \succ A_4 \succ A_2 \succ A_1 \succ A_3.$$

Resulting in A_5 as the most preferable option. Thus for a given value of parameter α, averaging parameters $\lambda's$ can effect the choice.

The numerical example shows that change in order of the rankings results by change in parameters λ & α establishing the significance of these parameters in multi-attribute sensitive decision making problems.

6 Conclusions

The paper provides a measure and application in multiple-attribute decision making problem under intuitionistic fuzzy environment. This study can lead to symmetric measure and resulting other insight into studying IFSs.

References

[1] Atanassov, K.T. (1986) Intuitionistic fuzzy sets, Fuzzy Sets and Systems, 20(1), 87-96.

[2] Atanassov K.T. (1999) Intuitionistic fuzzy sets. Springer Physica-Verlag, Heidelberg.

[3] Boran, F.E., Genç, S and Akay, D. (2011) Personnel selection based on intuitionistic fuzzy sets, Human Factors and Ergonomics in Manufacturing & Service Industries, 21(5), 493-503.

[4] Chen, S.M. and Tan, J.M. (1994) Handling multicriteria fuzzy decision-making problems based on vague set theory, Fuzzy Sets and Systems, 67(2), 163-172.

[5] De, S.K., Biswas, R. and Roy, A.R. (2001) An application of intuitionistic fuzzy sets in medical diagnosis, Fuzzy sets and Systems, 117(2), 209-213.

[6] Dengfeng, L. and Chuntian, C. (2002) New similarity measures of intuitionistic fuzzy sets and application to pattern recognitions, Pattern Recognition Letters, 23(1-3), 221-225.

[7] Dengfeng, L. (2004) Some measures of dissimilarity in intuitionistic fuzzy structures, Journal of Computer and System Sciences, 68(1), 115-122.

[8] Dengfeng, L. (2008) Extension of the LINMAP for multi-attribute decision making under Attanassov's intuitionistic fuzzy environment, Fuzzy Optimization and Decision Making, 7(1), 17-34.

[9] Endres, D.M. and Schindelin, J.E. (2003) A new metric for probability distributions, IEEE Transactions on Information Theory, 49(7), 1858-1860.

[10] He, Y., Hamza, A.B. and Krim, H. (2003) A generalized divergence measure for robust image registration, IEEE Transactions on Signal Processing, 51(5), 1211-1220.

[11] Hung, W.L. and Yang, M.S. (2008) On the *J*-divergence of intuitionistic fuzzy sets and its application to pattern recognition, Information Sciences, 178(6), 1641-1650.

[12] Hatzimichailidis, A.G., Papakostas, G.A. and Kaburlasos, V.G. (2012) A novel distance measure

of intuitionistic fuzzy sets and its application to pattern recognition problems, International Journal of Intelligent Systems, 27(4), 396-409.

[13] Kullback, S. and Leibler, R.A. (1951) On information and sufficiency, Annals of Mathematical Statistics, 22(1), 79–86.

[14] Khaleie, S. and Fasanghari M. (2012) An intuitionistic fuzzy group decision making method using entropy and association coefficient, Soft Computing, 16(7), 1197-1211.

[15] Lin, J. (1991) Divergence measure based on Shannon entropy, IEEE Transactions on Information Theory, 37(1), 145-151.

[16] Li, D.F. (2005) Multi-attribute decision-making models and methods using intuitionistic fuzzy sets, Journal of Computer and System Sciences, 70(1), 73-85.

[17] Liu, H.W. and Wang, G.J. (2007) Multi-criteria decision-making methods based on intuitionistic fuzzy sets, European Journal of Operation Research, 179(1), 220-233.

[18] Lin, L., Yuan, X.H. and Xia, Z.Q. (2007) Multicriteria fuzzy decision-making methods based on intuitionistic fuzzy sets, Journal of Computer and System Sciences, 73(1), 84-88.

[19] Mitchell, H.B. (2003) On the Dengfeng-Chuntian similarity measure and its application to pattern recognition, Pattern Recognition Letters, 24(16), 3101-3104.

[20] Pasi, G., Yager, R.R. and Atanassov, K. (2004) Intuitionistic fuzzy graph interpretations of multi-person multi-criteria decision-making: Generalized net approach. Proc. Second IEEE Int. Conf. Intell. Syst., Italy, 434- 439.

[21] Rènyi, A. (1961) On measures of entropy and information. Proceeding of the Forth Berkeley Symposium on Mathematics, Statistics and Probability, 1, 547-561.

[22] Rao, C. and Nayak, T. (1985) Cross entropy, dissimilarity measures and characterizations of quadratic entropy, IEEE Transactions on Information Theory, 31(5), 589-593.

[23] Shannon, C.E. (1948) A mathematical theory of communication, Bell System Technical Journal, 27, 379-423,623-656.

[24] Vlachos, I.K. and Sergiadis, G.D. (2007) Intuitionistic fuzzy information-application to pattern recognition, Pattern Recognition Letters, 28(2), 197–206.

[25] Verma, R., Sharma, B.D. (2012) On generalized intuitionistic fuzzy divergence (relative information) and their properties, Journal of Uncertain Systems, 6(4), 308-320.

[26] Szmidt, E. and Kacprzyk, J. (2000) Distances between intuitionistic fuzzy sets, Fuzzy Sets and Systems, 114(3), 505-518.

[27] Szmidt, E. and Kacprzyk, J. (2002) Using intuitionistic fuzzy sets in group decision making, Control and Cybernetics, 31, 1037-1053.

[28] Wei, G.W. (2008) Maximizing deviation method for multiple attribute decision making in intuitionistic fuzzy setting, Knowledge-Based Systems, 21(8), 833-836.

[29] Xu, Z.S. and Yager, R.R. (2008) Dynamic intuitionistic fuzzy multi-attribute decision making, International Journal of Approximate Reasoning, 48, 246-262.

[30] Xu, Z. and Hu, H. (2010) Projection models for intuitionistic fuzzy multiple attribute decision making, International Journal of Information Technology & Decision Making, 9 (2), 267-280.

[31] Yanhong, L., David, L.O. and Zhang, Q. (2007) Similarity measures between intuitionistic fuzzy (vague) sets: A comparative analysis, Pattern Recognition Letters, 28(2), 278-285.

[32] Zadeh, L.A. (1965) Fuzzy sets, Information and Control, 8(3), 338-353.

[33] Zhang, S.F. and Liu, S.Y. (2011) GRA-based intuitionistic multi-criteria decision making method for personnel selection, Expert Systems with Applications, 38(9), 11401-11405.

Algorithmic Tools for the Transformation of Petri Nets to DEVS

Mohammed Redjimi and Sofiane Boukelkoul
Université 20 Août 1955, Faculté des sciences, Département d'informatique
21000, Skikda, Algeria
E-mail: redjimimed@yahoo.fr, Bouk.sofiane@yahoo.fr

Keywords: DEVS, Petri nets, coupling models, multi-modeling, modeling and simulation

Received: May 1, 2013

Complex systems are characterized not only by the diversity of their components, but also by the interconnections and interactions between them. For modeling such systems, we often need several formalisms and we must concern ourselves with the coexistence of heterogeneous models. This objective can be achieved by using multi-modeling. The transformation of such models in a pivot model is a technique in this context. This paper introduces the DEVS 'Discrete Event System Specification' which model coupling approach is supported by a proposal for transformation of Petri nets in DEVS models. Petri Nets are universal formalisms which offer mathematical and graphical concepts for modeling the structure and the behavior of systems. We present mechanisms which can systematically transform the places and transitions in Petri nets to DEVS models. The coupling of these models generates a DEVS coupled model capable of running on platforms based on DEVS formalism.

Povzetek: Opisana je transformacija Petri mrež v formalizem DEVS.

1 Introduction

The diversity and the complexity of increasingly growing systems has forced the scientific community to implement tools for modeling and simulation [1] [2] [3] more and more efficient and meet the expressed requirements and constraints and support the heterogeneity and especially coupling systems in various disciplines. Now, it appears essential to use federative tools which offer extensive possibilities of abstraction and formalization. The multi-modeling consists of using several formalisms when one wants to model complex systems whose components are heterogeneous [4]. The idea developed in this paper is to determine a powerful formalism and abstraction that is as universal as possible to federate a set of concepts for the expression of different models. Once the formal model described, verified and validated it comes to transforming it into an executable form. In this article, we opted for Petri nets [5] [6] as tools for formal and abstract modeling of complex systems and DEVS "Discrete Event System Specification" [7] [8] [9] as universal formalism for the coupling of several transformation models. We detail in what follows mechanisms for transforming Petri nets (PN) in DEVS models [10].It consists of an algorithm permitting to systematically transform places and transitions to atomic DEVS models.

This paper begins by introducing the concept of multi-modeling. Then, we formally define DEVS and PN specifications. The following section shows the strength of DEVS as a universal system of multi-modeling followed by a formal approach to transform PN in DEVS models. We end this paper with a conclusion and perspectives.

2 Multi-modelling

Currently, systems can achieve large degrees of complexities and heterogeneities by combining multiple aspects which requires the use of several formalisms for their representation. Multi-modeling is used to represent these systems by using different formalisms. In this case, many models based on different formalisms can coexist in a single model. According to Hans Vangheluwe [2], the paradigm of multi modeling focuses on three axes:

- Different formalisms describe the coupling and the transformation of models.
- The relationship between the models at each level of abstraction is clearly defined.
- The meta-model focuses on the description of the classes of models (models of models).

In [11] there is a representation of various possible transformations by using formalism transformation graph "FTG".

3 Related works and motivations

In multi-modeling, several researches have focused on the study of the relationship between PN or other dynamic formalism and DEVS formalisms, since DEVS is considered as one of the basic modeling formalisms based on the unifying framework of general dynamic modeling formalism. Juan de Lara and al. proposed in [12] a modeling based multi-paradigm to generate PN and State-Charts. It consists of modeling at multiple levels of abstraction implemented in AToM^3 (A Tool for Multi-formalism and Meta-Modeling) [13] [14] [15], where is presented a graphical abstraction of meta-models of Sate charts and PNs. The use of CD++ to develop PN [16] [17] is close to our work. However it

only provides tools for generating PN by using library of predefining models for PN places and transitions. Therefore, one may be not finding the appropriate model for a given transition especially when it contains a big number of ports. Furthermore, in [17] we don't find a vital parallelism because firing transitions is scheduled. That means one never finds more than one transition in firing state, while the parallelism is one of the fundamental PN characteristics. Thus the conflict characteristic of PNs is silently absent, since without parallelism the problematic of conflict is not considered. So the value of our work is that is characterized by the development of algorithms that can automatically transform the existing PN in DEVS models [10]. Moreover, the most important characteristics of PNs such as parallelism, concurrency and conflict are well preserved in our approach.

4 DEVS formalism

DEVS was initially introduced by B. P. Zeigler [7] in 1976 for discrete event systems modeling. In DEVS, there are two kinds of models: atomic and coupled models. Atomic model is based on a continuous time inputs, outputs, states and functions. Coupled models are constructed by connecting several atomic models.

A DEVS atomic model is described by the following equation:

$$\text{AtomicDEVS} = (X, Y, S, \delta int, \delta ext, \delta con, \lambda, ta) \quad (1)$$

Where:

X is the set of external inputs. Y is the set of model outputs. S is the set of states. δint: S → S: represents the internal transition function that changes the state of the system autonomously. It depends on the time elapsed in the current state.

δext: S×X→S: is the external transition function occurs when model receives an external event. It returns the new state of the system based on the current state. δcon: X→SxS: is the transition function of conflict. It occurs if an external event happens when an internal system status changes. This feature is only present in a variant of DEVS: Parallel DEVS [8] [18]. λ: S → Y: is the output function of the model. It is activated when the elapsed time in a given state is equal to its life (ta (s) represents the life of a state "s" of the system if no external event occurs).

Coupled DEVS formalism describes a system as a network of components.

$$\text{CoupledDevs} = (X_{self}, Y_{self}, D, \{M_d / d \in D\}, EIC, EOC, IC) \quad (1)$$

Where Self: is the model itself. X_{self} is the set of inputs of the coupled model. Y_{self} is the set of outputs of the coupled model. D is the set of names associated with the components of the model, self is not in D. $\{M_d / d \in D\}$ is the set of components of the coupled model. EIC, EOC and IC define the coupling structure in the coupled model. EIC is the set of external input couplings. They connect the model inputs coupled to those of its own components. EOC is the external output couplings. They connect the outputs of the components to those of the coupled. IC defines internal coupling. It connects the outputs of components with entries from other components in the same coupled model.

In DEVS, both of atomic and coupled models can be represented graphically as illustrated in Fig. 1.

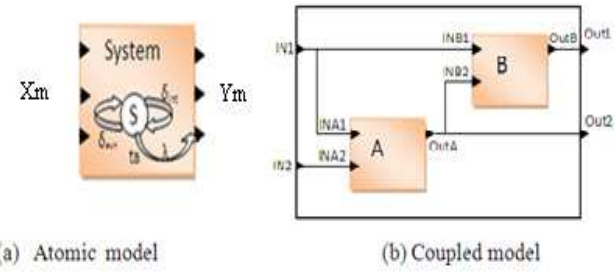


Figure 1: Representation of DEVS (a) atomic and (b) coupled models.

5 Petri nets (PN)

Petri Nets are a modeling formalism originally developed by C. A. Petri [5] [6]. They are very suitable for modeling dynamic systems.

Several types of nets can be used (timed Petri nets, colored Petri nets …) [19] [20]. We use classical Petri nets defined by the following 5-tuple:

$$PN = (P, T, PRE, POST, Mo) \quad (2)$$

P: is the set of places. T: is the set of transitions. PRE: the matrix generated by applying P x T → N. PRE [i, j] = n / n = 0 if the place is not upstream of the transition tj else n = τ / τ is the weight of the arc from pi to tj. POST: the matrix generated by applying T x P → N. POST [i, j] = n / n = 0 if the place pi is not downstream of the transition tj else n = τ / τ is the weight of the arc from tj to pi. M0: is the vector of initial marking. M[i] = k / k is the number of tokens in place pi. Fig. 2, shows a PN in the left (a) which consists of three places and one transition modeling action (T1) having two conditions (P1, P2) to be run. The result is put in place (P3).

6 PN to DEVS Transformation

6.1 Why DEVS?

DEVS provides a modular and hierarchical representation of dynamic models. Events generated by a model can take values in different areas and can be used as stimuli for other models. Also, according to B.P. Zeigler [7] [8], we can show that there is a DEVS model corresponding to each discrete event systems. We can go further, in fact, DEVS can be 'universal' [21] and allows the coupling of models and formalisms described with heterogeneous paradigms [11].

The main idea is that the models are considered as black boxes that have links with the outside world only through ports of inputs and outputs. Using this abstraction feature, several models can be coupled while enjoying the reuse of existing models. It is also possible to

perform the formal verification of DEVS models, which is a valuable aid in the design of systems [22] [23].
Several DEVS-based platforms are available such as VLE (Virtual Laboratory Environment)[24][25], DEVSJAVA [26] developed in Java, Cell-DEVS (Cellular DEVS) which is based on the formalism of cellular automata [27].
The coupling of models based on DEVS is a typical task. However, non-DEVS models require an extra effort to be coupled. Two methods exist to incorporate a non-DEVS model into a DEVS environment: co-simulation and transformation [28]. The transformation of non-DEVS models (PN in our case) in DEVS models requires to specifying models in a uniform language. In the case of a co-simulation, the communications between simulators is considered. Several works such as HLA (High Level Architecture) [29] take in account this way.

6.2 Mechanisms of PN to DEVS transformation

The idea of our approach is to have as result a DEVS coupled model (CDEVS) faithful to the input PN.

6.2.1 Structure of Resulting DEVS Model

The transformation of Petri provides a DEVS coupled model where places and transitions are replaced by atomic DEVS models. Fig.3, illustrates the CDEVS model corresponding to the PN example. The DEVS model corresponding to the "transition" of PN (TDEVS for "Transition DEVS") is characterized by an output port "control" (CT1) able to send events to places upstream and verify the number of tokens or inform them about its firing. However, TDEVS receives events from the models corresponding to places upstream (PDEVS "Place DEVS") with control ports as much as number of places (CPiT1).

TDEVS is not linked by its downstream CDEVS except by output port for each AT1Pi (in black) to inform them about its crossing. All TDEVS and PDEVS are provided with an output port OutTi and OutPi (in blue). These ports are coupled directly with the output ports for eventual CDEVS output. All PDEVS have an input port (InitPi) by which they are coupled with CDEVS via an input port InitP (in green) to initialize the marking of places. The arcs from place Pi to the transition Tj are translated into output ports APiTj (PDEVS) and input ports APiTj (TDEVS) corresponding to T (black). The creation of the structure of DEVS model corresponding to the PN is performed by algorithm1 which takes as input a PN= (P, T, PRE, POST, M0). The result is a DEVS model. Algorithm1 creates links corresponding to the arcs that link places by upstream transitions thanks to PRE matrix. The POST matrix is used for the coupling between TDEVS (transitions) and PDEVS (places) downstream of the transition.

Fig. 2 illustrates the elementary transformations of PN components to their equivalent objects in DEVS. Where (a) represents a single place with the minimum of ports it has to possess. (b) Illustrates a single given transition. (c) and (d) represents the minimum of IC between a place and a transition. (e) Corresponds to a graphical representation of IC in case of conflict between two transitions. Finally (f) represents the IC of typical transformation with parallelism.

Formally, the transformation is presented as follow:

PN = (P,T,PRE,POST,Mo) →
CDEVS=(X,Y,D,EIC,EOC,IC)
Where:
D = {P ∪ T}
X = {InitP, InitT}
Y = {OutDi / Di is atomic model representing Pi or Ti}
EIC = {(CDEVS.InitP, PDEVS.IntPi) ∪ (CDEVS.initT , TDEVS.IntTj)/ i ∈ N^+ & i < Number of places, j Є N^+ & j < Number of transitions }
EOC = {(Pi.OutPi, CM.OutPi) , (Tj.OutTj, CM.OutTj)/ i Є N+ & i<Number of places, jЄN^+ & j<Number of transitions }
IC = {
{(Pi.APiTj, Tj.APiTj) / PRE[i,j]>0 }
∪ {(Tj.ATjPi, Pi.ATjPi) / POST[i,j]>0 }
∪ { {Tj.CTj} X {Pi.CTjPi} / PRE[i,j]>0 }
∪ { (Pi.C PiTj , Tj.CPiTj} / PRE[i,j]>0 }
/i Є N^+ & I < Number of places, j Є N^+ & j < Number of transitions

}

Algorithm 1 : Transformation PN To DEVS

Main_PN_DEVS

Input PN= (P,T,PRE,POST,M0)
Output CDEVS //coupled model
Begin :
Create CDEVS as coupled DEVS model //void model
For all transition i **do**
create TDEVSi as atomic DEVS model
end for
for all places j **do**
create PDEVSj as atomic DEVS model
end for
for all PDEVSj **do**
add ‘InitPj’ as intput port and join it to
CDEVS.IN.InitP //starting tokens
add ‘OutPj’ as output port and join it to
CDEVS.OUT.OutPj //output stream
end for
for all TDEVSi **do**
add ‘InitTi’ as input port //initialize, stop, pause, release
join ‘InitTi’ port to CDEVS.IN. InitT port //coupling
add ‘OutTi’ as output port and join it to
CDEVS.OUT.OutTi //output stream
add ‘CTi’ as output port // control: check, reserve, decrement, cancel

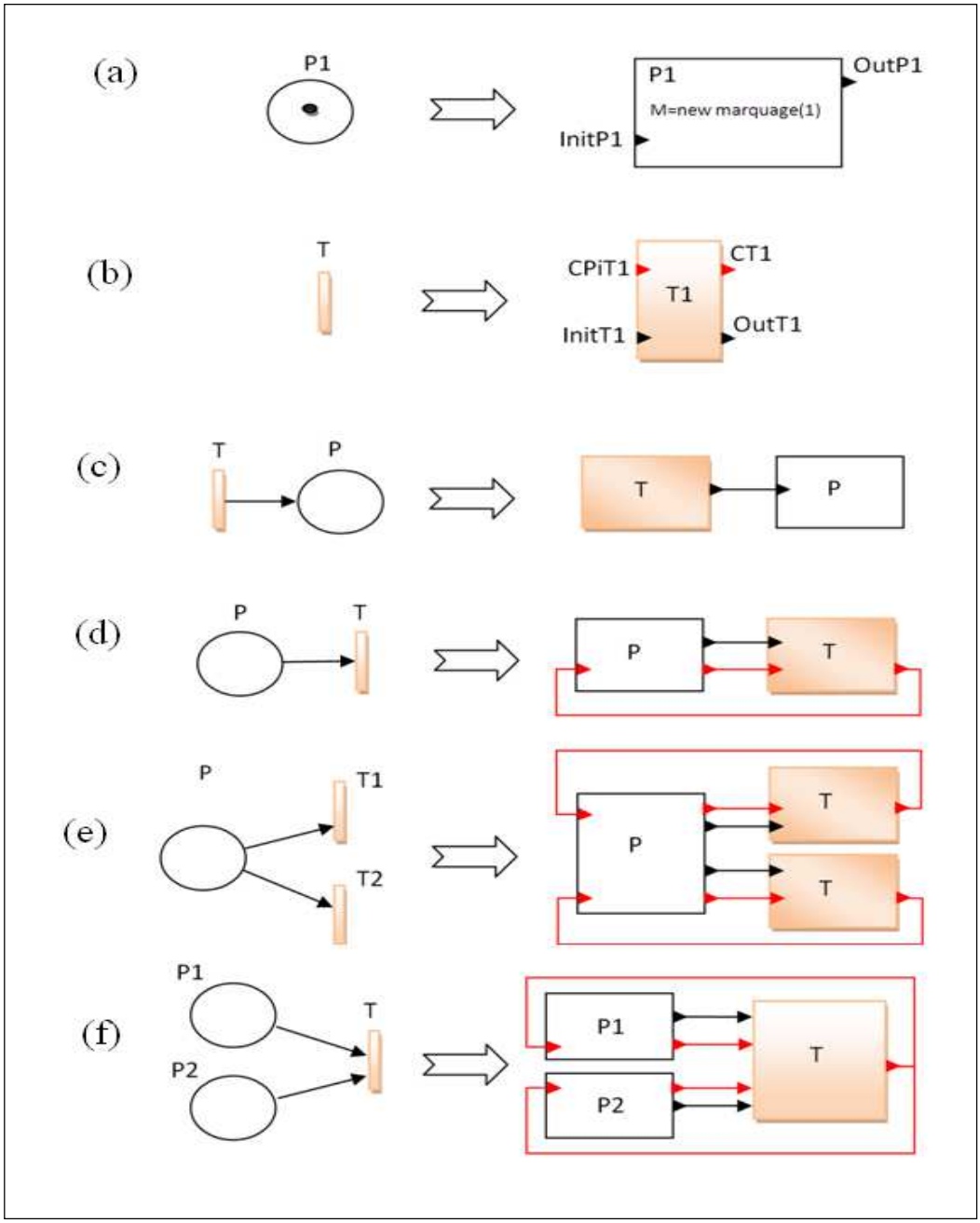


Figure 2**:** Graphical representation of elementary transformations and IC between generated DEVS models

for all PDEVSj **do**

if (PRE[i,j] > 0) //upstream place
add to PDEVSj 'CTiPj' as input port //check, reserve, decrement, cancel
join TDEVSi.OUT.CTi to PDEVSj.IN.CTiPj // coupling
add to PDEVSj 'CPjTi' as output port //ok, busy ,number_of_free_tokens
add to TDEVSi 'CPjTi' as input port //ok, busy ,number_of_free_tokens
join PDEVSj.OUT.CPjTi to TDEVSi.IN. CPjTi // coupling
add to PDEVSj 'APjTi' as output port //arc: value = PRE[i,j]
add to TDEVSi 'APjTi' as intput port //arc: value = PRE[i,j]
join PDEVSj.OUT. APjTi to TDEVSi.IN.APjTi // coupling
end if
if (POST[i,j] > 0) //downstream places
add to TDEVSi 'ATiPj' as output port //arc: value = POST[i,j]
add to PDEVSj 'ATiPj' as input port //arc: value = POST[i,j]
join TDEVSi.OUT.ATiPj to PDEVSj.IN.ATiPj // coupling
end if
end for
end for
end Main_PN_DEVS

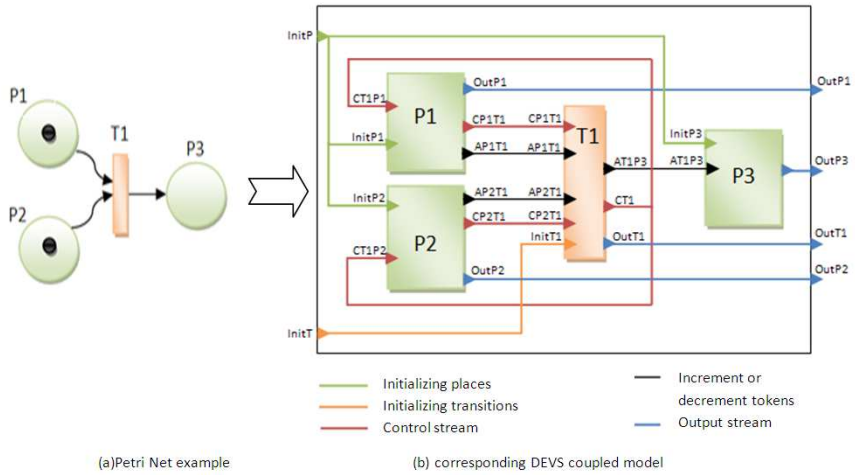


Figure 3**:** PN to coupled DEVS transformation.

6.2.2 Dynamic of Resulting DEVS Model

The dynamic of generated DEVS model is controlled by the functions of DEVS formalism which are δint, δext and λ. After initialization of places (PDEVS) by the initial marking and after launching the evolution of the model by the event "initialize" received by all transitions (TDEVS), the latter are in state "checking" (by δext) to see if the number of tokens in places upstream is sufficient to achieve a crossing. Event "check" is sent by λ. After receiving the event, PDEVS transmit the number of their free tokens (which are not reserved by other transition) with λ as well. If the number of tokens is sufficient to validate the transition (TDEVS), the status is changing from "checking" to "reserving" and the event "reserve" is sent with λ. The firing does not occur directly. It must go through a reservation status to avoid conflicts (if places are upstream of several transitions), as long as the transitions are in continuous competition. In this way the properties of PN in terms of dynamics and competition is faithfully preserved in our transformation approach.

When PDEVS receives the event "reserve" it returns "ok" if there is still enough free tokens, otherwise, it returns "fail". If TDEVS receives at least one "fail", it returns immediately the signal "cancel" to release the reserved tokens. It puts its state "Validated" otherwise. At this point, the transition can pass the crossing and therefore returns "decrement" to PDEVS which will destroy the tokens reserved by TDEVS in question. It sends simultaneously "increment" to PDEVS located downstream in order to increment the number of tokens with the value received by the input port (weight of arc). After firing a TDEVS, it rehabilitates "checking" and so on.

Functions δ_{ext}, δ_{int}, δ_{con} and λ, characterizing the models TDEVS, are summarized in Table2. The first two columns represent the inputs, which are the events and the current state. The other columns show the outputs of each function. The table rows are grouped separately for each current state and models PDEVS. Functions are shown in Table 2. By convention, if all events have the same impact, we write "all events". Empty cells indicate the absence of values, for λ that means the absence of events and for δ_{ext}, δ_{int} and δ_{con} that the function does not produce an output state. The "&" symbol indicates that the events are simultaneous.

Event	**Current state**	**δ_{ext}**	**δ_{int}**	**δ_{con}**	**λ (current state)**
initialize	all states	checking			Out
pause		Paused			Out
Stop		Stopped			Out
release		checking			Out
Free tokens	Reserving		reserving	Reserving	reserve
Ok		validated, reserving		validated, reserving	
fail		canceling		Canceling	
all events	Validated		checking		decrement & increment & out
all events	Canceling		checking		cancel

Table 1: The outputs of the TDEVS model functions.

6.2.3 Example of Transformation

Fig. 4 and 5 present an example of transformation of one of famous case study in PN training field: Producer-Consumer (Prod_Cons_PN).

The formal definition of this PN is:

Prod_Cons_PN = (P, T, PRE, POST, M0)

P = {P1, P2, P3, P4, P5, P6}

T = {T1, T2, T3, T4}

$$PRE = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} POST = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} M_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 7 \end{pmatrix}$$

P1 (Producer is ready to produce), T1 (Begin of production), P2 (Production is run), T2 (End of production), P3 (plug containing products, initially, plug is empty), P4 (Consumer is ready to consume), T3 (Begin of consummation), P5 (Consummation is run), T4 (End of consummation) and P6 (Number of free puts, initially: all puts in plug are free).

Event	**Current state**	**δ_{ext}**	**δ_{int}**	**δ_{con}**	**λ (current state)**
initialize	all states	Checking		Checking	Out
check	checking	Checking	Checking	Checking	free_ tokens
reserve		Reserving		Reserving	
increment		Incrementing		Incrementing	
decrement		Decrementing		Decrementing	
cancel		Checking		Checking	
check	reserving	Reserving	Checking	Reserving	ok, fail
reserve		Reserving		Reserving	
increment		Incrementing		Incrementing	
decrement		Decrementing		Decrementing	
cancel		Checking		Checking	
check	incrementing	Checking	Checking	Checking	Out
reserve		Reserving		Reserving	
increment		Incrementing		Incrementing	
decrement		Decrementing		Decrementing	
cancel		Incrementing		Incrementing	
check	decrementing	Checking	Checking	Checking	Out

Table 2: The outputs of the PDEVS model functions.

Fig.4 represents the coupled model faithful to the PN modeling Producer-Consumer. Fig.5 illustrates the corresponding coupled DEVS model. We conserve the same color signification as shown in Fig. 3: Color green to initialize places' tokens number. Color orange to initialize transitions. Color red: to illustrate control stream. Color black: to illustrate tokens incrementing or decrementing and color blue for outputs.

6.2.4 Discussion

Petri nets are formal tools modeling dynamic systems dealing perfectly with the aspect of competition, concurrency and parallelism. Therefore, they require gentle handling during mapping in order to not lose their specifications. In our approach, competition is preserved by the creation of temporary state transitions which is the reserving state. Thus, a token cannot participate at the same time, in the firing of two transitions in conflict. However, the transition must immediately release tokens if it fails to be validated in order to not paralyze other transitions which are in conflict with it.

In this paper we presented the generalized PN for the reader to understand the mechanism of transformation. However, other extensions such as coloured PN can also be processed. In this case, tokens will no longer be trivialized. We will need to extend the type of representation to comprise a list with different colours. Thus, during the broadcast of the event "check" with a transition. Places of upstream should check the port connecting to the transition in order to send only the number of free tokens with the same colour as specified

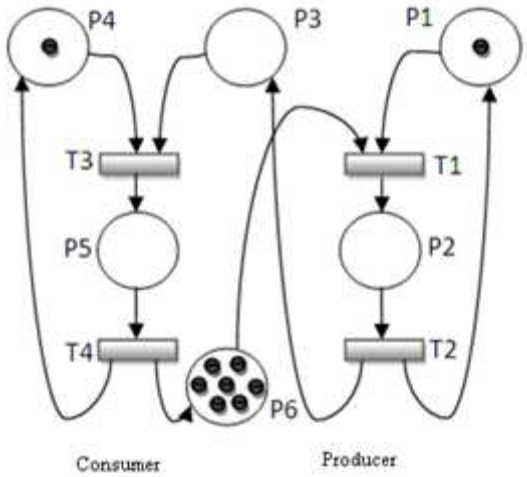


Figure 4: PN Producer-Consumer.

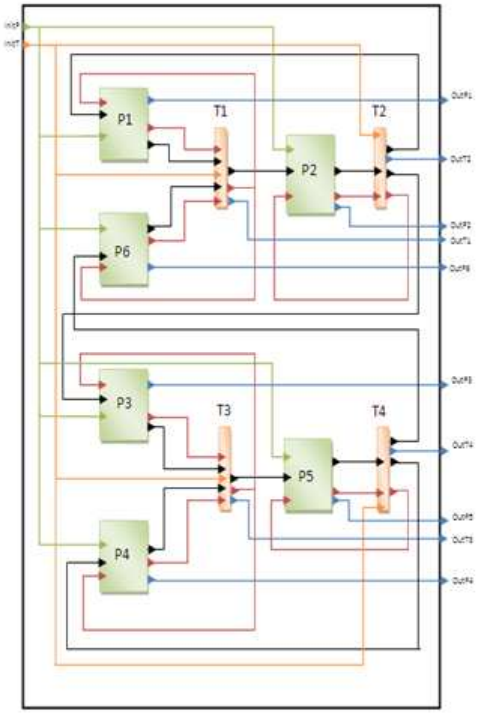


Figure 5: DEVS coupled models corresponding to

at this port.

In addition, the DEVS formalism provides flexibility in the internal structure of the models [30]. Models may disappear, others can take over. This aspect of dynamic structure related to DEVS will simplify the complexity of PN related to the representation of structural changes in systems. Therefore one DEVS model can represent several PNs at a time.

7 Conclusion and perspectives

In this paper we have presented a transformation approach of Petri nets to DEVS models, where places and transitions are transformed to atomic models. Coupling these models generates a coupled DEVS. This work falls within the framework of multi-modeling and transformation models based on multi-formalisms. Our choice of DEVS as focal formalism was based on its power in unifying and coupling models. Characterized by its abstraction, implementations independence and its ability to model complex systems in the form of a hierarchical model, DEVS is a formalism that can be the unifier of models.

By the transformation presented in this paper, the PN can enjoy the simulation on multiple DEVS based platforms.

Our perspectives focus on the implementation of such transformations to modelling complex industrial systems such as petroleum plants.

References

[1] Fishwick, P.A. (1995). Simulation Model Design and Execution: Building Digital Worlds. *Prentice Hall: Englewood Cliffs*, NJ.

[2] Vangheluwe, H. (2008). Foundations of modelling and simulation of complex systems. Electronic *Communications of the EASST, 10: Graph Transformation and Visual Modeling Techniques* .http://eceasst.cs.tuBerlin.de/index.php/eceasst/issue/view/19.

[3] Pidd, M. (2004). Systems Modelling: Theory and Practice. *John Wiley & Sons: Hoboken*, NJ.

[4] Fishwick, P.A. (2004). Toward an integrative multimodeling interface: A human-computer interface approach to interrelating model structures. *Simulation* **80**(9): 421.

[5] Murata, T. (1989). Petri Nets: Properties, Analysis and Applications. *Proceedings of the IEEE*, Vol.77,No.4 pp.541-580, April 1989.

[6] Peterson, J.L. (1977) : Petri nets, *Computing Surveys*. pp. 223–252

[7] Zeigler, B. P. (1976) : Theory of Modelling and Simulation, *Wiley InterScience*.

[8] Zeigler, B. P. Praehofer, H. and Kim, T. G.(2000): Theory of Modeling and Simulation, *Second edition. Academic Press, ISBN 0127784551*

[9] Shafagh, J. and Wainer, G.A. (2011). A Performance Evaluation of the Conservative DEVS Protocol in Parallel Simulation of DEVS-based Models., *Proceedings of 2011 Spring Simulation Conference (SpringSim11), DEVS Symposium*, page 103--110 - April 2011

[10] Boukelkoul, S. and Redjimi, M. (2013). Mapping Between Petri Nets and DEVS Models. *Proceeding of the 3rd International Conference on Information Technology & e-Service*, Sousse, Tunisia

[11] Vangheluwe, H. (2000). DEVS as a common denominator for multi-formalism hybrid systems modeling. *Conference IEEE International Symposium on Computer-Aided Control System Design*, Alaska, pp.129-134.

[12] De Lara, J. and Vangheluwe, H. (2002). Computer Aided Multi-Paradigm Modeling to Process Petri-Nets and Statecharts. *Lecture Notes in Computer Science*, *Springer* Volume 2505, pp 239-253.

[13] Home page: http://atom3.cs.mcgill.ca/ De Lara, J.,

[14] De Lara, J and Vangheluwe H. (2004). Meta-Modelling and Graph Grammars for Multi-Paradigm Modelling in AToM3. Manuel Alfonseca. Software and Systems Modeling, Vol 3(3), pp.: 194-209. *Springer-Verlag. Special Section on Graph Transformations and Visual Modeling Techniques.*

[15] De Lara, J., Vangheluwe, H. (2005): Model-Based Development: Meta- Modelling, Transformation and Verification, *The Idea Group Inc*, pp. 17 (2005).

[16] Wainer, G.A. and Mosterman, P. (2011) Discrete-Event Modeling and Simulation: Theory and Applications. *Taylor and Francis*.

[17] Jacques, C. J. D. and Wainer, G. A. (2002). Using the CD++ DEVS Tookit to Develop Petri Nets. *Proceedings of the SCS Summer Computer Simulation Conference*, San Diego, CA. U.S.A

[18] Shafagh, J. and Wainer, G.A.(2011). Conservative Synchronization Methods for Parallel DEVS and Cell-DEVS. *Proceedings of the 2011 ACM/SCS Summer Computer Simulation Conference*, The Hague, Netherlands.

[19] Genrich, H. J. and Lautenbach, K. (1981) System Modelling with High-Level Petri Nets. *Theoretical Computer Science*, vol. 13 (1981)

[20] Jensen, K. and Kristensen, L.M. (2009) Coloured Petri Nets Modelling and Validation of Concurrent Systems. *Springer*.

[21] Touraille, L., Traoré, M. K. and Hill, D. R. C. (2010). SimStudio: une Infrastructure pour la modélisation, la simulation et l'analyse de systèmes dynamiques complexes. *Research Report LIMOS/RR*-pp.10-13.

[22] Byun, J.H. Choi, C.B. and Kim, T.G. (2009) Verification of the devs model implementation using aspect embedded devs. *In Proceedings of the 2009 Spring Simulation Multiconference*, San Diego, USA, 2009

[23] Freigassner, R. Praehofer H. and Zeigler, B. P.(2000). Systems approach to validation of simulation models. *Cybernetics and Systems*, pp.52–57.

[24] Quesnel G. Duboz, R. and Ramat, E. (2009). The Virtual Laboratory Environment – An operational

framework for multi-modeling, simulation and analysis of complex dynamical systems. *Simulation Modeling Practice and Theory*, 17 :641–653.

[25] Quesnel, G. (2006). Approche formelle et opérationnelle de la multi-modélisation et de la simulation des systèmes complexes. PHD trésis Laboratoire d'Informatique du Littoral (LIL). Calais - France

[26] Sarjoughian, H. and Zeigler, B. P. (1998). Devsjava: Basis for a DEVS-based collaborative ms environment. *SCS International Conference on Web-Based Modeling and Simulation*, San Diego, CA, vol. 5, pp. 29-36.

[27] Ilachinski, A. (2001). Cellular Automata, a Discrete Universe, *World Scientific Publishing Co*, ISBN 981-02-4623-4.

[28] Schmidt, D. C (2006) .: Model-Driven Engineering *Guest Editor's Introduction IEEE Computer*, Vol. 39, No. 2, pp. 25-31.

[29] IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA)-Framework and Rules, Institute of Electrical and Electronics Engineers, IEEE (2000) 1516-2000

[30] Baati, L. (2007) : Approche de modélisation DEVS à structure hiérarchique et dynamique. *LSIS UMR-CNRS 6168, Domaine Universitaire de St Jérôme*.

Biologically Inspired Dictionary Learning for Visual Pattern Recognition

A. Memariani and C. K. Loo
University of Malaya, Kuala Lumpur, Malaysia
E-mail: ali_memariani@siswa.um.edu.my, ckloo.um@um.edu.my.

Keywords: holonomic brain theory, dictionary learning, sparse coding, quantum particle swarm optimization, complex-valued synergetic neural network, body expression.

Received: March 4, 2013

Holonomic brain theory provides an understanding of neural system behaviour. It is argued that recognition of objects in mammalian brain follows a sparse representation of responses to bar-like structures. We considered different scales and orientations of Gabor wavelets to form a dictionary. While previous works in the literature used greedy pursuit based methods for sparse coding, this work takes advantage of a locally competitive algorithm (LCA) which calculates more regular sparse coefficients by combining the interactions of artificial neurons. Moreover the proposed learning algorithm can be implemented in parallel processing which makes it efficient for real-time applications. A complex-valued synergetic neural network is trained using a quantum particle swarm optimization to perform a classification test. Finally, we provide an experimental real application for biological implementation of sparse dictionary learning to recognize emotion using body expression. Classification results are promising and quite comparable to the recognition rate by human response.

Povzetek: Z zgledovanjem po bioloških sistemih je predstavljena je metoda učenja vizualnih vzorcev.

1 Introduction

Neural structure has been one of the inspirations of machine learning. However, the concept of axonal discharge is misunderstood. Pribram's holonomic brain theory, proposes the term neuromodulator rather than neurotransmitter to refer to the electrical gab in junctions (axodendritic and dendo-dendritic) caused by chemical synapses. Accordingly arrival patterns of a nerve impulse are described as sinusoidal fluctuating hyperpolarizations (-) and depolarizations (+) which are inadequately large to make a nerve impulse discharge instantly [1]. Maps of these hyper and polarizations are called receptive fields. These receptive fields of visual cortex contain multiple bands of excitatory and inhibitory areas which act as line detectors. Thus neurons are tuned to a limited bandwidth of frequencies to provide harmonic features; In other words neurons behave like active filters sensitive to oriented lines, movements and colours rather than Euclidean-based geometric features. A specific shape could be represented as a combination of filter responses (2-D convolution integrals). A set of filters is called a dictionary, since elements of a dictionary are not orthogonal to each other, there are many redundant features to represent an image (overcomplete approximation). A more sparse representation is obtained by selecting the best features among those with high correlation with each other. And remove others. Following an iterative strategy, the sparse coded representation is generated in which selected features satisfy the orthogonality assumption.

This paper applies a locally competitive algorithm (LCA) [2] to extract the sparse coded definition of visual patterns. A synergetic neural network (SNN) is used to learn the visual features of a class of objects. SNN parameters are optimized with a quantum particle swarm approach.

1.1 Holonomic brain theory

The fact that for a harmonic oscillation we can either specify frequency or time (i.e. Heisenberg's principle of indeterminacy) has linked psychophysics and quantum mechanics. Gabor function is described as the modulation product of an oscillation with a given frequency (carrier) and an envelope in the form of normal distribution function. A biologically-plausible model for the visual pathway (retina, LGN, striate cortex) is described as a triple of convolutions. These triple convolutional preprocessing provides maximal coding of information. Biological Infomax visual cognition models such as independent component analysis (ICA) [3] and sparseness-maximization net [4] have better performance than classical Principal Component Analysis (PCA) or Hebbian models[5]. Relations between sparseness-maximization net and dendritic fields describes a dendritic implementation of sparseness-maximization net [6]; Though the dendritic

implementation is limited by infomax process which could be originated from top down lateral inhibition. Olshausen and Field formulate the reconstruction of stimuli in receptive fields of simple cell using sparse coding [4, 7]. Advantages of combining Gabor responses as in [4, 7] over ICA-like shapes are described by [6].

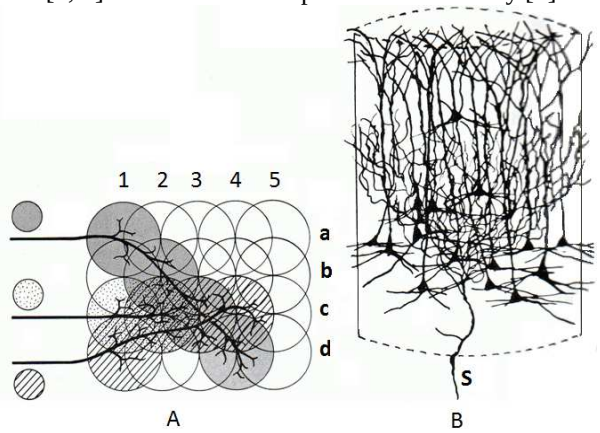


Figure 1: Microstructure of synaptic domains in cortex[1]. Overlapping line detectors (vertical and horizontal circles) combined to represent a stimulus (A) and interacting polarizations producing the dendritic fields (B).

Since then sparse coding is improved by many researchers. Though, most of them used greedy approaches to compute a sparse representation [8-10]. Accordingly, Biological realization of sparsification was unknown. However, in the recent work [2] a locally competitive algorithm (LCA) is proposed which is based on biological inhibition in neural circuits.

Table 1: Comparison of basic holonomic approaches[1].

	HNeT	Quantum Associative Net	ICA	field computing
Effectiveness	Very Effective	Effective	Very Effective	A general model with potentially very effective "sub-branches"
biological plausibility	Fundamental level only	Fundamental level only	bio-implausible but plausible output	fundamental level only
Possible quantum implementation	indirect similar core as QAN	direct	not yet known	partially direct
Main weakness	a mixture of natural and artificial features	limited to assoc. memory and pattern recognition	unknown bio implementability	Consciousness still missing

The striate cortex (V1) is the area of conscious visual perception in brain. Experimental results from functional magnetic resonance imaging (fMRI) supported that effect of the visual cortex in V1 in response to a stimuli can be estimated by a 2D Gabor function. A Gabor field I is the superposition of different Gabor functions' responses:

$$I = \sum_{j=1}^{M} a_j GW_j \quad (1)$$

where a_j and GW_j are Gabor coefficient and elementary Gabor function corresponding to the J_{th} element in the dictionary. The superposition of Gabor fields is in analogy to dictionary learning that represent the equation in similar form [4, 7, 9]. Therefore, the selection of Gabor coefficients can be performed by a sparse coding algorithm such as LCA so that an image is represented with minimum subset of Gabor elementary functions.

Output of V1 is projected to peri-striate cortex (V2) where probably retinal images are reconstructed. Triple-stage convolution in visual pathway has inspired convolutional neural networks acting as a course to fine process; though the research has focused mostly on magnitude data [11]. Some of the works included phase information to form an associative memory network [12]. Table 1 compares some of basic approaches of holonomic phase-magnitude encoding approaches.

Here we proposed a recognition algorithm based on the holonomic brain theory. Experimental results are compared to the state of the art algorithms. Furthermore, we applied the algorithm to recognize emotions based on body expression data which is inspired by the action based behavior in psychology. Classification results are compared to those of human recognitions.

2 Sparse coding

Representing an image with a few elementary functions is widely used in image processing and computer vision. Determining image component is useful to remove the noise. Also decomposition is used for compression by simplifying image representation.

In computer vision decomposition is a tool for feature extraction. An elementary function is called basis and set of bases functions is a dictionary. In early models choice of dictionary elements was subject to orthogonality condition. A complete representation of image is a linear combination of bases in the dictionary, derived by projection of image into bases. However, poor quality of representation in complete solutions resulted in relaxation of orthogonality condition and applying overcomplete dictionaries. Due to useful mathematical characteristics obtained by orthogonality (e.g. computing decomposition coefficients with projection), overcomplete dictionaries are still meant to be partially orthogonal. A common approach is to use an orthogonal subset of a large dictionary containing all possible elements.

Early works applied gradient descent to train the dictionary. Bayesian approaches also have been used to represent an image based on the MAP estimation of the dictionary components[13].

Textons are developed as a mathematical representation of basic image objects[14]. First images are coded by a dictionary of Gabor and Laplacian of Gaussian elements; Responses to the dictionary elements is Combined by transformed component analysis. Furthermore, sparse approximation helps to find a more general object models in terms of scale and posture[15].

Active basis model [16] provides a deformable template using Gabor wavelets as dictionary elements. They also proposed a shared sketch algorithm (SSA) inspired by AdaBoost.

2.1 Gabor wavelets

Biological models in object recognition are based on the findings of functional magnetic resonance imaging (fMRI) of mammalian brain. process of images in receptive fields (V1) is more sensitive on bar-like structures [17]. Responses of V1 are combined together by extrastriate visual areas and passed to inferotemporal cortex (IT) for recognition tasks. Research in computational neuroscience argued that recognition of objects in mammalian brain follows a sparse representation of responses to bar-like structures [4, 18]. Gabor wavelets are widely used as biologically inspired basis to model information encoding in receptive fields. 2D Gabor function cantered at (x_0, y_0) is:

$$G(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}e^{-\left[\frac{(x-x_0)^2}{\sigma_x^2}+\frac{(y-y_0)^2}{\sigma_y^2}\right]}e^{i[\xi_0 x+\upsilon_0 y]} \quad (2)$$

where (ξ_0,υ_0) is optimal spatial frequency. Using wavelet transform a Gabor function can be rotated, dilated or translated. General form of Gabor wavelet function is:

$$GW(x,y,\omega,\theta)=\frac{\omega}{\sqrt{2\pi}k}e^{\frac{\omega^2}{8k^2}(4(x\cos\theta+y\sin\theta)^2+(-x\sin\theta+y\cos\theta)^2)} \cdot\left[e^{i\omega(x\cos\theta+y\sin\theta)}-e^{-\frac{k^2}{2}}\right] \quad (3)$$

where ω is the radial frequency and θ is the wavelet orientation. κ is a constant representing bandwidth frequency [19]. Approximation of $\kappa\approx\pi$ and $\kappa\approx 2.5$ are common for 1 and 1.5 octave bandwidth (ϕ) respectively. Generally κ is:

$$\kappa=\sqrt{2\ln 2}\left(\frac{2^\phi+1}{2^\phi-1}\right) \quad (4)$$

A dictionary of Gabor wavelets (as shown in Fig.2), including n orientations and m scales is in the form of: $GW_j(\theta,\omega)$, $j=1,\ldots,m\times n$, where

$$\theta=\left\{\frac{k\pi}{n},k=1,\ldots,n-1\right\}, \quad (5)$$

and $\omega=\frac{\sqrt{2}}{i},i=1,\ldots,m.$

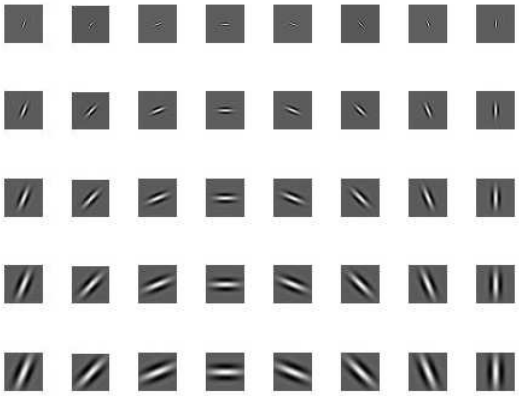


Figure 2: A dictionary of Gabor wavelets.

2.2 Sparse coding using locally competitive algorithm

Response to a dictionary of Gabor wavelets is an overcomplete representation. Sparse coding is the method of selecting a proper subset of responses to represent the image (signal). In addition to biological motivations, sparse coding is necessary to avoid redundant information. Having a fixed number of features, redundancy may cause loss of essential information which is going to be encoded in the lower levels (Fig.4).

Figure 3: Edge detection using Gabor wavelets, A. Original image[1], B. edge detected image with a large number of features without sparsity, C. edge detected image with a small number of features where sparsity is enforced.

Assuming an image (I_o) its sparse approximation I is derived according to (1). Optimal sparse coding tries to minimize the number of nonzero coefficients a_j, which is an NP-hard optimization problem.

We applied a locally competitive algorithm (LCA) [2] to enforce local sparsity. Unlike classical sparse coding algorithms, LCA uses a parallel neural structure inspired by biological model. LCA is applied to minimize the mean square error combined with a cost function in the local neighbourhood:

$$E(t)=\frac{1}{2}\left\|I(t)-I_0\right\|^2+\lambda\sum_j C\left(a_j(t)\right). \quad (6)$$

Thresholds are useful to generate coefficients with exact zero value.

For a threshold function $T_{(\alpha,\gamma,\lambda)}(.)$, cost function C is:

$$C_{(\alpha,\gamma,\lambda)}(a_j)=\frac{(1-\alpha)^2\lambda}{2}+\alpha\left|a_j\right|, \quad (7)$$

$$T_{(\alpha,\gamma,\lambda)}(u_j)=\frac{u_j-\alpha\lambda}{1+e^{-\gamma(u_j-\lambda)}} \quad (8)$$

Limit of T as $\gamma \to \infty$ is called ideal thresholding function. $T_{(0,\infty,\lambda)}(.)$ is hard tresholding function and $T_{(1,\infty,\lambda)}(.)$ is soft tresholding function.

In previous works, there is no real application that has been applied using LCA, although some simulation results are shown. Here an empirical experiment based real application of body expression recognition, is proposed to provide an evidence for the practical utility of Holonomic Brain Model as dictionary learning method by LCA.

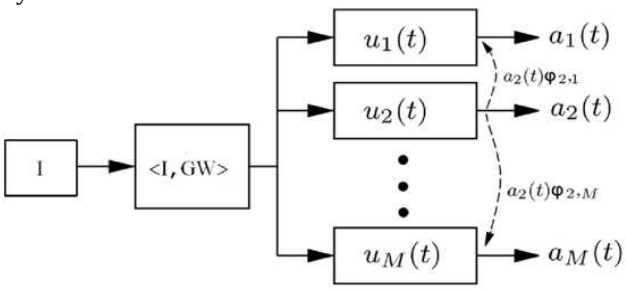


Figure 4: LCA structure [4].

LCA structure acts as a set of integrate and fire neorons. response to a dictionary of filter charges the internal state of the neurons and leads to the activity of the neuron. Neurons with higher charge (internal state) become active and fire signals to inhibit other neurons. A firing signal keeps other neurons that are highly corelated with the coresponding active neuron from being active by defusing their charge in an unidirectional inhibition.

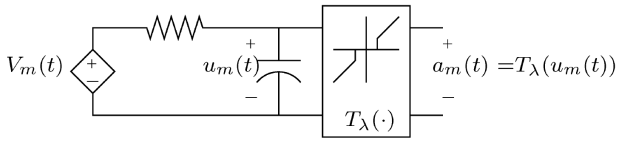


Figure 5: integration (charge up) and fire in a neoral circuit [2].

3 Locally competitive active basis recognition

We applied a supervised algorithm to recognize two types of objects in images; First a pixel-wise approach for aligned objects which combines the learned samples of objects in each class to form a prototype and second a feature based approach for non-aligned objects in which Gabor wavelets are localized to represent a potential match between specific scale and orientation and edges of objects. Both approaches are fed into a synergetic neural network to perform a classification task.

Images are scaled to have the exact same size. Each image is convolved with all the elements in the dictionary. Then sparse coding is enforced to minimize the representing elements for each pixel. Finally, remaining parts are reconstructed to generate the sparse superposition of the image. For pixel values in the local area LCA has the following steps:

1. Compute the response (convolution) of I_i with all the elements in the dictionary.

$$C_j = \langle GW_j, I_j \rangle \quad (9)$$

(Set t = 0 and $u_j(0) = 0$, for j = 1,…, n).

2. Determine the active nodes by activity thresholding.
3. For each pixel calculate internal state u_j of element j.

$$u_j = \frac{1}{\tau}\left[C_j(t) - u_j(t) - \sum_{j \neq k} \Phi_{j,k}.a_j(t)\right] \quad (10)$$

$$\Phi_{j,k} = \langle GW_j.GW_k \rangle \quad (11)$$

4. Compute sparse coefficients $a_j(t)$ for $u_j(t)$.

$$a_j(t+1) = T_\lambda(u_j(t)) \quad (12)$$

$$T_{(\alpha,\gamma,\lambda)}(u_j) = \frac{u_j - \alpha\lambda}{1 + e^{-\gamma(u_j - \lambda)}} \quad (13)$$

5. If $a_j(t-1) - a_j(t) > \delta$ then $t \leftarrow t+1$ and go to step 2, otherwise finish.

Original SNN used pixel-wised features to represent an object which is not robust in case objects are in a variable shapes (e.g. different body emotions of human). In this case, we construct a template model as a collection of Gabor wavelet features included in the dictionary which represents the general characteristics of all body posture classes. Test images are convolved with the components of the template model. Sparsity is then enforced to catch the best fit over the specific posture. LCA thresholding strategy enables us to remove redundancies effectively (producing sparse coefficients with exactly zero values). Number of output Gabor wavelets are fixed in order to make the comparison with trained prototype of each class. Features are selected based on their highest response to the training images; furthermore, each feature is allowed to perturb slightly in terms of location and orientation. In this aspect our template construction is a modification of shared sketch algorithm [16]. For each image i feature value v_{ij} corresponded to the selected Gabor wavelet j, is determined as the following:

$$v_{ij} = \gamma_i C_{ij} - \log(Z(\gamma_i)) \quad (14)$$

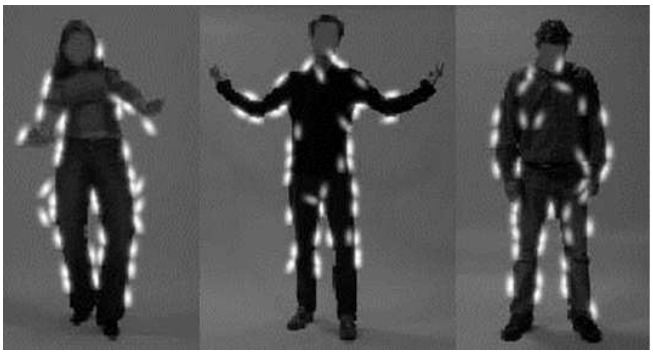


Figure 6: Gabor wavelet features detecting the edge pattern of different body postures.

where γ_i is derived by maximum likelihood estimation and Z is the partition function. Therefore, boundaries of

object are segmented out before the result is given to SNN.

3.1 Complex-valued synergetic neural network

Synergetic neural network (SNN) developed by Haken [20] describes the pattern recognition process in the human brain which tries to achieve the learned model with fast learning and no false state rather than traditional neural networks [21, 22] [20].

A common approach to combine learned samples is averaging the feature values. One way to deal with inflexibility is to use learning object in the same view which will restrict the classification task. A melting algorithm is proposed by [23] to combine objects in deferent poses. Suppose a learned sample object $\hat{I}$ consists of n pixel values. $\hat{I}$ is reshaped to a column vector v_i and normalized so that:

$$\sum_{j=1}^{n} v_{ij} = 0 \tag{15}$$

$$\sum_{j=1}^{n} \left| v_{ij} \right|^2 = 1 \tag{16}$$

A prototype $V^{\dagger}$ is the Hermitian conjugate of V:

$$V^{\dagger} = (V^T V)^{-1} V = C(v) + iS(v) \tag{17}$$

A test samples q corresponding to a test image is normalized and compared to the prototype of each class, using the order parameters. For each prototype k order parameters ε_k is initialized as:

$$\varepsilon_k = v_k^{\dagger}.q \ , \ k = 1,...,m. \tag{18}$$

where $v_k^{\dagger}$ is the k_{th} row in the Hermitian conjugate $V^{\dagger}$. Order parameters are updated derived iteratively with the synergetic dynamics:

$$\dot{\varepsilon}_k = \frac{1}{D}(\lambda_k - D + B\varepsilon_k^2)\varepsilon_k + \varepsilon_k \tag{19}$$

$$D = (B + C)\sum_k \varepsilon_k^2 \tag{20}$$

where λ_k is the attention parameter for class k; B, C are constants [24]. Attention parameters could be considered balance (equal and mostly unit) or unbalance.

Attention parameters in the model are trained using a quantum particle swarm optimization in order to minimize the overall classification error in the test set.

3.2 Centroidal Voronoi Tessellation (CVT)

As mentioned in section 3.1 unbalance attention parameters should be tuned. We applied a CVT in order to cover the whole feasible space in the initial state of the random search. A set of generators are considered as a group of points in the space forming a Tessellation. Generators are associated with subsets and points are nearer to its corresponding generators rather than any of other generators according to the distance function (*e.g.*, the l_p norm). Note that the generators are not quite evenly distributed throughout the space. Dividing the feasible space into the partitions, several generators set at almost precisely the same point in the space. CVT overcomes the poor and non-uniform distribution of some Voronoi cells by choosing the generators at centre [25-27]. Assuming $\lambda_{\max}$ as the maximum potential attention parameter search space is defined as:

$$0 < \lambda_i < \lambda_{\max}, i = 1,...,m. \tag{21}$$

Given a set of Voronoi regions $T_\xi (\xi = 1,...,\Xi)$ in the space $\Omega \subset R^m$, each initial position p_ξ is the Centroid of its region.

$$T_\xi = \left\{ x \in \Omega, \left| x - p_\xi \right| - \left| x - p_{\bar{\xi}} \right| \ for \bar{\xi} = 1,...,\Xi, \bar{\xi} \neq \xi \right\} \tag{22}$$

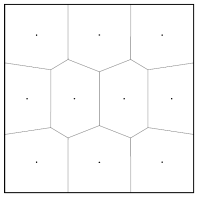


Figure 7: Centroidal Voronoi tessellation dividing a square into 10 regions[28].

3.3 Quantum particle swarm optimization (QPSO)

Initial attention parameters are tuned using a QPSO in order to minimize the overall classification error in the test set. Each particle position X, is updated based on the movement framework in the quantum mechanics.[29] State of the particle is described by a wave function,

$$\psi(Y) = \frac{1}{\sqrt{L}} e^{-\frac{|Y|}{L}} \tag{23}$$

$$Y = X - p \tag{24}$$

$$L = \frac{h^2}{my} \tag{25}$$

where γ is called intensity of the potential well at point p, m is the particle mass and h is a constant. Finally, for particle i, j_{th} element of the position $x_{i,n}^j$ can be updated as:

$$X_{i,n+1}^{j} = p_{i,n}^{j} \pm \frac{L_{i,n}^{j}}{2} \ln\left(\frac{1}{u_{i,n}^{j}}\right), \tag{26}$$

$$u_{i,n+1}^{j} \approx U(0,1), \tag{27}$$

$$L_{i,n}^{j} = 2\alpha \left| X_{i,n}^{j} - C_n^j \right| , \tag{28}$$

$$C_n^j = \frac{1}{M} \prod_{i=1}^{M} p_{i,n}^{j} , \tag{29}$$

where C_n^j is the average of all particle positions.α is a positive real number which could be constant or change dynamically in total N iteration as:

$$\alpha = \frac{0.5*(N-n)}{N} + 0.5 \tag{30}$$

To improve the accuracy an adaptive penalty function [30] is added to the overall error:

$$Z = f(\Lambda) + \sum_{i=1}^{m} k_j v_j \tag{31}$$

$$k_j = \left| f(\Lambda) \right| + \frac{v_j(\Lambda)}{\sum_{i=1}^{m} v_j(\Lambda)^2} \tag{32}$$

$$\Lambda = (\lambda_1, ..., \lambda_k) \tag{33}$$

Figure 8 shows an overview of the recognition method. QPSO is used to iteratively tune the attention parameters $\lambda_i, i = 1, ..., k$ where k is the number of classes.

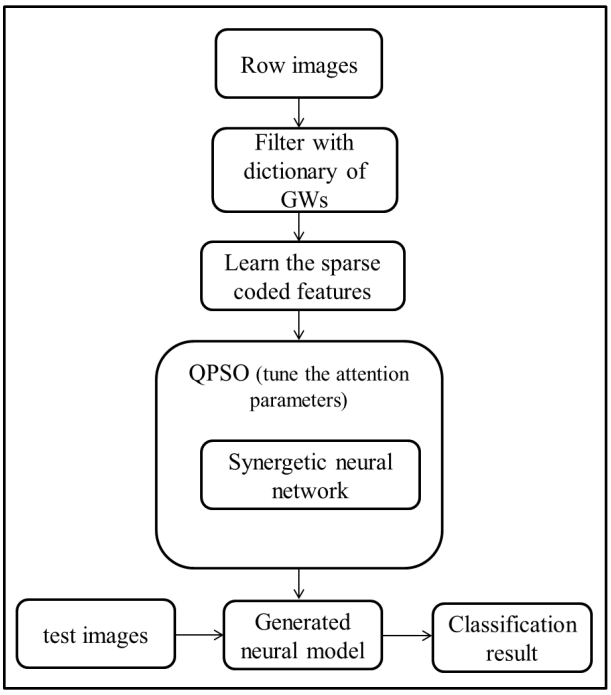


Figure 8: Scheme of proposed visual recognition model.

4 Emotion recognition using body expression and results

Even though most of the works in the area of emotion recognition has been focused on facial expressions, some of psychological theories considered emotional appraisals that are not facially expressive [31-33]. In that sense, emotions are described based on the state of action readiness that they cause in the whole body (either impulsive or intentional)[31]. Intentional actions might differ person to person though impulsive actions only depend of the nature of their action readiness.

Accordingly, impulsive actions can be used to recognize emotions considering the body expressions. Facial expression has been combined with upper body gestures to recognize emotions [34]. Movements of hands are detected using color segmentation and represented by centroid of the area; face components is also detected using skin detection techniques. Facial features (eyebrows, mouth, chin, etc.) are then combined with hand movements to set up the features. similar works has consider body feature along with facial features for fear detection [35] and anger detection [36, 37].

Body gestures are also merged with speech based features derived by acoustic analysis. Together with facial expressions [38] developed a framework in which face and body data was recorded with different resolutions and synchronized with subjects' speech interaction. They applied a Bayesian classifier to recognize the emotions.

Kleinsmith et al [39] argued that emotions can be recognized by humans from body postures when their face is removed. They also developed a recognition model to recognize the affection of faceless avatars in computer games.

Figure 9: Extracted features for four classes of emotions top down as, anger, fear, happiness, sadness.

Human actions caused by emotions could be detected using point-light animations [40]. Ross et al perform a test to compare recognition ability of students in primary and secondary schools and adults [41]. Faces of the test subjects were covered and recognition task performed on both full-light display and a point-light display where only main parts of the body postures are shown in a black and white format. Their result shows that adults have a better ability of bodily emotion recognition and display full-light is more expressive than in point-light for the task.

Table 2: Classification Accuracies for different QPSOs.

	Anger (%)	Fear (%)	Happiness (%)	Sadness (%)	Overall Error (%)
QPSO1	92.31	68.97	72.0	93.10	18.35
QPSO2	36.54	93.10	62.0	93.10	27.52
QPSO3	92.31	72.41	74.0	93.10	16.97
QPSO4	36.54	93.1	64.0	93.10	27.06
QPSO5	92.31	86.21	60.0	93.10	16.51
QPSO6	92.31	68.97	72.0	93.10	18.35
QPSO7	36.54	94.83	64.0	93.10	26.61
QPSO8	82.69	89.66	66.0	93.10	16.51
BEAST (Human Recognition)	93.6	93.9	85.4	97.8	

In order to validate the perception of body expression tests have been developed and validated by human recognition. Atkinson et al developed a dataset for both static and dynamic body expressions; The dataset contains 10 subjects (5 female) and covers five emotions (anger, disgust, fear, happiness and sadness)[42]. The bodily expressive action stimulus test (BEAST) [43] provides a dataset for recognizing four types of emotions (anger, fear, happiness, sadness) which is constructed using non-professional actors (15 male, 31female). Body expressions are validated with a human recognition test.

We applied a supervised approach to recognize two types of objects in images; First a pixel-wise approach for aligned objects which combines the learned samples of objects in each class to form a prototype and second a feature based approach for non-aligned objects in which Gabor wavelets are localized to represent a potential match between specific scale and orientation and edges of objects (figure 9). Both approaches are fed into a synergetic neural network to perform a classification task.

We applied the BEAST data set[1] to classify four classes of basic emotions. Gabor wavelets are generated in a (20, 20) matrix and images are resized to have 500 pixels in row and relatively scaled pixels in column. Images are divided into train and test sets for each class 10 images are selected randomly to form the train data and the rest are included for test. Different scenarios are considered to train the model:

1. Static QPSO with $\alpha = 0.75$ and randomly initialized.
2. Static QPSO with synergetic melting prototype [44].
3. Dynamic QPSO where α changes according to (29) and randomly initialized.
4. Static QPSO with $\alpha = 0.75$ and initialized with CVT.
5. Dynamic QPSO as (29) and initialized with CVT.
6. Dynamic QPSO as (29), initialized with CVT and a synergetic melt prototype.
7. Static QPSO with $\alpha = 0.75$, initialized with CVT and penalized with (30).
8. Dynamic QPSO as (29), initialized with CVT and penalized with (30).

Classification accuracies of different trained SNNs are compared with results of human recognition (table2). In some cases happiness and anger are misclassified as fear, this happened more frequently in static learning. However regardless of the learning scenario, happiness turns to be the most difficult one to detect and the reason is not clear for the authors.

Figures 10 and 11 show the learning rate for each scenarios during the learning iterations. CVT has improved the accuracy with Dynamic learning scenario.

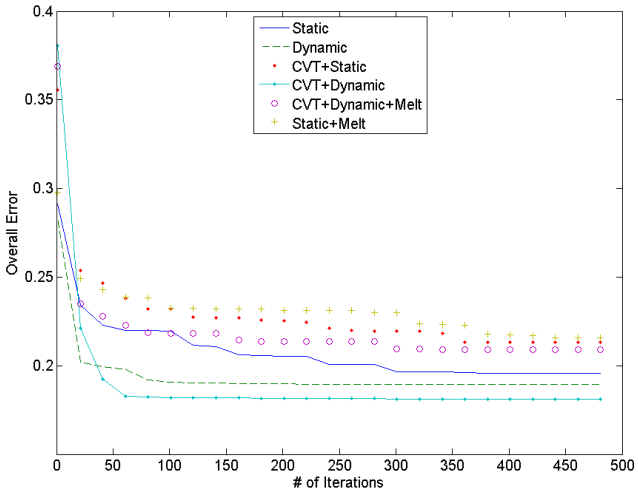


Figure 10: Average learning rates for different QPSOs.

[1] http://www.beatricedegelder.com/beast.html

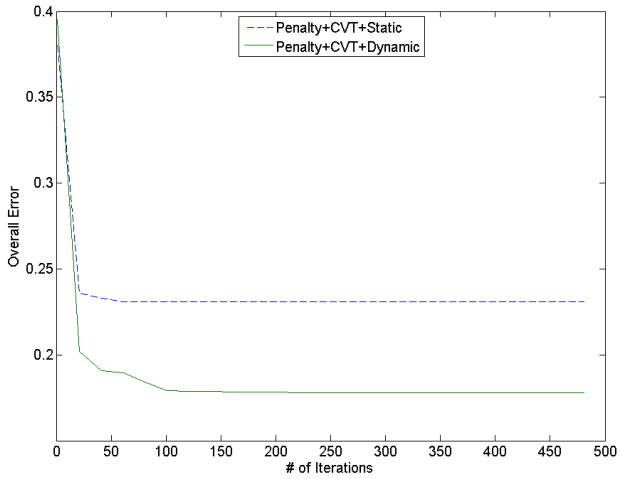


Figure 11: Average learning rates for penalized objective functions.

5 Conclusion

We proposed a biologically-plausible approach for recognition of aligned and non-aligned objects. Our dictionary learning algorithm is inspired by the holonomic brain theory. LCA is applied to enforce sparsity on a dictionary of Gabor wavelets. Regarding the parallel structure of the learning method, implementation could be optimized via parallel processing which is essential for real-time applications.

Furthermore, synergetic neural network is combined with Gabor wavelet features which make it applicable for recognition of non-aligned objects. Gabor features also enhance the SNN to use images with different size for both construction of the Hermitian conjugate and test images. Effect of background is also removed because of recognition is based on the pattern of edges; Though sparse coding is robust in presence of classical noise since dot noise does not follow any meaningful shape pattern intrinsically.

Experimental results supported the real application of Holonomic Brain Model as dictionary learning method using a biological implementation.

Acknowledgment

This work is supported by Flagship research grant of University of Malaya (FL006-2011) “PRODUCTIVE AGING THRU ICT” and HIR-MOHE research grant (H-22001-00-B000010) of University of Malaya.

References

[1] Pribram, K.H., *Brain and perception : holonomy and structure in figural processing*1991, Hillsdale, N.J.: Lawrence Erlbaum Associates. xxix, 388 p.

[2] Rozell, C.J., et al., *Sparse coding via thresholding and local competition in neural circuits.* Neural Computation, 2008. **20**(10): p. 2526-2563.

[3] Makeig, S., et al., *Independent component analysis of electroencephalographic data.* Advances in neural information processing systems, 1996: p. 145-151.

[4] Olshausen, B.A. and D.J. Field, *Emergence of simple-cell receptive field properties by learning a sparse code for natural images.* Nature, 1996. **381**(6583): p. 607-609.

[5] Peruš, M. and C.K. Loo, *Biological and Quantum Computing for Human Vision: Holonomic Models and Applications*2010: Medical Information Science Reference.

[6] Peruš, M., *Image processing and becoming conscious of its result.* Informatica, 2001. **25**: p. 575-592.

[7] Olshausen, B.A. and D.J. Field, *Sparse coding with an overcomplete basis set: A strategy employed by V1?* Vision Research, 1997. **37**(23): p. 3311-3325.

[8] Chen, S.S.B., D.L. Donoho, and M.A. Saunders, *Atomic decomposition by basis pursuit.* Siam Journal on Scientific Computing, 1998. **20**(1): p. 33-61.

[9] Olshausen, B.A. and D.J. Field, *Sparse coding of sensory inputs.* Current Opinion in Neurobiology, 2004. **14**(4): p. 481-487.

[10] Mallat, S.G. and Z.F. Zhang, *MATCHING PURSUITS WITH TIME-FREQUENCY DICTIONARIES.* Ieee Transactions on Signal Processing, 1993. **41**(12): p. 3397-3415.

[11] Haykin, S.S., *Neural networks: a comprehensive foundation*1994: Macmillan.

[12] Hopfield, J.J., *Neural networks and physical systems with emergent collective computational abilities.* Proceedings of the National Academy of Sciences, 1982. **79**(8): p. 2554-2558.

[13] Kreutz-Delgado, K., et al., *Dictionary learning algorithms for sparse representation.* Neural Comput., 2003. **15**(2): p. 349-396.

[14] Zhu, S.C., et al., *What are textons?* International Journal of Computer Vision, 2005. **62**(1-2): p. 121-143.

[15] Figueiredo, M.A.T., *Adaptive sparseness for supervised learning.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2003. **25**(9): p. 1150-1159.

[16] Wu, Y.N., et al., *Learning Active Basis Model for Object Detection and Recognition.* International Journal of Computer Vision, 2010. **90**(2): p. 198-235.

[17] Riesenhuber, M. and T. Poggio, *Neural mechanisms of object recognition.* Current Opinion in Neurobiology, 2002. **12**(2): p. 162-168.

[18] Daugman, J.G., *TWO-DIMENSIONAL SPECTRAL-ANALYSIS OF CORTICAL RECEPTIVE-FIELD PROFILES.* Vision Research, 1980. **20**(10): p. 847-856.

[19] Tai Sing, L., *Image representation using 2D Gabor wavelets.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1996. **18**(10): p. 959-971.

[20] Haken, H., *Synergetic Computers and Cognition: A Top-Down Approach to Neural Nets*2004: Springer.

[21] Koruga, D., et al. *Synergy of classical and quantum communications channels in brain: neuron-astrocyte network*. in *Neural Network Applications in Electrical Engineering, 2004. NEUREL 2004. 2004 7th Seminar on*. 2004.

[22] Bin, L. and T. Yuru. *The research of learning algorithm of synergetic neural network.* in *Computer Science and Information Processing (CSIP), 2012 International Conference on*. 2012.

[23] Hogg, T., D. Rees, and H. Talhami. *Three-dimensional pose from two-dimensional images: a novel approach using synergetic networks.* in *Neural Networks, 1995. Proceedings., IEEE International Conference on*. 1995.

[24] Gao, J., et al., *Optical-electronic shape recognition system based on synergetic associative memory.* 2001: p. 138-148.

[25] Richards, M. and D. Ventura. *Choosing a starting configuration for particle swarm optimization*. in *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*. 2004.

[26] Du, Q., V. Faber, and M. Gunzburger, *Centroidal Voronoi Tessellations: Applications and Algorithms.* SIAM Rev., 1999. **41**(4): p. 637-676.

[27] Nguyen, H., et al., *Constrained CVT meshes and a comparison of triangular mesh generators.* Comput. Geom. Theory Appl., 2009. **42**(1): p. 1-19.

[28] Burkardt, J., et al., *User manual and supporting information for library of codes for centroidal Voronoi point placement and associated zeroth, first, and second moment determination.* SAND Report SAND2002-0099, Sandia National Laboratories, Albuquerque, 2002.

[29] Sun, J., et al., *Quantum-Behaved Particle Swarm Optimization: Analysis of Individual Particle Behavior and Parameter Selection.* Evolutionary Computation, 2012. **20**(3): p. 349-393.

[30] Barbosa, H.J.C. and A.C.C. Lemonge, *A new adaptive penalty scheme for genetic algorithms.* Inf. Sci., 2003. **156**(3-4): p. 215-251.

[31] Frijda, N.H., *THE LAWS OF EMOTION.* American Psychologist, 1988. **43**(5): p. 349-358.

[32] Frijda, N.H., *Not Passion's Slave.* Emotion Review, 2010. **2**(1): p. 68-75.

[33] Frijda, N.H., *Impulsive action and motivation.* Biological Psychology, 2010. **84**(3): p. 570-579.

[34] Gunes, H. and M. Piccardi, *Bi-modal emotion recognition from expressive face and body gestures.* Journal of Network and Computer Applications, 2007. **30**(4): p. 1334-1345.

[35] van Heijnsbergen, C.C.R.J., et al., *Rapid detection of fear in body expressions, an ERP study.* Brain Research, 2007. **1186**(0): p. 233-241.

[36] Pollick, F.E., H. Paterson, and P. Mamassian, *Combining faces and movements to recognize affect.* Journal of Vision, 2004. **4**(8): p. 232.

[37] Paterson, H.M., F.E. Pollick, and E. Jackson., *Movement and faces in the perception of emotion from motion*, in *Perception, ECVP Glasgow Suppl*2002. p. 232.

[38] Kessous, L., G. Castellano, and G. Caridakis, *Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis.* Journal on Multimodal User Interfaces, 2010. **3**(1): p. 33-48.

[39] Kleinsmith, A., N. Bianchi-Berthouze, and A. Steed, *Automatic Recognition of Non-Acted Affective Postures.* Ieee Transactions on Systems Man and Cybernetics Part B-Cybernetics, 2011. **41**(4): p. 1027-1038.

[40] Blake, R. and M. Shiffrar, *Perception of human motion*, in *Annual Review of Psychology*2007, Annual Reviews: Palo Alto. p. 47-73.

[41] Ross, P.D., L. Polson, and M.-H. Grosbras, *Developmental Changes in Emotion Recognition from Full-Light and Point-Light Displays of Body Movement.* PLoS ONE, 2012. **7**(9): p. e44815.

[42] Atkinson, A.P., et al., *Emotion perception from dynamic and static body expressions in point-light and full-light displays.* Perception, 2004. **33**(6): p. 717-746.

[43] de Gelder, B.V.d., Stock J., *The Bodily Expressive Action Stimulus Test (BEAST). Construction and Validation of a Stimulus Basis for Measuring Perception of Whole Body Expression of Emotions.* Frontiers in Psychology, 2011. 2:181. doi:10.3389/fpsyg.

A Novel Similarity Measurement for Iris Authentication

Mohamed Mostafa Abd Allah
Minia University, Faculty of Engineering, Egypt
Department of Electrical, Communications and Electronics section
E-mail: mmustafa@yic.edu.sa

Keywords: iris authentication, genuine and impostor pairs, similarity measurement

Received: July 7, 2013

This paper introduces a novel similarity measurement which derives the likelihood ratio between two eyes. The proposed method takes into consideration the individual and system error rates of eye features. It handles two kinds of individual probabilities: (consistent Probability (CP), the Inconsistent Probability (IP),) to achieve the best matching approach between two feature sets. While calculating the probabilities, we assume that a reasonable alignment approach should be obtained before the matching approach introduced. The proposed matching algorithm is theoretically proved to be optimal, and experimental results show that the proposed method has more efficient performance on separating genuine and impostor pairs

Povzetek: Predstavljena je nova metoda za prepoznavanje identitete očes.

1 Introduction

The iris is the color part of the eye behind the eyelids, and in front of the lens. It is the only internal organ of the body which is normally externally visible. Whose unique pattern is stable after age one. Compared with other biometric features such as the face and the fingerprint, iris patterns are more stable and reliable. Iris recognition systems are non-invasive to their users, but require a cooperative subject. For this reason, iris recognition is usually used for verification or identification purposes, rather than for a watch list that is, a large database with which individuals are compared to determine if they belong to a selected group, such as terrorists. Iris recognition is gaining acceptance as a robust biometric for high security and large-scale applications [1][2]. Most classical algorithms verify a person's claimed identity by measuring the features between two iris [2], which consist of two stages: alignment and matching. The alignment stage employs a special pattern matching approach to achieve the best alignment between two feature sets. The matching stage compares the feature sets under the estimated transformation parameters and returns a similarity score using a constructed similarity measurement. If the similarity score is larger than an acceptance threshold, the two irises are recognized as a genuine pairs, otherwise the claimed identity is rejected. Associating with the similarity threshold, there are two error rates: False Match Rate (FMR) and False Non-match Rate (FNMR). FMR denotes the probability that the score of an impostor pair is larger than the threshold. FNMR denotes the probability that the score of a genuine pair is less than the threshold. The overall FMR and FNMR for a set of eyes are the integration or average of the FMR and FNMR for all individual eyes in the data set. Conventional methods construct the similarity measurement with simple decisions [3] or multi-decisions based on fusing the similarity scores of different features [5], which use one unified threshold for all eyes to make the final decision. Their similarity thresholds are experimentally determined to assure that the average error rates are lower than a required level, while the individual error rates of some eyes are higher than this required level although the average error rates for all eyes are sufficient. The difficulty of constructing the similarity measurement is that the threshold which balances the tradeoff between the overall FMR and FNMR may not be optimal for each individual eye and thus not optimal for the overall FMR and FNMR of all eyes.The rest of this paper is organized as follows. In section II, iris alignment algorithm regard transformation parameters have been presented. In section III, iris matching algorithm presents the estimation of consistent Probability (CP) and Inconsistent Probability (IP) under the assumption of No/High correlation. Section IV conducts several experiments to evaluate the proposed method. Conclusion has been presented in section V.

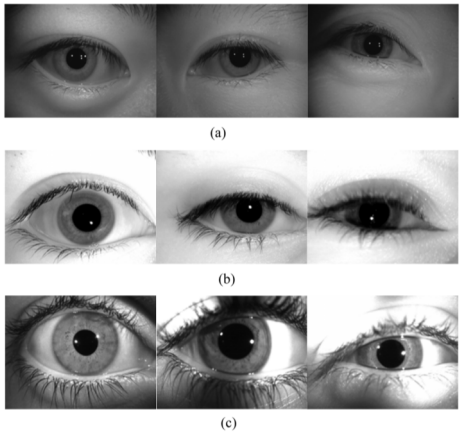


Figure1: Examples of iris images.

2 Alignment approach

Most previous matching methods, suffer from memory requirement, time consuming and computationally exhaustive processes. This is because that the distribution of matching scores is evaluated in every possible transformation. This paper, assume that a reasonable alignment approach should be obtained before the Matching, to overcome such problems and provides a fast and memory-efficient matching process.

The proposed method, defines vector representation of Template iris features (T), Input iris features (I), and Transformed iris features (S') as following:

$\mathbf{T}=\{\mathbf{t}_1, \mathbf{t}_2,\ldots, \mathbf{t}_m \mid i=1..m\}$, $\mathbf{I}=\{\mathbf{s}_1, \mathbf{s}_2,\ldots, \mathbf{s}_n \mid j=1..n\}$, and $S'=(s'^{j}_{x}, s'^{j}_{y}, s'^{j}_{\theta})$.

Let, $F_{\Delta x,\Delta y,\Delta\theta}$ that formulated in Eq. 1., be the geometrical transformation function that maps $\mathbf{s_j}$ (input iris features) into $\mathbf{s_j}'$ (transformed iris features).

$$\begin{pmatrix} s'^{j}_{x} \\ s'^{j}_{y} \\ s'^{j}_{\theta} \end{pmatrix} = \begin{pmatrix} \cos\Delta\theta & -\sin\Delta\theta & 0 \\ \sin\Delta\theta & \cos\Delta\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} s^{j}_{x} \\ s^{j}_{y} \\ s^{j}_{\theta} \end{pmatrix} + \begin{pmatrix} \Delta x \\ \Delta y \\ \Delta\theta \end{pmatrix} \quad (1)$$

Hough transform alignment approach [9] uses an accumulator array $A(p, q, r)$ to counts and collect alignment scores of each transformation parameter $\Delta x,\Delta y,\Delta\theta$ respectively. In practice, each transformation parameter is discredited into a finite set of values: $\Delta x = \{\Delta x_1, \ldots, \Delta x_P\}$, $\Delta y = \{\Delta y_1, \ldots, \Delta y_Q\}$ and $\Delta\theta = \{\Delta\theta_1, \ldots, \Delta\theta_R\}$. A direct implementation of a 3-D Hough transform alignment approach [8] is infeasible for embedded devices with limited memory budget. Suppose that P=256, Q=256 and R=128, then 8,388,608 memory units are required for such implementation. Obviously, to overcome such problems and provides a memory-efficient process, a new alignment technique should be proposed.

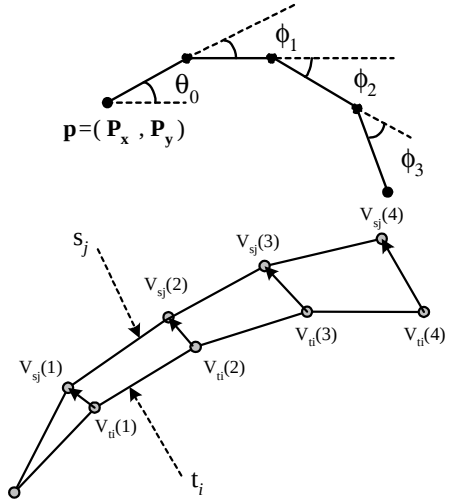


Figure 2: The iris distances representation between two iris features vectors.

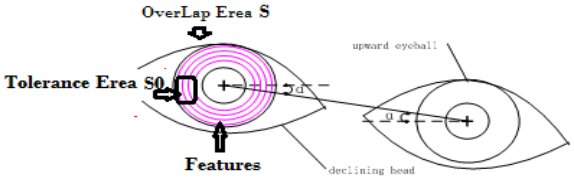


Figure 3: The distribution of features positions after reasonable alignment.

2.1 Proposed alignment approach

The proposed alignment approach with multi-resolution accumulator array could greatly reduce the amount of required memory units. For each value of $\Delta\theta_r$, there is an exactly one shift vector $(\Delta x_p, \Delta y_q)$ of each pair $(\mathbf{t_i}, \mathbf{s_j})$ such as given by Eq. 2. Therefore, 2-D accumulator array B with entry $B(p, q)$ is enough to evaluate accumulation of alignment at rotation $\Delta\theta_r$. For all possible rotation that could done in a specific tolerance area S_0, the proposed approach accumulate evidence value into the array B and the maximum alignment score will represent the best geometrical transformation alignment between **I** & **T**. Applying this computation method reduces memory requirement to 4,096 memory units.

$$\begin{pmatrix} \Delta x_p \\ \Delta y_q \end{pmatrix} = \begin{pmatrix} t^{j}_{x} \\ t^{j}_{y} \end{pmatrix} - \begin{pmatrix} \cos\Delta\theta_r & -\sin\Delta\theta_r \\ \sin\Delta\theta_r & \cos\Delta\theta_r \end{pmatrix} \begin{pmatrix} s^{j}_{x} \\ s^{j}_{y} \end{pmatrix} \quad (2)$$

Memory optimization result is not only giving advantages for smaller memory requirement of the proposed approach but also offering a faster alignment peak detection process. Detecting alignment peak value in a smaller Hough space is faster than one in the conventional method [4].

2.2 Two vector similarity measure

Although several kinds of features can be extracted from Iris image [3][6][8][9], the proposed approach introduces a novel measurement of iris contour features. The feature representation in this proposal offers alternative matching criteria between two vectors called the *similarity measure* (*sM*). The proposed matching criteria are derived by accumulating spatial differences between the corresponding trace points of two vectors. As shown in Figure. 2(a), the proposed Iris features is approximated represent by piece-wise linear segments extracted along the iris contour [4]. The vector representation of Iris contour feature **S** can be given as:

$$\mathbf{S} = (\mathbf{P_x}, \mathbf{P_y}, \theta_0, \phi_1, \phi_2, \phi_3) \quad (3)$$

Where, (P_x, P_y) represent as feature position, s_θ as the contour orientation and ($s_{\Phi 1}$, $s_{\Phi 2}$, $s_{\Phi 3}$) as the orientation differences of two adjacent linear segments. As shown in Figure .3, if T ($\mathbf{P_{tx}}$, $\mathbf{P_{ty}}$, θ_{t0}, ϕ_{t1} , ϕ_{t2} , ϕ_{t3}), and S ($\mathbf{P_{sx}}$, $\mathbf{P_{sy}}$,θ_{s0}, ϕ_{s1} , ϕ_{s2} , ϕ_{s3}) represent the template and input irises vectors in the tolerance overlapped area *O*. A pair vector (*K*) from T are considered to be mated with corresponding features from S if and only if their accumulating spatial differences (*aD*) is equal or smaller than the tolerance threshold D_0 and the *direction difference* (*dD*) between them is smaller than an angular tolerance θ_0.

These tolerance thresholds (D_0 and θ_0) are necessary to compensate the unavoidable errors from image processing and features extraction algorithm. From the accumulated distances, $aD=\sum_k V(k)$, we derive the similarity *sM* as follows:

$$aD(\mathbf{t}_i,\mathbf{s}'_j) = f(Dist, \Delta\phi_1, \Delta\phi_2, \Delta\phi_3)$$

$$sM(\mathbf{t}_i,\mathbf{s}'_j) = \begin{cases} f(aD) & aD(\mathbf{t}_i,\mathbf{s}'_j) \le D_0 \\ 0 & \text{others} \end{cases} \quad (4)$$

where,

$\Delta\phi_1 = \phi_{s1} - \phi_{t1}$

$\Delta\phi_2 = \phi_{s2} - \phi_{t2}$

$\Delta\phi_3 = \phi_{s3} - \phi_{t3}$

$Dist(\mathbf{s}, \mathbf{t}) = |\Delta\phi_1| + |2\Delta\phi_1 + \Delta\phi_2| + |3\Delta\phi_1 + 2\Delta\phi_2 + \Delta\phi_1|$ (5)

The *sM* function returns value from 0 (different) to a constant positive value *maxSim* (same).

3 Probability matching approach

While calculating the probabilities, we assume in the overlapped area *O*, there are M features from template iris, and N features from input iris. A tolerance area of features spatial distance is assigned as S_0. The probabilities that a randomly distributed *M* features from template iris corresponds with one of the N features from input iris in the overlapped area *O* can be estimated by two aspects: Iris consistent probability and Iris inconsistent probability.

3.1 Iris consistent probability

Assume that template and input irises are originated from different eyes and have no correlation between each other. If the consistent probability result is large enough, the two eyes are represented as an impostor pair. Therefore, if there are $i-1$ arbitrarily features from T located in *O*, and all of which are mated with features from S, the rest overlapped area can be represented with $O-(i-1)S_0$, and the unmated randomly distributed features number of S in *O* is represented with $N-(i-1)$. In additional, the probability that the $i-th$ randomly distributed features from T in *S* corresponds to one of the $N-(i-1)$ features from S in the overlapped area *O* can be denoted with:

$$\frac{N-(i-1)}{E-(i-1)} \ (i=1\ldots\ldots K) \ \&\ (E=\frac{O}{S0}) \quad (6)$$

Since the corresponding pairs K between T & S under the estimated transformation parameters, the rest consistent probability can be considered as unmated features. The probability that the $(K+1)-th$ randomly distributed features from T does not correspond to any *features* from S in the overlapped area *O and* can be represented by:

$$\frac{E-N}{E-K} \quad (7)$$

The probability that the $(K+j)-th$ feature from T is randomly distributed in the rest overlapped area $O-(K+j)S_0$

and does not correspond to any feature from S in *O* can be calculated with:

$$\frac{E-(N+j)}{E-(k+j)} \ (j=1\ldots\ldots M\text{-}K) \quad (8)$$

Therefore, the *Iris Consistent Probability* between template and input irises under the assumption that T and S have no correlation can be given as:

$$Pcp(S \neq T) = C_M^K \prod_{i=1}^{k} \frac{N-(i-1)}{E-(i-1)} \prod_{j=1}^{M-K} \frac{E-(N+j)}{E-(k+j)} \quad (9)$$

3.2 Inconsistent probability

Assume that template and input irises are originated from the same eye and have high correlation between each other. If the inconsistent probability result is large enough, the two irises are represented as a genuine pair. Considering that the poor quality irises detected during iris acquisition and feature extraction may cause some truth features to be missing or spurious features to be detected, we assume the truth features from iris T and S in the overlapped area *O* are m and n, respectively. Thus, the spurious features counts in iris T and S are $M-m$ and $N-n$. For the truth features between T and S, there should be someone to one correspondence between each other. But due to the existence of eye deformation, features position change and features missing, there are position gaps between the corresponding features of two irises even for genuine pairs. The position gaps of the missing truth features are treated as ∞. We assume that the truth features, which located inside the tolerance

threshold r_0 are h and the truth features, which located outside r_0 are g. then conditions $g + h \leq \min(m, n)$ are satisfied.

Where $h \in [0, \min(K,m, n)]$, and $g \in [0, \min(m, n) - h]$. (10)

For the spurious features from T and S in O, there may happen that some spurious features of T located inside the tolerance area of some of those in S. Since the number of corresponding features pair between T and S is Q, the mated spurious features can be represented by:

$Q - h \leq \min(M - m, N - n)$ Could be satisfied . . (11)

Consider all the features count in the overlapped area O, the identical truth features $\geq \max(m, n)$ and can be calculated as $m + n - (h + g)$. The identical spurious features) $\geq \max(M - m, N - n)$ and can be calculated as $(M - m) + (N - n) + (K - h)$. In practice, the total features count in O is thus calculated with $M + N - K - g$.

3.3 Probability distribution

Since the Probability Distribution of the positional differences in corresponding features extracted from mated irises is similar to Gaussian distribution [3] [8].

The probability that the position difference with respect to the corresponding features exceeds the tolerance threshold r_0 to be represented with:

$$1 - \int_0^{r0} G(r)dr \quad (12)$$

where $G(r)$ is the probability of position difference for mated features. Therefore, the probability that truth features (h) *that* are located inside r_0 and truth features (g)that are located outside r_0 is calculated by:

$$P_{TF} = C_{h+g}^{h} P_{PD}(sd \leq r0) P_{PD}(sd > r0) \quad (13)$$

For the spurious features, since there is no one to one correspondence between each other, the probability calculation can be accomplished by replaced M by $M - m$, N is replaced by $N - n$ and S is replaced by $S + [h - (m + n)]S_0$. Therefore, the probability that the $i - th$ randomly distributed spurious features of $M - m$ from T in $S + [h - (m + n)]S_0$ corresponds to one of the $(N - n) - (i - 1)$ spurious features from I is denoted with:

$$\frac{(N-n)-(i-1)}{E+\big(h-(m+n)\big)-(i-1)} \quad (i = 1 \dots\dots\dots K-h) \quad (14)$$

For the un-mated spurious features, M is replaced by $M - m$, N is replaced by $N - n$, S is replaced by $S + [h - (m + n)]_0$, and K is replaced by $K - h$. The probability that the $(K - h + j) - th$ spurious features of $M - m$ from T is randomly distributed in the rest overlapped area $S + [h - (m + n)]S_0 - (K - h + j)S_0$ and does not correspond to any spurious features of $N - n$ from I in $S + [h - (m + n)]S_0$ is derived by:

$\frac{E+(h-(m+n))-((N-n)+j)}{E+(h-(m+n))-((K-h)+j)}$ (j=1……((M-m)-(K-h))) . . (15)

Therefore, the probability that $K - h$ spurious features are mated and $(M - m) - (K - h)$ spurious features are un-mated between $M - m$ and $N - n$ spurious features from T and I is calculated as:P_{SF}=

$$C_{M-m}^{K-h} \prod_{i=1}^{k-h} \frac{(N-n)-(i-1)}{E+\big(h-(m+n)\big)-(i-1)} \prod_{j=1}^{M-m-K-h} \frac{E+\big(h-(m+n)\big)-((N-n)+j)}{E+\big(h-(m+n)\big)-((K-h)+j)} \quad (16)$$

The IP between T and I under the assumption that T and I are highly correlated is given by:

$$Pip(I = T) = \sum_{m=0}^{M} \sum_{n=0}^{N} \sum_{h=0}^{b} \sum_{g=0}^{a-h} p(m, n, h, g) \quad (17)$$

where

$$p = \begin{cases} C_{\geq max(m,n)+\geq max(M-m,N-n)}^{\geq max(m,n)} C_{\geq max(m,n)}^{h+g} \text{PTF} x C_{\geq max(M-m,N-n)}^{K-h} P_{SF} \\ 0 \; else \end{cases}$$

4 Experimental results

The proposed technique has been tested over 4320 images. The iris data are captured from 60 people by using three different kinds of iris sensors (BERC, CASIA V1.0, and CASIA-Iris V3). 24 iris image samples per person for each sensor are captured. That mean the total field test data were 60person x 8Iris x 3samples x 3 sensor = 4320 iris Image. The size of Iris is 128×128pixels. In the feature extraction process [4], a pattern is extracted from each iris image using the linear predictive analysis of an 8-pole filter. Firstly, we compare the proposed approach with two existing methods [8] and [9]. The three methods are implemented into a same Iris-based verification system. We use total field test data to construct the evaluation, in which there are number of genuine and impostor matches. The performances of different methods are shown in a representation of the ROC curves, which are plotted as FAR against FRR, as shown in Figure .4. From the ROC curves, it can be observed that the proposed algorithm causes the most improvement. With a given FAR, the proposed approach can help the system to obtain the lowest FRR. Statistically, compared with the other two systems, the proposed algorithm can reduce the system FRR when FAR=0.01%. Secondly, we investigate evaluating iris image quality. , and the measure becomes larger in clear iris image, and smaller in faded image. Figure 5 shows ROC curves correspond to application of image-quality parameter. Under the terms of (a) (without examination in image-quality), which means we don't reject faded images. (b)Examining both registered and verification data (all Iris images). (c)Examining the images, which should be registered only?. Recognition rate is improved from 95.6% to 99.3%.

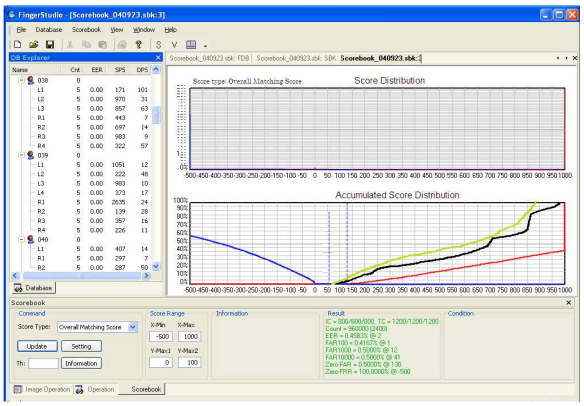


Figure 4: FAR & FRR evaluation result of the proposed approach.

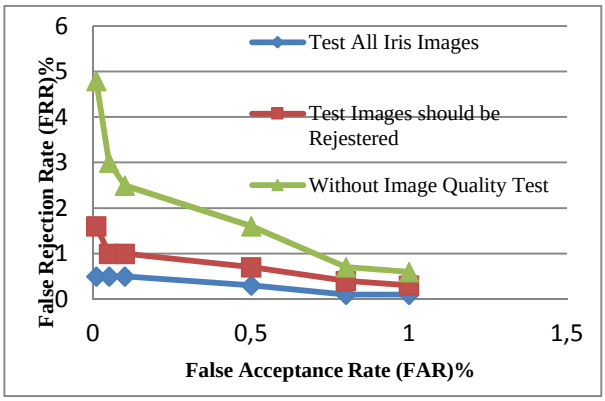


Figure 5: ROC curve vs. application of Image Quality Parameters.

5 Conclusion

The proposed alignment approach which using features vector representation generates a higher peak in Hough space than a conventional vector representation. Hence, an accumulator array with lower resolution could be employed without suffering difficulty of alignment. The proposed approach evaluation result FAR & FRR as shown in Figure .4, work as better as some previously presented approaches. We have been Applied the proposed discriminate algorithm to iris verification device which operates in real world. This evaluation makes it possible that the proposed approach can be implemented into an embedded system, such as DSP-based iris identification module. As shown at figure 5, Comparing with other methods, the proposed method can obtain the best performance for separating the genuine and impostor, which benefits from the utilization of CP and IP to construct the likelihood ratio. This paper invent a method to utilize parameters groups that has a relation with iris image quality and iris image information to got a perfect enrollment procedure results in the capture of the highest quality iris image(s). Another merit of the proposed approach is that it does not depend on the sensor type. Therefore, the proposed approach is more robust and implemental in practice.

References

[1] A. K. Jain, A. Ross, and S. Prabhakar (2004). An Introduction to Biometric Recognition, IEEE Trans. Circuits and Systems for Video Tech., vol. 14, pp. 4 – 20.

[2] K. W. Bowyer, K. Hollingsworth, and P. J. Flynn (2008). Image understanding for iris biometrics: A survey Computer Vision and Image Understanding, 110(2):281 – 307.

[3] Y. Du (2006). Using 2d log-gabor spatial filters for iris recognition. In Proc. of the SPIE Biometric Technology for Human Identification III, pages 62020:F1–F8.

[4] Yukun Liu, Dongju Li, Tsuyoshi Isshiki and Hiroaki Kunieda, (2010) ”A Novel Similarity Measurement for Minutiae-based Fingerprint Verification”," IEEE Trans. on Circuits and Systems for Video Technology,vol.14, No.1, pp.86-94.

[5] K. P. Hollingsworth, K. W. Bowyer, and P. J. Flynn. (2009). The best bits in an iris code. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(6):964–973.

[6] L. Ma, T. Tan, Y. Wang, and D. Zhang (2004). Efficient Iris Recogntion by Characterizing Key Local Variations. IEEE Transactions on Image Processing, 13(6):739–750.

[7] L. Masek (2003). Recognition of human iris patterns for biometric identification. Master’s thesis, University of Western Australia.

[8] C. Rathgeb, A. Uhl, and P. Wild (2011). Shifting score fusion: On exploiting shifting variation in iris recognition. In Proc. Of the 26th ACM Symposium On Applied Computing (SAC’11), pages 1–5.

[9] A. Uhl and P. Wild (2010). Enhancing iris matching using levenshte in distance with alignment constraints. In Proc. of the 6th Int. Symp. on Advances in Visual Computing (ISVC’10),pages 469–479.

[10] S. Ziauddin and M. Dailey (2008). Iris recognition performance enhancement using weighted majority voting. In Proc. of the 15th Int. Conf. on Image Processing (ICIP ’08), pages 277– 280.

[11] A. M. Bazen, R. N. J. Veldhuis, (2004). Likelihood-Ratio-Based Biometric Verification, IEEE Trans. on Circuits and Systems for Video Technology, vol.14, No.1, pp.86-94.

Usability Testing Tools for Web Graphical Interfaces

Carlos Teixeira, Bernardo Santos and Ana Respício[1]
Department of Informatics, University of Lisbon 1749-016 Lisboa, Portugal
[1]Operations Research Center, University of Lisbon 1749-016 Lisboa, Portugal
E-mail: cjteixeira@fc.ul.pt

Keywords: usability testing tools, user-centered interaction design, graphical web, multi-criteria decision analysis

Received: February 16, 2013

Software design and development following a user-centered approach can benefit from the adoption of adequate usability testing tools. However, the choice of a suitable tool for a particular purpose can be a difficult task, due to the multiplicity of such tools, each one offering a variety of different features. This paper surveys usability testing tools for web graphical interfaces, selects a set of appropriate tools and evaluates them. A set of relevant evaluation features is identified and aggregated into criteria. A multi-criteria additive utility function and the Analytical Hierarchy Process are proposed as evaluation methods and for establishing a ranking of a selected set of usability testing tools. Results of both methods are presented and compared.

Povzetek: Prispevek predstavlja pregled orodij za spletne grafične vmesnike.

1 Introduction

The user-centered design process relies on the involvement of users in every dimension that could be related to the success of the product. As human issues are always a main source of complexity for engineering, the size and heterogeneity of designers' team is often a requirement and another source of problems in itself. In order to overcome this small additional source of complexity, designers should cooperate according to some common guidelines built on their experience and a vast literature of recommendations, in a productive way that should provide convergence of results toward the final product (Norman, 2002).

Long lasting design teams have their own stabilized strategies, tactics and tools, partly established on the acquired experience with previous projects. New teams or teams with several new collaborators can take extra benefits from commercial off-the-shelf, well documented frameworks of integrated computer tools. When it concerns user-centered design of web interfaces, advanced prototypes, the final product and the users, can be directly accessed by robust common frameworks. These frameworks are repeatedly used, project after project, by the same teams. Even if teams are often remixed in their composition, a reliable framework, well understood by all the personal, will decrease the distance in the gulf that separates the evaluation protocols and the corresponding collected data from the team intuition about the problems and the innovations for their solutions.

Evaluation of a product relying on users tests (usability testing) is an irreplaceable technique in user-centered design (Shneiderman, 1998; Nielsen, 1993), since it gives direct input on how real users interact with the system (Nielsen, 1993).

There are many usability testing tools (UTTs) available nowadays, with different features and capacities. This paper is an attempt to organize the concerned information and choose a suitable usability testing tool for web interfaces (Nielsen, 1999; Dix et al., 2003), with particular emphasis on graphical interaction.

The evaluated UTT issues and features and the corresponding preferences were established by a restricted number of experts with the aim of conveying the usability tests of interfaces designed for prototypes developed by the World Search Project (World Search Project, 2010). This is a Portuguese project of almost 2 million euros investment which is responsible for the design of search interfaces for dedicated areas of public concern, namely in the health area. The goal of the World Search Project is the research and development of innovative web search technologies in Portugal as well as the research and development of generic and business information with semantic relevance and with the proper knowledge of the Portuguese language, culture and market.

The second section presents the issues and features considered for evaluation and comparison of UTTs. The third section surveys usability testing tools and presents the selected set of UTTs. The evaluation methods adopted are described in Section 4. The fifth section presents and discusses the results obtained insofar. Final section presents conclusions and some directions of future work.

2 Main issues and features for UTT evaluation

Many issues and features are relevant for building a comprehensive usability testing tool. Figure 1 is a tentative graphical representation of the main issues considered. These were represented as a flow as close as possible from the temporal order where designer's plans must be implemented.

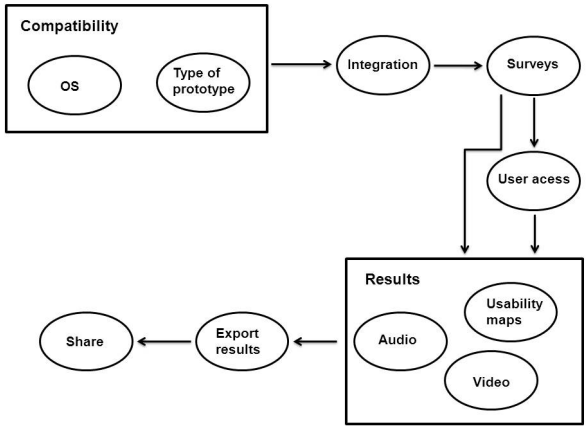


Figure 1: Usability testing tools issues.

When adopting any new software, anyone first concern will go into compatibility issues such as OS compatibility. Specifically for the design process, it is important to integrate several types of possible prototypes, giving a wide space of freedom to the designers, while facilitating several options of integration with all kind of surveys, questionnaires and alerts. A flexible integration can promote high quality testing. For instance, integrating the tests within the application (ex: using Javascript) can increase the dynamics of the usability tests as well as the quality of possible tests when compared to submitting screenshots to the UTT. Another relevant issue is the type of surveys produced by the UTT, the extent and the kind of questions allowed in the surveys that will be used to produce results. Our aim is to perform usability tests, having access to users located across the country or even abroad. Thus, user access is also a main issue to be considered. Concerning collecting results, three types of input are relevant: usability maps, which contribute to the analysis of users' interaction with the application; video recording, that is fundamental for tracing users' actions in the display and simultaneously recording facial expressions while interacting; and, audio recording, for collecting voice information produced by the user along with the interaction and consequently producing annotations (essential for the think-aloud protocol). As our goal is to evaluate interfaces with graphical interaction, a higher importance is given to features concerning collecting video from display, as well as generating usability maps including clicks and mouse movements. Finally, it is aimed that the format used to export the results is adequate for the subsequent analysis. Features concerning results' formats are aggregated by the issue "Export results", which also includes features related with the possibilities of sharing results ("Share") with the developers and designers teams (project partners). The survey of Vraa (Vraa, 2009) identifies important features and functionalities relevant for UTT evaluation. Many of these were also considered in present contribution.

To summarize, the following lines enumerate main issues (criteria) considered and the features (sub-criteria) within each of them:

1. OS Compatibility: Windows; Linux; Mac OS.
2. Supported types of prototypes: Applications; Prototypes; Screenshots of the interface; Wireframes; Mock-up's.
3. Interface integration with the UTT: Offline program (off-line test generation and managing); Website post (the URL to be tested is submitted to the UTT website); Uploaded images (screenshots submission); JavaScript code (that forwards information to an on-line account of the UTT website); Online wizard (all details of the interface; associated tasks are submitted to the UTT website in a pre-specified order).
4. User access (to the usability tests): Local; Remote; On-line.
5. Creation and submission of surveys and tasks for the users: Complete survey; Screen aligned questions (kind of pop-up with questions during specific passages of the usability test); Screen aligned text (kind of pop-up with questions during specific passages of the usability test);
6. Collecting audio: Record (both user and wizard-of-Oz /prototypes/ etc.); Annotations.
7. Collecting video: Display; Facial Expressions; Eye Tracking; Annotations.
8. Usability maps: Clicks; Mouse move; Scroll reach; Attractive zones; Interest zones; Attention zones; Form inputs.
9. Export: XLS/CSV/TSV; XML; Database; Share (online access management to results for the development team).

3 Selected UTT

This section describes the process of selecting the UTT candidates for the present study, which was inspired by several interesting web articles starting with Vraa (Vraa, 2009), Fadeyev (Fadeyev, 2009) and Tomlin (Tomlin, 2009). In the following years related articles were also published on-line by Walker (Walker, 2010), Gube (Gube, 2011), Jules (Jules, 2011) and LeMerle (LeMerle, 2012).

Table 1 displays in the first row our list of 23 candidates and the considered UTT reviews in the first column. Each UTT discussed by a given review is highlighted with an 'x' mark in the corresponding cell.

The list of candidates elected for evaluation was mainly based on the review of Tomlin (Tomlin, 2009) that extensively describes UTT in terms of features, presenting several plans of prices. Some of the Tomlin UTTs are not included in our candidates. The *Clixpy* and *Simple Mouse Track* websites were not found. The *Google Website Optimizer* and the *UserVue* were merged

review	UTT	*Concept Feedback*	*Chalkmark*	*ClickHeat*	*ClickTale*	*Crazyegg*	*Ethnio*	*Feng-GUI*	*Fivesecondtest + NavFlow + ClickTest*	*Feedback Army*	*Loop 11*	*Mechanical Turk*	*Morae (include User Vue)*	*Open Hallway*	*Silverback*	*Usabilla*	*Userfly*	*User Testing*	*Google Analytics (WebSiteOptimizer)*	*Intuition HQ*	*4Q Survey*	*Mouse Flow*	*Attention Wizzard*	*Click density*
(Tomlin, 2009)		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x					
(Vraa, 2009)											x						x							
(Fadeyev, 2009)			x	x	x				x	x					x		x	x	x					
(Walker, 2010)		x			x	x	x	x	x	x			x	x	x	x	x		x					
(Gube, 2011)			x						x	x	x					x				x				
(Jules, 2011)		x	x	x	x	x		x	x	x	x		x		x	x	x	x			x	x	x	x
(LeMerle, 2012)			x	x	x				x	x					x		x	x						

Table 1: Usability testing tools reviews and candidates selected for evaluation.

into *Google Analytics* and the *Morae*, respectively. The *Website Grader* was conceived in order to enhance online marketing websites, which is not within the scope defined in this paper introduction. *Fivesecondtest* is now available with two complementary applications *NavFlow* and *ClickTest*, which can be seen as a single UTT (the *UsabilityHub* from *Angry Monkeys*).

Vraa (Vraa, 2009) presents and discusses the best "Do's and Don'ts for Web Design and Usability" naming "16 crucial web design and usability best practice compilations and tools".

Though Vraa only reviews two UTT, the extended discussion on crucial UTT features inspired us in the identification of evaluation criteria and relevant features.

In the same year, Fadeyev (Fadeyev, 2009) surveys ten affordable UTT, claiming that "testing for usability is the only reliable way to find out how well a website works". Walker (Walker. 2010) also describes some of the already reviewed UTT and added a few more, whose main goals were to improve the visibility of websites for marketing purposes and thus were not included in our list of candidates. Gube (Gube, 2011) reviews the "22 essential tools for testing your website's usability" by classifying them into six categories.

1. User Task Analysis: *Intuition HQ*, *Usabilla*, *Loop11* and *Fivesecondtest*.
2. Readability: "*Juicy Studio: Readability Test*", *WordsCount* and *Check My Colours*.
3. Site Navigability: *Websort.net*, *OptimalSort*, *Chalkmark*, *WriteMaps*, *NavFlow* and *PlainFrame*;
4. Accessibility: "*Juicy Studio: Local Tools*", *VisCheck*, *W3C Markup Validation Service*, *WebAnywhere* and *Browsershots*.
5. Website Speed: *Pingdom Tools* and *Page Speed Online*.
6. User Experience: *Feedback Army* and *UserVoice*.

OptimalSort was already considered as part of the *Chalkmark* package. Other UTT referred were discarded, mainly because they were designed to evaluate specific aspects and not to support a significant coverage of all required usability issues.

Jules (Jules, 2011) presents the "best website usability testing tools and services", reviewing four UTT of our list that hadn't been previously discussed. The ten "essential website usability tools" discussed by LeMerle (LeMerle, 2012) were also analysed during this study.

Besides the preliminary analysis of the descriptions in web pages articles, the official websites for each of the selected candidates were also analysed. In order to assure the presence (or absence) of the features under assessment, all the content available was analysed, namely the videos demonstrating the UTT features.

4 Evaluation method

A simple additive utility function was used for providing a score on each UTT.

$$UF(UTT) = \sum_{j=1}^{m} w_j \sum_{k=1}^{n_j} w_{j,k}\, s_{j,k}\,(UTT)$$

This function linearly weights binary attributes $s_{j,k}(UTT)$ corresponding to the presence of elementary UTT features (0 for inexistent / 1 for implemented) using a two level hierarchy of weights. The second level $w_{j,k}$ weights the presence of feature *k* within the main issue *j*. Considering that issue j aggregates n_j features $\sum_{k=1}^{n_j} w_{j,k} = 1$. The first level aggregates the evaluation of *m* main issues where w_j is the weight determining the impact of the *j*-th main issue on the evaluation of the given UTT, where $\sum_{j=1}^{m} w_j = 1$.

The highest values found for this function should indicate the most suitable UTTs for our usability evaluations.

4.1 Utility model

The preferences (scores) for the main issues as well as for the features were set using an integer quantitative

scale. Table 2 displays the correspondence between the quantitative values used and their qualitative importance.

Quantitative	Qualitative
5	Crucial
4	Important
3	Significant
2	Minor
1	Irrelevant

Table 2: Quantitative versus qualitative scale for setting preferences.

Weights were obtained by normalizing preferences into the interval [0;1]. Considering the preference for feature *k* within an issue *j*, represented by $p_{j,k}$, the corresponding weight is obtained by $w_{j,k} = p_{j,k} \,/\, \sum_{k=1}^{n_j} p_{j,k}$, where n_j is the number of features aggregated in issue *j*. This normalization ensures the equality $\sum_{k=1}^{n_j} w_{j,k} = 1$. Similarly, the weight for a main issue was computed as its relative contribution for the sum of all issues' preferences, thus ensuring $\sum_{j=1}^{m} w_j = 1$.

Preferences were obtained in two rounds by a team of three experts working for the project and having responsibilities in the task of interface design. All of them have a large experience in the development of software (ten or more years). In the first round each expert set up his/her own preferences in a printed form. The resulting printed forms were shared among the team. In a second round all the experts together discussed their scores until they agreed in a final number according to the quantitative scale of Table 2. In the remaining text we will refer to the above described scoring system as the Utility Model (UM).

4.2 Analytical hierarchy process

UM assumes criteria to be preferentially independent. The Analytical Hierarchy Process (AHP) (Saaty, 2005) also uses a linear additive model, but instead of giving absolute weights, the experts are questioned for pairwise comparisons of criteria and alternatives. This seems to be a much reasonable approach, namely because absolute values given in a single evaluation have very few references for providing the desired overall balanced result. Our AHP results were computed using a free trial version of commercial software (*Expert Choice Comparison*, 2012). This software considers all scores and makes all weights computations using a percentual scale. The pairwise comparison scale uses a judgment of preferences including nine categories: "extremely" preferred, "very strongly to extremely ", "very strongly", "strongly to very strongly", "strongly", "moderately to strongly", "moderately", "equally to moderately" and "equally" preferred.

A rating scale was used to score sub-criteria: the null value was assigned whenever a feature is absent; otherwise the score was set to 1. Though the AHP model has been criticized due to inconsistencies that can arise from weighting and scoring, we found easy to overcome them through a careful analysis and comparison setting. Again the preferences were set up in a collaborative meeting.

5 Results

5.1 Utility model

Table 3 presents the most significant results obtained by using the UM. The first column displays the main issues considered and the features aggregated under each issue. The second column presents the preferences specified for issues and features, on a 1-5 scale according to Table 2.

The UTTs under evaluation are presented in the first line and have been ranked according to their final scores, which were computed using the utility function and normalized to a 1-10 scale (last line). The column for each UTT also displays information about the presence or absence of each feature, represented by a 1 or a null value in the corresponding cell, respectively; and the values of relative scores for issues.

The best scored UTT, *Morae*, although providing limited user access was not excluded from our analysis because it presents good scores in almost all the other issues. However, this limitation may restraint remote or online usability tests, which is a major requirement in this project. Final decision about the election of the UTT to adopt should be based on testing the UTT since, at the present stage, our evaluation was mainly supported by industrial advertising information. Analogously, *Loop 11*, ranked in second place presents high preferences in the majority of issues. It was not excluded from the evaluation, despite not offering features for collecting audio – another important feature. The best ranked next three UTTs, *User Testing*, *Userfly* and *Usabilla*, also present good scores, offering all the required functionalities, even in a limited way. *Usabilla* is an exception as it does not provide audio collecting or video recording, which can be too confining.

Collecting additional information and testing the UTTs would be advantageous to support a final decision, as this study was mainly supported by industrial advertising information. Even considering the limitations above, Table 3 still provides a fair ranking suggestion for UTT selection, but then we present a new model based on the results of comparison.

Criteria and features	**preferences / UTT**	*Morae*	*Loop 11*	*User Testing*	*Userfly*	*Usabilla*	*Silverback*	*Click density*	*Intuition HQ*	*4Q Survey*	*Mouse Flow*	*Open Hallway*	*Google Analytics*	*ClickTale*	*Chalkmark*	*Fivesecondtest*	*Ethnio*	*Mechanical Turk*	*Concept Feedback*	*Crazyegg*	*Feng-GUI*	*Feedback Army*	*Attention Wizzard*	*ClickHeat*
OS compatibility	**3**	**4**	**10**	**10**	**10**	**10**	**3**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**
- windows	**4**	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
- mac os	**3**	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
- linux	**4**	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Types of interfaces supported	**4**	**5**	**5**	**5**	**5**	**8**	**5**	**5**	**8**	**5**	**5**	**5**	**9**	**5**	**5**	**8**	**5**	**5**	**9**	**5**	**1**	**5**	**5**	**5**
- applications	**4**	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1	0	1	1	1
- prototypes	**5**	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1
- screenshots of website	**2**	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0
- wireframes	**2**	0	0	0	0	0	0	0	1	0	0	0	1	0	1	1	0	0	1	0	0	0	0	0
- mockups	**4**	0	0	0	0	1	0	0	1	0	0	0	1	0	1	1	0	0	1	0	0	0	0	0
Interface integration	**4**	**3**	**2**	**2**	**3**	**3**	**3**	**3**	**1**	**3**	**3**	**2**	**3**	**3**	**1**	**1**	**3**	**2**	**1**	**3**	**1**	**1**	**1**	**3**
- offline program	**4**	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
- online post /URL submission	**2**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0
- upload images	**1**	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0
- Javascript code	**4**	0	0	0	1	1	0	1	0	1	1	0	1	1	0	0	1	0	0	1	0	0	0	1
- online wizard	**3**	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
Usability test access	**5**	**1**	**9**	**5**	**5**	**9**	**1**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**
- local	**1**	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
- remote	**4**	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
- online	**5**	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Surveys	**4**	**10**	**10**	**6**	**6**	**6**	**6**	**0**	**3**	**6**	**0**	**0**	**0**	**0**	**6**	**3**	**7**	**6**	**0**	**0**	**0**	**3**	**0**	**0**
- complete survey	**5**	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
- screen aligned questions	**4**	1	1	1	1	1	1	0	1	1	0	0	0	0	1	1	0	1	0	0	0	1	0	0
- screen aligned text	**3**	1	1	1	1	1	1	0	0	1	0	0	0	0	1	0	1	1	0	0	0	0	0	0
Collecting audio	**4**	**10**	**0**	**10**	**6**	**0**	**10**	**0**	**0**	**0**	**0**	**6**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
- audio record	**5**	1	0	1	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
- annotations	**4**	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Collecting video	**5**	**10**	**3**	**5**	**3**	**0**	**8**	**3**	**5**	**0**	**3**	**3**	**0**	**3**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
- display	**5**	1	1	1	1	0	1	1	1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0
- facial expressions recording	**4**	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
- eye tracking	**4**	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
- annotations	**4**	1	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Usability map types supported	**4**	**6**	**5**	**0**	**5**	**2**	**0**	**5**	**2**	**0**	**7**	**0**	**0**	**6**	**2**	**2**	**0**	**0**	**0**	**5**	**5**	**0**	**1**	**2**
- clicks	**5**	1	1	0	1	1	0	1	1	0	1	0	0	1	1	1	0	0	0	1	0	0	0	1
- mouse move	**5**	1	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0
- scroll reach	**4**	1	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0
- attractive zones	**4**	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
- interest zones	**4**	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
- attention zones	**4**	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0
- form inputs	**5**	1	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0
Export results	**5**	**6**	**6**	**2**	**2**	**3**	**0**	**5**	**2**	**5**	**2**	**2**	**5**	**0**	**2**	**2**	**0**	**0**	**2**	**0**	**3**	**0**	**2**	**0**
- XLS / CSV / TSV	**5**	1	1	0	0	1	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
- XML	**4**	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
- database	**3**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
- online mng. results access	**3**	0	0	1	1	0	0	1	1	1	1	1	1	0	1	1	0	0	1	0	0	0	1	0
UM		**8**	**7**	**6**	**6**	**5**	**5**	**5**	**5**	**5**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**3**	**3**	**3**	**3**	**3**	**3**

Table 3: Main issues and features preferences in a 1-5 scale. Relative and final scores in a 0-10 scale.

5.2 Analytical hierarchy process

Table 4 presents the results obtained from AHP study. All numbers are displayed as percentages. The first column ranks the UTTs according to AHP results. Each of the columns 1-9 displays a criterion (main issue), its weight (second row) and the importance of each UTT in this criterion. Column "AHP" displays the relative importance of the UTT obtained by AHP, while "AHP (%)" displays the corresponding normalization considering 100% for the best alternative scores. Their counterparts "UM" and "UM (%)" display the same numbers obtained by the UM.

The best scored UTT, *Loop 11*, presents high preferences for criteria considered crucial (4, 7 and 9). In addition, it reached satisfactory scores for the other criteria. It does not offer the features of criterion 6, however, this will not exclude it from our choice. The second UTT, Morae, provides limited user access, which may restraint remote usability tests. However, this UTT presents good scores in almost all the other criteria and, consequently, was not excluded. Considering that this evaluation was mainly supported by industrial advertising information, additional information is needed.

The next four UTT, *Usabilla*, *Click Density*, *Userfly* and *User Testing* present good scores, offering all the required functionalities, even in a limited way, with the exception of *Usabilla* that does not provide audio and video recording.

Sensitivity analysis allowed us to conclude that the "User Access" weight strongly influences the relative importance of *Morae*.

AHP produced results finer tuned than the previously obtained by the UM, highlighting the relative differences between UTTs. This is also disclosed by the standard deviation values displayed in the last line. The pairwise comparison of criteria is also more comprehensive than the normative assignment of marks, either in a quantitative or qualitative scale. Though small differences were found in the relative positions, the most significant difference concerns the first two UTTs, which can be explained by the tuned comparison of criteria preferences. These results should be interpreted carefully. Besides the limited type of sampling, most of the features were reduced to binary evaluation.

Scalability, for instance in the number of surveys or usability tests, seems often just a question of pricing. However, some of the features, even when present, may have some limitations when compared to a similar implementation in another UTT. Ultimately, some very specific features which can be highly valuable are only provided by few UTT. It should also be noted that all the preferences were defined by a small number of experts and considering the requirements of a specific project (World Search Project). Pricing can obviously be an important restriction for any product, which in this case was decided to be considered separately. It is still interesting to find some correlation between the price and the number of features or their specificity. Again, scalability can produce very significant pricing differences.

Criteria	1. OS compatibility	2. Supported types of prototypes	3. Interface integration w/ the UTT	4. User access	5. Creation & submission of surveys	6. Collecting audio	7. Collecting video	8. Usability maps	9. Export	AHP	UM	AHP (%)	UM (%)
UTT /Weights	3	7	7	27	7	7	18	7	18				
Loop 11	5	4	3	8	13	0	8	9	15	8	7	100	90
Morae	2	4	7	1	13	24	16	12	15	7	8	93	100
Usabilla	5	6	7	8	8	0	0	4	11	6	5	77	70
Click density	5	4	7	4	0	0	8	7	12	6	5	71	62
Userfly	5	4	7	4	8	14	8	11	2	6	6	70	74
User Testing	5	4	3	4	8	24	11	0	2	6	6	70	78
Mouse Flow	5	4	7	4	0	0	8	13	2	5	4	60	58
Intuition HQ	5	5	1	4	4	0	11	4	2	5	5	59	61
4Q Survey	5	4	7	4	8	0	0	0	12	5	5	59	59
ClickTale	5	4	7	4	0	0	8	12	0	4	4	57	51
Google Analytics	5	6	7	4	0	0	0	0	12	4	4	56	53
Open Hallway	5	4	3	4	0	14	8	0	2	4	4	55	54
Silverback	1	4	7	1	8	24	14	0	0	4	5	55	63
Fivesecondtest	5	6	1	4	4	0	0	4	2	3	4	44	49
Ethnio	5	4	7	4	9	0	0	0	0	3	4	42	49
Crazyegg	5	4	7	4	0	0	0	9	0	3	3	41	42
Mechanical Turk	5	4	3	4	8	0	0	0	0	3	4	40	46
Chalkmark	5	2	1	4	8	0	0	4	2	3	4	39	49
ClickHeat	5	4	7	4	0	0	0	4	0	3	3	38	38
Concept Feedback	5	6	1	4	0	0	0	0	2	3	3	38	43
Feng-GUI	5	0	1	4	0	0	0	7	5	3	3	37	40
Feedback Army	5	4	1	4	4	0	0	0	0	3	3	36	40
Attention Wizzard	5	4	1	4	0	0	0	1	2	3	3	36	39
							Std deviation			1,4	1,3	18	16

Table 4: AHP results – compared with previous UM results.

6 Conclusions and future work

Our team main concern in the World Search Project (World Search Project, 2010) is to enforce a user-centered design approach in a set of advanced information search demonstrators for specific domains. This kind of approach can benefit from using integrated usability testing tools (UTTs) for new applications design and development. Experience teams working regularly with a suitable UTT can better concentrate on solving usability issues and proposing innovative products. New team members can also find a good reference for integration by sharing such UTT capabilities with more experienced member teams. To the best of our knowledge, our study is the first quantitative evaluation and comparison of a significant number of UTTs within the context of Web graphical interfaces design. A special effort was given to include in our list all UTTs adequate to this context. A simple linear utility function and AHP model was used to score and rank 23 UTTs. Weighting and scoring was performed by a small team of experts.

The presented results should be considered with caution, due to the limited type of evaluation, namely

almost exclusively based on the vendor's descriptions. Future work is expected in three different directions. The first direction will investigate and test other suitable multiple criteria decision analysis methods (Cechich et al., 2003; Figueira et al., 2004). A second direction will increase the number of experts for getting more reliable preferences and perhaps including new features. A third direction will verify features in lab for the preferred set of candidates. There will be an extra concern on usability tests/ UTTs features for applications running in mobile devices.

7 Acknowledgement

This work was partially supported by project QREN – I&D / ADI Nº 11495 (World Search Project, 2010) and by National Funding from FCT - Fundação para a Ciência e a Tecnologia, under the project: PEst-OE/MAT/UI0152.

References

[1] Cechich A., Piattini M., and Vallencillo A., 2003. *Component-Based Software Quality*, Springer Verlag, Berlin, Heidelberg.

[2] Dix A., Finlay J., Abowd G., Beale R., 2003. *Human Computer Interaction*, Prentice-Hall, Upper Saddle River, NJ, USA.

[3] Expert Choice, 2012. [online]. Available: http://expertchoice.com [10 October 2012].

[4] Fadeyev, D.,2009. *10 Tools to Improve Your Site's Usability on a Low Budget*, [Online], Available: http://www.webdesignerdepot.com/2009/06/10-tools-to-improve-your-site%E2%80%99s-usability-on-a-low-budget/ [23 July 2012].

[5] Figueira J., Greco S., Ehrgott M., 2005. *Multiple Criteria Decision Analysis*, Springer Verlag, Boston Dordrecht, London.

[6] Gube, J., 2011. *22 Essential Tools for Testing Your Website's Usability*, [Online], Available: http://mashable.com/2011/09/30/website-usability-tools/ [24 July 2012].

[7] Jules, 2011. *Best Website Usability Testing Tools and Services*, [Online], Available: http://www.quertime.com/article/arn-2011-04-06-1-best-website-usability-testing-tools-and-services/ [24 July 2012].

[8] LeMerle, R., 2012. *10 Essential Website Usability Tools*, [Online], Available: http://blog.ineedhits.com/tips-advice/10-essential-website-usability-tools-045711224.html [24 July 2012].

[9] Nielsen J., 1993. *Usability Engineering*, Academic Press, Boston.

[10] Nielsen J., 1999. *Designing Web Usability*, New Riders Publishing, Thousand Oaks, CA, USA.

[11] Norman D., 2002. *The Design of Everyday Things*, Basic Books, New York.

[12] Preece J., Rogers Y., Sharp H., Benyon D., Holland S., Carey T., 1994. *Human Computer Interaction*, Addison Wesley. Reading, Massachusetts.

[13] Saaty, T.L., 2005. The Analytic Hierarchy and Analytic Network Processes for the Measurement of Intangible Criteria and for Decision-Making. In J. Figueira, S. Greco, and M. Ehrgott (eds.) MCDA: State of the Art Surveys, Springer Verlag.

[14] Shneiderman B., 1998. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, Addison Wesley, Reading, Massachusetts.

[15] Tomlin, W., 2009. *24 Usability Testing Tools*, [Online], Available: http://www.usefulusability.com/24-usability-testing-tools/ [23 July 2012].

[16] Vraa, L., 2009. *16 Crucial Webdesign and Usability Best Practice Compilations and Tools*, [Online], Available: http://www.tripwiremagazine.com/2009/06/16-crucial-webdesign-and-usability-best-practice-compilations-and-tools.html [23 July 2012].

[17] Walker, T., 2010. *20 Fantastic Usability & Conversion Analysis Tools*, [Online], Available: http://spyrestudios.com/usability-conversion-analysis-tools/ [24 July 2012].

[18] World Search Project, 2010. [Online], Available: http://www.microsoft.com/portugal/mldc/worldsearch/en/ [March 2011].

Frequent Spatiotemporal Association Patterns Mining Based on Granular Computing

Gang Fang[1, 2] and Yue Wu[1]
[1] School of Computer Science & Engineering
University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, P. R. China
[2] School of Computer Science & Engineering
Chongqing Three Gorges University, Wanzhou, Chongqing, 404000, P. R. China
E-mail: gangfang07@sohu.com, ywu@uestc.edu.cn

Keywords: spatiotemporal association patterns, star association model, granular computing, mixed radix notation system, spatiotemporal data mining

Received: May 17, 2013

In order to discover multi-dimensional spatiotemporal association patterns, and improve the efficiency of traditional mining algorithms for spatiotemporal association patterns, this paper firstly constructs a star association model based on event, which can show more spatiotemporal information on the basis of the present star association model. Besides traditional attributes association with point, line and plane, the model also can fully and flexibly express temporal association, orientation association, and topology association, namely, it can quickly and simply form multi-dimensional spatiotemporal association patterns. And then for the star association model based on event, an algorithm of discovering frequent spatiotemporal association patterns based on granular computing is proposed, which is different from traditional association patterns mining algorithms. One is that the algorithm breaks traditional thinking of generating candidate frequent itemsets, namely, it generates candidate frequent itemsets by updating the mixed radix numeral. The method is quick and simple to avoid redundant complicated calculations for adopting complex FP-tree data structure or generating candidate by joining frequent itemsets. The other is that the algorithm for discovering frequent spatiotemporal association patterns only needs to read database once via granular computing, in other words, it discovers each frequent spatiotemporal association pattern via constructing a spatiotemporal information granule, where the intension can be mapped to the mixed radix numeral from the mixed radix notation system based on spatiotemporal information system. Finally, this paper further discusses the characteristics and the optimal application environments of the algorithm. Experimental results indicate that the algorithm is simpler and faster than these traditional frequent patterns mining algorithms on the optimal application environments.

Povzetek: Opisana je nova metoda za rudarjenje prostorsko-časovnih vzorcev.

1 Introduction

Discovering spatial association patterns from spatial database is one of important tasks for spatial data mining and knowledge and discovery. Spatial association patterns have been applied to some valuable domains, such as Urban Traffic [1], Bioscience [2], Social Security [3], Climate forecasting [4], and Demographic survey [5]. In recent decade, there is some research work for mining spatial association patterns. Reference [6] focuses on this specificity of spatial data mining by showing the suitability of join indices to this context. It describes the join index structure and shows how it could be used as a tool for spatial data mining; Reference [7] discusses the multiple level association rules mining, and further indicates spatiotemporal association rules mining should address issues of data integration, data classification, the representation and calculation of spatial relationships, and strategies for finding 'interesting' rules; Reference [8] has proposed a generalized framework to effectively discover different types of spatial and spatiotemporal patterns in scientific data sets, which can be used to capture a variety of interactions among objects of interest and the evolutionary behaviour of such interactions. Based on the feature of geographic elements, the research work can be divided into the following two groups.

One group is discovering frequent region association patterns for numeric geographic elements with point, line and plane, i.e. firstly, these numeric attributes are turned into Boolean attributes with geographic elements, and spatial association patterns are discovered by transaction frequent patterns mining methods. The group is suitable for mining spatial association patterns based on spatial location. Reference [7] uses association rules to discover spatiotemporal relationships among a set of variables that characterize socioeconomic and land cover changes, but it only refers to the region. Reference [9] proposes a robust geospatial multivariate association rules mining framework, where the attributes for geographic elements with point, line and plane can be turned into the region. Reference [10] proposes a novel framework to mine

regional association rules based on a given class structure.

The other is discovering frequent spatial association patterns for discrete geographic elements with spatial objects and layout relationships, i.e. firstly, these discrete geographic elements are turned into the category set, and transaction frequent patterns mining methods are used to extract spatial association patterns. Reference [8] and [11] discuss star association patterns, sequence association patterns and clique association patterns based on spatial distance for these spatial objects relationships and layout relationships.

However, these research have the following some shortcoming, firstly, these mining objectives are mainly from spatial database, where these algorithms do not fully regard temporal relationship with spatial association patterns; Secondly, their geographic elements in spatial association patterns are most one-dimensional, namely, the form of spatial association patterns is quite single. Finally, for these traditional frequent patterns mining algorithms, such as Apriori, FP-growth, and their improved algorithms have some disadvantages as follows:

One is the mining framework based on Apriori, i.e. these mining algorithms discover frequent patterns via the thinking of the algorithm Apriori, called the Apriori Framework. The mining framework needs to repeatedly read database for discovering frequent itemsets.

The other is the mining framework based on FP-growth, i.e. these mining algorithms discover frequent patterns via data structure FP-tree, called the FP-growth Framework. The mining framework uses complex data structure to save reading database, but it needs to cost much memory for discovering frequent patterns.

These mining frameworks have some disadvantages for more details seeing references [12-16].

The main contributions in our research work can be summarized as follows:

One is constructing a star association model based on event, which not only expresses traditional attributes association with point, line and plane; the model also can fully flexibly express multi-dimensional spatiotemporal association patterns including the orientation association, the temporal association and the topology association.

The other is proposing an algorithm of discovering frequent spatiotemporal association patterns based on granular computing. For discovering frequent spatio-temporal association patterns, it only needs to read the database once; and then it generates candidate frequent itemsets via updating the mixed radix numeral, where granular computing is introduced to save reading the spatiotemporal database.

The remainder parts are organized as follows:

In Section 2, we introduce the related research work; In Section 3, we construct a star association model based on event; In Section 4, we propose an algorithm of discovering frequent spatiotemporal association patterns based on granular computing; In Section 5, we use some experiments to verify the algorithm, and then discuss its the optimal application environments. In Section 6, we summary research results and discuss future work.

2 Related research work

Based on the notions of granularity [17] and abstraction [18], the ideas of granular computing have been widely investigated in artificial intelligence [19]. In this paper, we adopt a partition model of granular computing to construct information granule [19], which depends on rough set theory [20] and quotient space theory [21]. Here, we introduce the following related definitions.

Definition 2.1 An information table is a quintuple $S = (U, A, \{V_a \mid a \in A\}, L, \{I_a \mid a \in A\})$, where

U, called universe of discourse, is a finite nonempty set for objects;

A, called attributes set, is also a finite nonempty set for attributes;

V_a, called domain set, is a finite set of values for $a \in A$, where V_a is defined as a discrete category set;

L, called descriptive language, a language is defined by attributes in A;

For describing an object of U via the language, it can be denoted as $L = \{\ell \mid V_{a_*^1} \times V_{a_*^2} \times ... \times V_{a_*^{|A^*|}}, a_* \in A^* \subseteq A\}$;

I_a, called information function, is a total function that maps an object of U to exactly one value in V_a, namely $I_a: U \to V_a$.

Definition 2.2 Information granule is a two-tuple $IG = (\xi, \varphi(\xi))$, where

ξ, called the intension of information granule, consists of all attributes that are valid for all those objects to which information granule applies; in other words, the intension is an abstract description of common features or properties shared by elements in the extension, which is expressed as $\xi = (\xi_1, \xi_2, ..., \xi_{|\xi|})$, where

$\xi_k \in V_{a_*^k}$, $a_* \in A^* \subseteq A$, $k = 1, 2, ..., |\xi|$, $\xi \in L$;

$\varphi(\xi)$, called the extension of information granule, is the set of objects which information granule applies, in other words, the extension consists of concrete examples of information granule, which is expressed as follows:
$\varphi(\xi) = \{x \in U \mid I_{a_*^1}(x) = \xi_1, I_{a_*^2}(x) = \xi_2, ..., I_{a_*^{|\xi|}}(x) = \xi_{|\xi|}\}$;

Definition 2.3 Atomic information granule is a two-tuple $AIG = (\xi, \varphi(\xi))$, where

ξ, called the intension of $AIG = (\xi, \varphi(\xi))$, is expressed as $\xi = (\xi_a)$ $(\xi_a \in V_a, a \in A, \xi \in L)$;

$\varphi(\xi)$, called the extension of $AIG = (\xi, \varphi(\xi))$, is expressed as $\varphi(\xi) = \{x \in U \mid I_a(x) = \xi_a\}$.

Definition 2.4 Intersection operation of information granule is denoted by $\otimes$, which is described as follows:

Let two information granules be $IG_\alpha = (\xi_\alpha, \varphi(\xi_\alpha))$ and $IG_\beta = (\xi_\beta, \varphi(\xi_\beta))$, respectively; if $(\exists\, \xi_\alpha^i \in \xi_\alpha \wedge \xi_\alpha^i \in V_a) \wedge (\exists\, \xi_\beta^j \in \xi_\beta \wedge \xi_\beta^j \in V_a)$ then $\xi_\alpha^i = \xi_\beta^j$; and then the intersection operation $\otimes$ can be expressed as follows:
$IG = (\xi, \varphi(\xi)) = IG_\alpha \otimes IG_\beta = (\xi_\alpha \cup \xi_\beta, \varphi(\xi_\alpha) \cap \varphi(\xi_\beta))$.

Definition 2.5 Star association model is expressed as $M =< e_c, \{e_1, e_2,..., e_m\}, \prec >$, where

e_c, called the core element of star association model, is a sole core element;

$e_i (i = 1, 2,..., m)$, called the non-core element of star association model, at least there is a kind of association between each non-core element and the core element;

$\prec$, called time series relationship of star association model, this model only has two types of time series, namely, $\{e_c \prec e_1 \wedge e_c \prec e_2 \wedge ... \wedge e_c \prec e_m\}$ or $\{e_1 \prec e_c \wedge e_2 \prec e_c \wedge ... \wedge e_m \prec e_c\}$.

Definition 2.6 Star association pattern is denoted by $P = \{r_1, r_2,..., r_k\}$, where the association between the core element e_c and the non-core element $e_i (i \in [1, 2,..., k])$ can be denoted by $r_i = R < e_c, e_i > (i \in [1, 2,..., k])$, which consists of the temporal association, the orientation association and the topology association. And then, star association patterns mining is defined as discovering frequent star association patterns from spatiotemporal database for the given minimal support.

3 A star association model based on event

In this paper, on the basis of definition 2.5, we propose a star association model based on event. The model is applied to transform spatiotemporal events and discover frequent spatiotemporal association patterns in Section 4.

Definition 3.1 Star association model based on event is denoted as $EM =< e, e_c, A, E_s, F, P >$, where

e, called a spatiotemporal event, is from a spatiotemporal database, which consists of orientation factor, time factor and topology factor, besides these traditional attributes with point, line and plane;

e_c, called the core element of star association model based on event, is a subject object in the event;

A, called attributes set of the core element e_c, is also a finite nonempty set for the attribute, which can be traditional attribute with point, line and plane;

E_s, called non-core elements set of star association model based on event, is a set of spatial entity objects denoted by $E_s = \{e_1, e_2,..., e_m\}$. Here is only a kind of spatial location association between the core element e_c and each non-core element e_i $(e_i \in E_s)$;

F, called spatiotemporal factors set for describing this event e, is expressed as follows:

$F = \{time, orientation, topology\}$;

P, called predicates set for F, is a finite set of values for $f \in F$, denoted by $P = \{P_{time}, P_{orientation}, P_{topology}\}$.

In this paper, P_f $(f \in F)$ is defined as follows:

$P_{time}(e_c, e_t) = \{before(e_t), after(e_t), equal(e_t)\}$, where e_t is a temporal element;

$P_{orientation}(e_c, e_o) = \{east(e_c), south(e_c), southeast(e_c), west(e_c), southwest(e_c), northwest(e_c), northeast(e_c), north(e_c)\}$, where e_o is an orientation element;

$P_{topology}(e_c, e_s) = \{disjoint(e_c, e_s), coveredby(e_c, e_s), cover(e_c, e_s), contain(e_c, e_s), touch(e_c, e_s), inside(e_c, e_s), overlap(e_c, e_s)\}$, where e_s is a spatial entity objects;

For example, there are three spatiotemporal events from a spatiotemporal database, and then we use the star association model based on event to describe them as follows:

(1) e: a taxi (No.t2) with passenger eastward fast left the school (No.s1) along the riverside (No.r1) at 5 PM.

e_c: $taxi(2)$, is a subject object in this event;

A_1: $\{load, rate\} = \{true, fast\}$;

E_s: $\{school(1), river(1)\}$;

$time$: $\{at\ 5\ PM\}$;

$orientation$: $\{east\}$;

$topology$: $\{a\ taxi\ touch\ the\ river, a\ taxi\ disjoint\ the\ school\}$.

We can use traditional attributes and these predicates to describe the star association pattern for the event, which can be expressed as follows:

$P(1) = \{load = true, rate = fast, before(night), equal(afternoon), east(taxi(2)), disjoint(taxi(2), school(1)), touch(tax(2), river(1))\}$;

(2) e: a taxi (No.t5) with passenger southward slow droved into the school (No.s2), and parked at the gate of the bank (No.b3) at 11 AM.

e_c: $taxi(5)$, is a subject object in this event;

A_1: $\{load, rate\} = \{true, slow\}$;

E_s: $\{school(2), bank(3)\}$;

$time$: $\{at\ 11\ AM\}$;

$orientation$: $\{south\}$;

$topology$: $\{a\ taxi\ is\ inside\ the\ school, a\ taxi\ touch\ the\ bank\}$;

Via traditional attributes and these predicates, the star association pattern for the event can be expressed as follows:

$P(2) = \{load = true, rate = slow, before(afternoon), equal(morning), south(taxi(5)), touch(taxi(5), bank), inside(taxi(5), school(1))\}$;

(3) e: a taxi (No.t6) without passenger southwest slow left the bank (No.b4), and droved into the business street (No.b3) at 8 night.

e_c: $taxi(6)$, is a subject object in this event;

A_1: $\{load, rate\} = \{false, slow\}$;

E_s: $\{business\ street(3), bank(4)\}$;

$time$: $\{at\ 8\ night\}$;

$orientation$: $\{southwest\}$;

$topology$: {*a taxi is covered by the business street*, *a taxi disjoint the bank*} .

Via traditional attributes and these predicates, the star association pattern for the event can be expressed as follows:

$P(3) = \{load = true, rate = slow, after(afternoon), equal(night), south(taxi(6)), touch(taxi(6), bank(4)), inside(taxi(6), business\ street)\}$;

Definition 3.2 Spatiotemporal association patterns mining is defined as discovering frequent spatiotemporal association patterns from spatiotemporal database based on event, namely, frequent star association patterns, whose support is the same as traditional association rules.

In the course of mining spatiotemporal association patterns, there are two key problems as follows:

One is turning an event into a spatiotemporal association patterns, namely, star association patterns. We have solved the problem via definition 3.1;

The other is discovering frequent spatiotemporal association patterns. We use the algorithm as described in Section 4.2 to solve the problem.

4 Frequent spatiotemporal association patterns mining

In this section, firstly, we introduce granular computing based on the star association model, and then propose an algorithm of discovering frequent spatiotemporal association patterns based on granular computing; finally, we compare the algorithm with these traditional mining algorithms, particularly, the Apriori Framework and the FP-growth Framework

4.1 Granular computing based on the star association model

Definition 4.1 A spatiotemporal information system based on the star association model is a six-tuple $STIS = (U, F, A, \{V_a \mid a \in A\}, L, \{I_a \mid a \in A\})$, where

U , called universe of discourse, is a finite nonempty set of events, where each event has a sole core element;

F , called spatiotemporal factor set for describing the event e in U , is expressed as follows:

$F = \{time, orientation, topology\}$;

A , called joined attributes set of an event, is denoted by $A = A_t \cup E_t \cup \{orientation\} \cup E_s$,

Where

A_t , called traditional attributes with point, line and plane;

E_t , is a given group of time division; such as $E_t = \{morning, afternoon, night\}$ or $E_t = \{Monday, Tuesday, Wensday, Thursday, Friday, Saturday, Sunday\}$;

E_s , called non-core generalization set for U , is a finite nonempty set of non-core element category for the star association model based on event;

For example, there are three spatiotemporal events in Section 3; and we can get these spatial entities for them as follows:

$\{school(1), river(1), school(2), bank(3), business\ street(3), bank(4)\}$;

And then, we have the following E_s :

$E_s = \{school, river, bank, street\}$;

There are more details about A in table 1.

V_a , called domain set, is a nonempty finite set of values for attribute a ($a \in A$) , which can be expressed as follows:

$$V_a = \begin{cases} V*_a, & a \in A_t \\ P_{time}(e_c, a), & a \in E_t \\ P_{orientation}(e_c, a), & a = orientation \\ P_{topology}(e_c, a), & a \in E_s \end{cases}$$, where

$V*_a$, called domain of traditional attribute with point, line and plane, is defined as a discrete category set; the others are the predicate sets as described in definition 3.1.

L , called a kind of logical descriptive language, is defined to describe a spatiotemporal event through these traditional attributes and predicates; L can be expressed as $L = \{\ell \mid V_{a_*^1} \times V_{a_*^2} \times ... \times V_{a_*^{|A^*|}}, a_* \in A^* \subseteq A\}$;

I_a , called information function, is a total function that maps an event of U to exactly one value in V_a , namely $I_a: U \to V_a$.

Based on definitions 3.1 and 4.1, for the three spatiotemporal events in Section 3, and we let a time division be $E_t = \{morning, afternoon, night\}$, then we can construct a spatiotemporal information system based on the star association model as follows:

$STIS = (U, F, A, \{V_a \mid a \in A\}, L, \{I_a \mid a \in A\})$, where

$A_t = \{load, rate\}$, is an attribute set of the taxi (called a point entity);

$E_s = \{school, river, bank, street\}$;

So we can create the following mining database as described in table 1.

A / *T_ID*	Taxi(2)	Taxi(5)	Taxi(6)
Load	True	True	False
Rate	Fast	Slow	Slow
Morning	--	Equal	--
Afternoon	Equal	Before	After
Night	Before	--	Equal
Orientation	East	South	Southwest
School	Disjoint	Inside	--
River	Touch	--	--
Bank	--	Touch	Touch
Street	--	--	Inside

Table 1: Mining database.

Definition 4.2 Spatiotemporal information granule is a two-tuple $STIG = (\zeta, \psi(\zeta))$, where

ζ, called the intension of spatiotemporal information granule, is an abstract description of common values of joined attributes shared by events in the extension, which is expressed as $\zeta = (\zeta_1, \zeta_2, ..., \zeta_{|\zeta|})$ $(\zeta_k \in V_{a_*^k}, a_* \in A^* \subseteq A, k = 1, 2, ..., |\zeta|, \zeta \in L)$;

$\psi(\zeta)$, called the extension of spatiotemporal information granule, is a set of events which spatiotemporal information granule applies, which is expressed as follows:

$\psi(\zeta) = \{u \in U \mid I_{a_*^1}(u) = \zeta_1, I_{a_*^2}(u) = \zeta_2, ..., I_{a_*^{|\zeta|}}(u) = \zeta_{|\zeta|}\}$.

Definition 4.3 Atomic spatiotemporal information granule is a two-tuple $ASTIG = (\zeta, \psi(\zeta))$, where

ζ, called the intension of atomic spatiotemporal information granule, is denoted by $\zeta = (\zeta_a)$ $(\zeta_a \in V_a, a \in A, \zeta \in L)$;

$\psi(\zeta)$, called the extension of atomic spatiotemporal information granule, is denoted by the following:

$\psi(\zeta) = \{u \in U \mid I_a(u) = \zeta_a\}$.

Definition 4.4 Intersection operation of spatio- temporal information granule is denoted by Θ. Suppose two spatiotemporal information granules are $STIG_\alpha = (\zeta_\alpha, \psi(\zeta_\alpha))$ and $STIG_\beta = (\zeta_\beta, \psi(\zeta_\beta))$, respectively; if $(\exists \zeta_\alpha^i \in \zeta_\alpha \wedge \zeta_\alpha^i \in V_a) \wedge (\exists \zeta_\beta^j \in \zeta_\beta \wedge \zeta_\beta^j \in V_a)$ then $\zeta_\alpha^i = \zeta_\beta^j$; and so the intersection operation Θ can be expressed as $STIG = (\zeta, \psi(\zeta)) = STIG_\alpha \Theta STIG_\beta$

$= (\zeta_\alpha \cup \zeta_\beta, \psi(\zeta_\alpha) \cap \psi(\zeta_\beta))$.

Definition 4.5 A mixed radix notation system based on a spatiotemporal information system is a triple $M = \{STIS, m, < w_1, w_2, ..., w_m >\}$, where

$STIS$, called a spatiotemporal information system, is expressed as follows:

$STIS = (U, F, A, \{V_a \mid a \in A\}, L, \{I_a \mid a \in A\})$;

m, called the number of bit for the mixed radix notation system, is denoted by $m = |A|$;

$< w_1, w_2, ..., w_m >$, called the weight set of bit for the mixed radix notation system, each element w_i is defined as $w_i = |V_{a_i}| + 1$ $(i = 1, 2, ..., m)$, and $V_{a_i}^k \leftrightarrow k$ $(k = 1, 2, ..., |V_{a_i}|)$, and $|M| = \prod_{i=1}^{m} w_i - 1$.

For example, for the spatiotemporal information system of table 1, we can get the following mixed radix notation system based on a spatiotemporal information system $M = \{STIS, 10, < 3, 3, 4, 4, 4, 9, 8, 8, 8, 8 >\}$.

Definition 4.6 Combinatorial number ratio based on a spatiotemporal information system is defined as $\rho = log_{|U|}^{|M|} > 0$, where

$|U|$, is the number of spatiotemporal events in the spatiotemporal database, which is mapped to the spatiotemporal information system $STIS$;

$|M|$, is a combinatorial number for attribute values in the spatiotemporal database, which is mapped to the spatiotemporal information system $STIS$.

4.2 Discovering frequent spatiotemporal association patterns

In this section, we propose an algorithm of discovering frequent spatiotemporal association patterns based on granular computing, which is denoted by DFSTAP, and then we use the following pseudo code to describe the algorithm DFSTAP.

STD, is a spatiotemporal database based on event;

s, is the given minimal support;

F: saving these maximal frequent spatiotemporal association patterns;

NF: saving these non frequent spatiotemporal association patterns;

Input: STD and s;

Output: F;

(1) $F = \Phi$;
(2) $NF = \Phi$;
(3) Read STD; //reading once database
(4) Create $STIS_{STD} = (U, F, A, \{V_a \mid a \in A\}, L, \{I_a \mid a \in A\})$; //def. 4.1, creating a $STIS$ by the STD
(5) Compute each $ASTIG = (\zeta, \psi(\zeta))$; // def. 4.3
(6) Create $M = \{STIS, m, < w_1, w_2, ..., w_m >\}$; // def. 4.5
(7) For $\forall i \in [1, |M|]$ do {
(8) $M(i) = (\omega_m \omega_{m-1} ... \omega_1)_M$; //a decimal integer i is turned into a mixed radix numeral $(\omega_m \omega_{m-1} ... \omega_1)_M$
(9) $\zeta_{M(i)} = \zeta_m \cup \zeta_{m-1} \cup ... \cup \zeta_1$; // $\zeta_{M(i)}$ is a set of items, each ζ_k is mapped to the ω_k
(10) If $(\forall \varpi \in NF, \varpi \not\subset \zeta)$ then {
(11) Construct $STIG = (\zeta_{M(i)}, \psi(\zeta_{M(i)}))$; // def. 4.4
(12) If $|\psi(\zeta_{M(i)})| \geq s$ then {
(13) Delete σ $(\forall \sigma \in F, \sigma \subset \zeta_{M(i)})$; //deleting all subsets of $\zeta_{M(i)}$ in F
(14) Write $\zeta_{M(i)}$ to F;} //saving frequent itemset
(15) Else
(16) Write $\zeta_{M(i)}$ to NF; //saving non frequent itemset
(17) }
(18) $i++$;}
(19) Output F;

The interval $[1, |M|]$ in the algorithm is the search range of candidate frequent patterns. In other word, the algorithm updates the mixed radix numeral to generate candidate frequent itemsets.

The algorithm discovers frequent spatiotemporal association patterns through constructing spatiotemporal information granule.

4.3 Performance comparison

Based on the introduction in Section 4.2, we know the algorithm DFSTAP is different from traditional frequent patterns mining algorithms, particularly, the Apriori Framework and the FP-growth Framework.

For discovering frequent association patterns, the Apriori Framework is a representative algorithm with candidate, and the FP-growth Framework is a typical algorithm without candidate, and then we compare the algorithm DFSTAP with the Apriori Framework and the FP-growth Framework. The comparative results can be expressed as the following table 2.

Based on the comparison as described in table 2, we can draw the following conclusions:

The Apriori Framework needs to read the database repeatedly, and it joins two frequent itemsets to generate candidate; and so there are lots of calculated amount for discovering frequent patterns. However, the algorithm DFSTAP updates the mixed radix numeral to generate candidates; the speed of which for the latter is faster than the former; additionally, the DFSTAP only needs to read the database once. Hence, the computational complexity of the algorithm DFSTAP is lower than the Apriori Framework. In other words, the algorithm avoids these disadvantages of the Apriori Framework.

In addition, the algorithm DFSTAP uses simple data structure as array to express single format of candidate, and traverses an interval to discover frequent association patterns; so it uses less memory; and it is easy to program and maintain the algorithm. Namely, the DFSTAP has these advantages of the Apriori Framework.

However, for mining frequent association patterns, the FP-growth Framework only needs to read database twice, its advantage is saving reading database, the DFSTAP also has the advantage. But the FP-growth Framework needs to traverse a complex FP-tree; so its computational complexity is higher than the DFSTAP, meanwhile, it also needs to cost more memory, and it is no picnic to program and maintain it. Obviously, the algorithm DFSTAP avoids these disadvantages of the FP-growth Framework.

Comparative items	DFSTAP	Apriori Framework	FP-growth Framework
Reading Database	Once	Many times	Twice
Data structure	Simple	Simple	Complex
Programming	Simple	Simple	Complex
Computational complexity	Low	High	High
Memory usage	Less	Less	More
Generating candidate	Yes	Yes	No
Format of candidate	Digit	Itemset	--
Speed of generating candidate	Fast	Slow	--

Table 2: Performance comparison.

In conclusion, this algorithm is better than traditional mining algorithm in theory.

5 Experimental result

In this section, we design two types of experiments as follows:

One is evaluating the performances of the proposed mining algorithm for discovering frequent spatiotemporal association patterns on different datasets.

The other is discussing the application environments for the proposed mining algorithm.

The first data set is from the GPS data of taxi in a city, for the GPS interval point with a taxi, an event is made of speed, loading, time, and space layout. There are 323080 spatiotemporal events after data filtering; the dataset can be mapped to a spatiotemporal information system $STIS_1$, and then we can create a mixed radix notation system based on the spatiotemporal information system $STIS_1$, which can be expressed as follows:

$M_1 = \{STIS_1, m, < w_1, w_2, ..., w_m >\}$, where

$STIS_1$, is mapped to the spatiotemporal database for the taxi;

$m = 7$, there are seven attributes in the database;

$< w_1, w_2, w_3, w_4, w_5, w_6, w_7 >=< 4, 4, 4, 4, 3, 5, 9 >$.

And we have $\rho = log_{|U|}^{|M|} = log_{323080}^{34559} = 0.824$.

The second data set is from the GPS data of bus in a city, for the GPS interval point with a bus, an event is made of speed, time, grade of service, and space layout. We deal with the dataset to form 40600 spatiotemporal events; the dataset can be mapped to a spatiotemporal information system $STIS_2$, and then we also can create a mixed radix notation system based on the $STIS_2$, which can be expressed as follows:

$M_2 = \{STIS_2, m, < w_1, w_2, ..., w_m >\}$, where

$STIS_2$, is mapped to the spatiotemporal database for the bus;

$m = 6$, there are six attributes in the database;

$< w_1, w_2, w_3, w_4, w_5, w_6, w_7 >=< 3, 5, 4, 6, 7, 8 >$.

And we also have $\rho = log_{|U|}^{|M|} = log_{40600}^{20159} = 0.934$.

Experimental environment is Microsoft Window XP Professional with Intel (R) Core (TM)2 Duo CPU (T6570 @) 2.10 GHz 1.19GHz) and 1.99 GB memory. The software development environment is based on C# with Microsoft Visual Studio 2008.

5.1 The experiments of performance comparison

Here, for discovering frequent spatiotemporal association patterns on the two datasets, we compare the algorithm DFSTAP with the Apriori Framework and the FP-growth Framework. Based on the performance comparison in Section 4.3, we respectively design three groups of experiments on the two datasets.

1. Testing on the first dataset

For the first dataset, we compare the performance as the number of frequent association pattern increases, and the test results are expressed as figure 1; as the maximal length of frequent association pattern increases, and the test results are expressed as figure 2; as the minimal support of frequent association pattern increases, and the test results are expressed as figure 3.

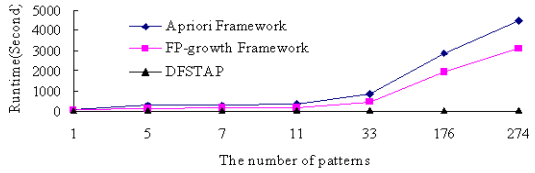


Figure 1: Performance comparison as the number of frequent association pattern increases.

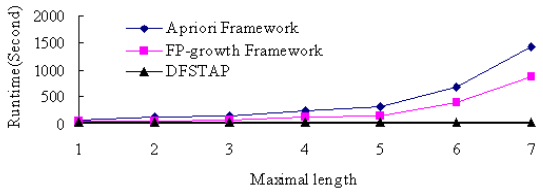


Figure 2: Performance comparison as the maximal length of frequent association pattern increases.

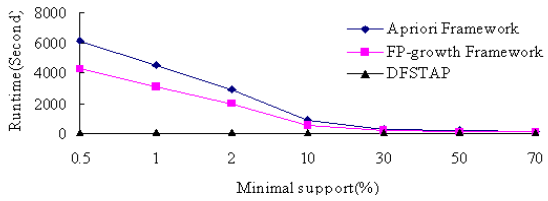


Figure 3: Performance comparison as the minimal support of frequent association pattern increases.

2. Testing on the second dataset

For the second dataset, we compare the performance from three aspects also; in other words, with the number of frequent association pattern, the maximal length, and the minimal support; and their experimental results are expressed as figures 4, 5, and 6, respectively.

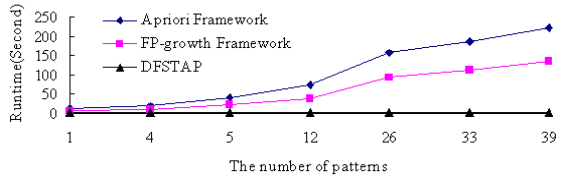


Figure 4: Performance comparison as the number of frequent association pattern increases.

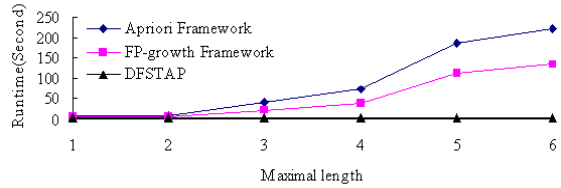


Figure 5: Performance comparison as the maximal length of frequent association pattern increases.

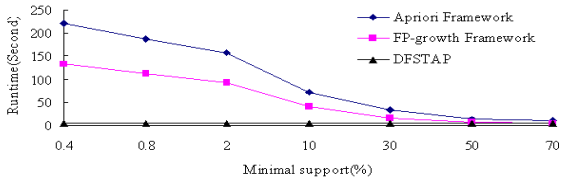


Figure 6: Performance comparison as the minimal support of frequent association pattern increases.

Based on these comparison results from figure 1 to 6, we can draw two conclusions as follows:

One is that the algorithm DFSTAP is better than the Apriori Framework and the FP-growth Framework on the type of mining dataset ($\rho \le 1$).

The other is that the performance of the algorithm DFSTAP does not depend on the number of frequent association pattern, the maximal length, and the minimal support parameter.

5.2 The experiments of discussing the optimal application environments

In this part, we mainly discuss the relationships between the performance and the following parameters:

$|U|$, is the number of spatiotemporal events;

$|M|$, is the combinatorial number for attribute values;

ρ, is the combinatorial number ratio.

1. Testing on the first dataset

For the first dataset, we change database events or database structure to create eight new datasets as table 3.

Name	Weight set	ρ
Data_T 1	<4,4,4,5,9>	$log_{403850}^{2879} = 0.617$
Data_T 2	<4,4,4,5,9>	$log_{323080}^{2879} = 0.628$
Data_T 3	<4,4,4,4,3,5,9>	$log_{403850}^{34559} = 0.810$
The first dataset	<4,4,4,4,3,5,9>	$log_{323080}^{34559} = 0.824$
Data_T 4	<4,4,4,4,3,5,9,3>	$log_{403850}^{103679} = 0.895$
Data_T 5	<4,4,4,4,3,5,9,3>	$log_{323080}^{103679} = 0.910$
Data_T 6	<4,4,4,5,9>	$log_{3231}^{2879} = 0.986$
Data_T 7	<4,4,4,4,3,5,9>	$log_{3231}^{34559} = 1.293$
Data_T 8	<4,4,4,4,3,5,9,3>	$log_{3231}^{103679} = 1.429$

Table 3: Changing description of the first dataset.

As we all know, the performance of the FP-growth Framework is better than the Apriori Framework, so we do not directly compare them in these experiments.

Here, if the minimal support is less than 1%, then we regard it as the lower support; if the minimal support is greater than 30%, then we regard it as the higher support.

(1) The relationship between the performance and ρ (the combinatorial number ratio)

As the minimal support increases, we compare the algorithm DFSTAP with the Apriori Framework and the

FP-growth Framework on the eight datasets. Their results are respectively expressed as figures 7-14.

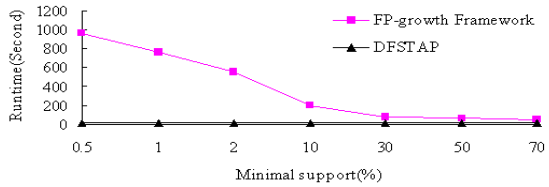


Figure 7: Performance comparison on Data_T 1 ($\rho = 0.617$)

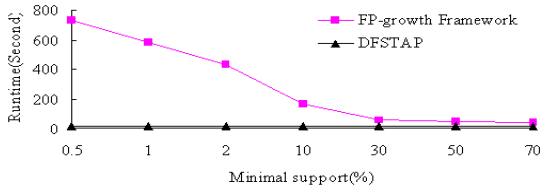


Figure 8: Performance comparison on Data_T 2 ($\rho = 0.628$)

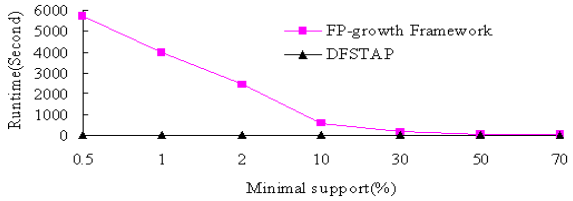


Figure 9: Performance comparison on Data_T 3 ($\rho = 0.810$)

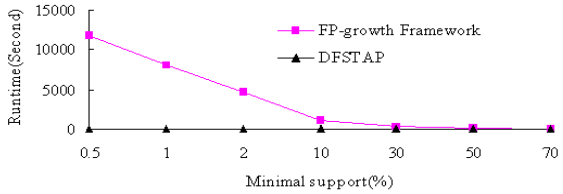


Figure 10: Performance comparison on Data_T 4 ($\rho = 0.895$)

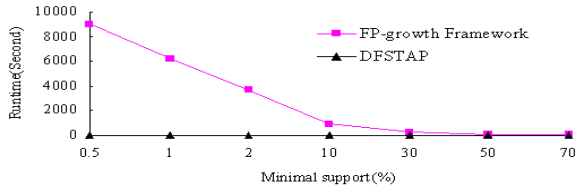


Figure 11: Performance comparison on Data_T 5 ($\rho = 0.910$)

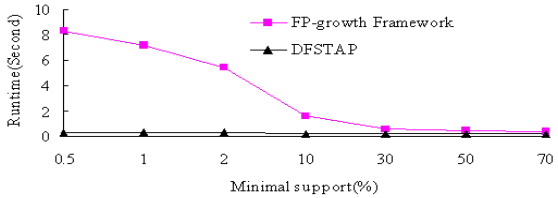


Figure 12: Performance comparison on Data_T 6 ($\rho = 0.986$)

Based on figures 7-12, when $\rho \le 1$, we can know that the performance of the algorithm DFSTAP is better than the Apriori Framework and the FP-growth Framework.

Based on figures 13 and 14, when $\rho > 1$, we can know that the performance of the algorithm DFSTAP is better than the Apriori Framework and the FP-growth Framework for the lower support; but for the higher support, it is not better than them.

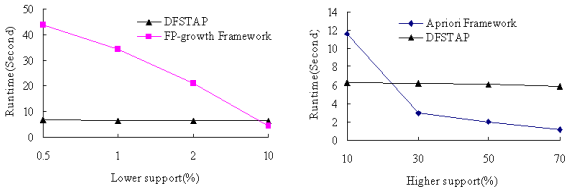


Figure 13: Performance comparison on Data_T 7 ($\rho = 1.293$)

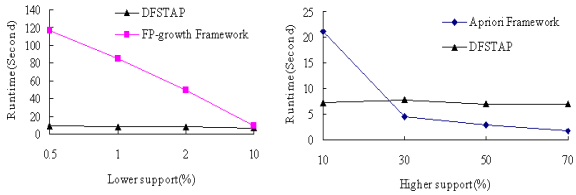


Figure 14: Performance comparison on Data_T 8 ($\rho = 1.429$)

(2) The relationship between the performance and $|M|$ (the combinatorial number for attribute values)

Here, we discuss the variation trend of runtime as the combinatorial number $|M|$ increases when the number of events $|U|$ is invariant. These experimental results can be expressed as figure 15.

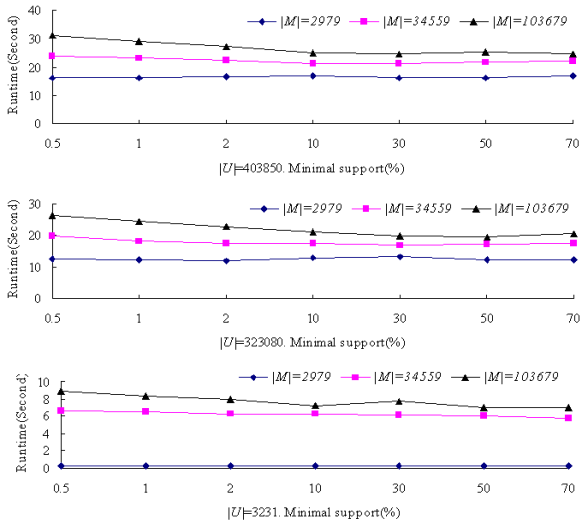


Figure 15: Performance comparison with $|M|$ varying

Based on these results from figure 15, we can know that the runtime of the algorithm DFSTAP is ascending as $|M|$ increases when $|U|$ is invariant.

(3) The relationship between the performance and $|U|$ (the number of events)

When $|M|$ is invariant, we discuss the variation trend of runtime as $|U|$ increases. These experimental results are expressed as figure 16.

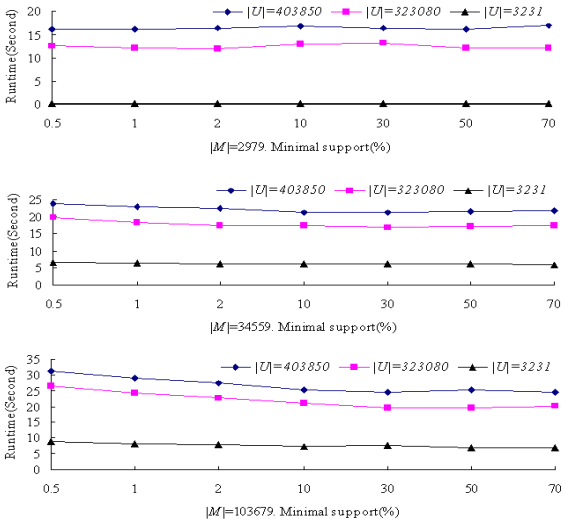


Figure 16: Performance comparison with $|U|$ varying

Based on these results from figure 16, when $|M|$ is invariant, we can know that the runtime of the algorithm DFSTAP is ascending as $|U|$ increases.

2. Testing on the second dataset

For the second dataset, we also use the same method to create eight new datasets as table 4, and compare the performance on these datasets.

Name	Weight set	ρ
Data_B 1	<3,5,6,7,8>	$log_{121800}^{5039} = 0.728$
Data_B 2	<3,5,6,7,8>	$log_{40600}^{5039} = 0.803$
Data_B 3	<3,5,4,6,7,8>	$log_{121800}^{20159} = 0.846$
The second dataset	<3,5,4,6,7,8>	$log_{40600}^{20159} = 0.934$
Data_B 4	<3,5,4,6,7,8,3,6>	$log_{121800}^{362879} = 1.093$
Data_B 5	<3,5,6,7,8>	$log_{2030}^{5039} = 1.119$
Data_B 6	<3,5,4,6,7,8,3,6>	$log_{40600}^{362879} = 1.206$
Data_B 7	<3,5,4,6,7,8>	$log_{2030}^{20159} = 1.301$
Data_B 8	<3,5,4,6,7,8,3,6>	$log_{2030}^{362879} = 1.954$

Table 4: Changing description of the second dataset.

(1) The relationship between the performance and ρ (the combinatorial number ratio)

We use the same method to test on the eight datasets. Their results are respectively expressed as figures 17-24.

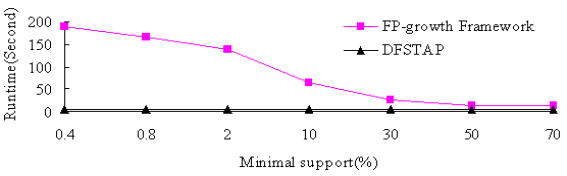


Figure 17: Performance comparison on Data_B 1 ($\rho = 0.728$)

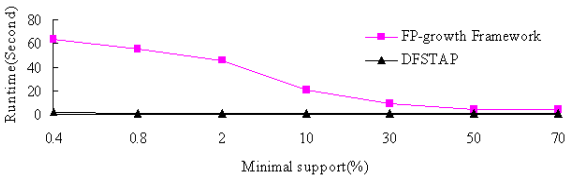


Figure 18: Performance comparison on Data_B 2 ($\rho = 0.803$)

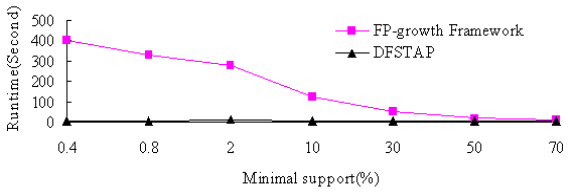


Figure 19: Performance comparison on Data_B 3 ($\rho = 0.846$)

Based on figures 17-19, when $\rho \le 1$, we can know the performance of the algorithm DFSTAP is better than the Apriori Framework and the FP-growth Framework.

Based on figures 20-24, when $\rho > 1$, we can know the performance of the algorithm DFSTAP is better than the Apriori Framework and the FP-growth Framework for the lower support; but for the higher support, it is not better than them.

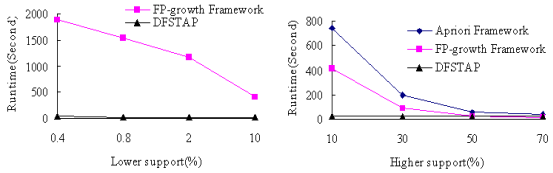


Figure 20: Performance comparison on Data_B 4 ($\rho = 1.093$)

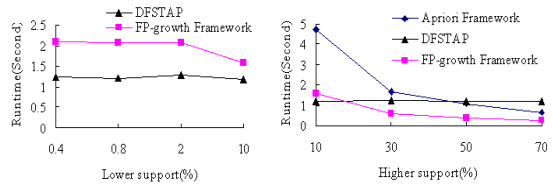


Figure 21: Performance comparison on Data_B 5 ($\rho = 1.119$)

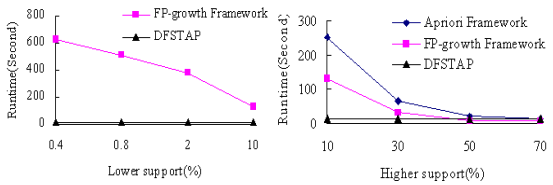


Figure 22: Performance comparison on Data_B 6 ($\rho = 1.206$)

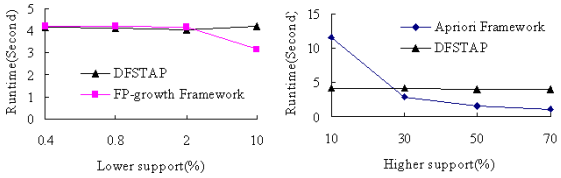


Figure 23: Performance comparison on Data_B 7 ($\rho = 1.301$)

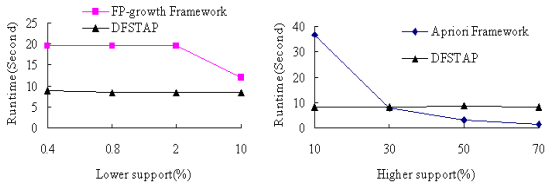


Figure 24: Performance comparison on Data_B 8 ($\rho = 1.954$)

(2) The relationship between the performance and $|M|$ (the combinatorial number for attribute values)

Here, when $|U|$ is invariant, we discuss the variation trend of runtime as $|M|$ increases. These experimental results are expressed as figure 25.

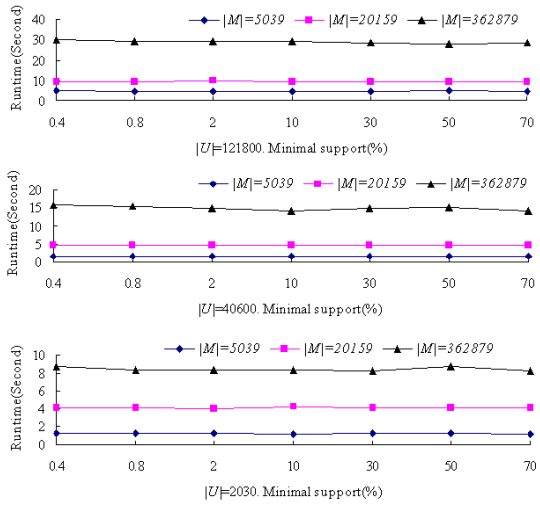


Figure 25: Performance comparison with $|M|$ varying

Based on figure 25, we can know that the runtime of the algorithm DFSTAP is ascending as $|M|$ increases when $|U|$ is invariant.

(3) The relationship between the performance and $|U|$ (the number of events)

When $|M|$ is invariant, we discuss the variation trend of runtime as $|U|$ increases. These experimental results are expressed as figure 26.

Based on figure 26, we can know that the runtime of algorithm DFSTAP is ascending as $|U|$ increases when $|M|$ is invariant.

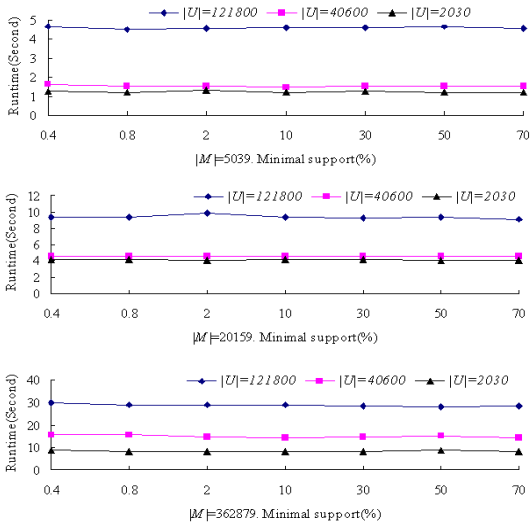


Figure 26: Performance comparison with $|U|$ varying

According to all these experimental results, we can draw the following conclusions:

(1) When the number of events $|U|$ is invariant, the runtime of the algorithm DFSTAP is ascending as the combinatorial number for attribute values $|M|$ increases. Namely, the performance is inversely proportional to the combinatorial number for attribute values $|M|$.

(2) When the combinatorial number for attribute values $|M|$ is invariant, the runtime of the DFSTAP is ascending as the number of events $|U|$ increases. Namely, the performance is inversely proportional to the number of events $|U|$.

(3) For mining frequent spatiotemporal association patterns, the performance of the DFSTAP is better than the Apriori Framework and the FP-growth Framework on the type of datasets ($\rho \le 1$).

On the type of datasets $\rho > 1$, the algorithm DFSTAP is suitable for mining frequent spatiotemporal association patterns with the lower support; but it is unsuitable for mining frequent spatiotemporal association patterns with the higher support.

(4) Since the computing environments generally has the performance bottleneck, when $|M| > \mu$ and $\rho \le 1$ (μ is a parameter with the computing environments), the performance of the DFSTAP also become much worse than the other. For our computing environments in this paper, if $|M| > \mu = 2^{25}$, the interval $[1,|M|]$ is too large, the performance will become much worse.

Hence, the optimal application environments for the algorithm DFSTAP is $|M| \le \mu, \rho \le 1$ (μ is a parameter with the computing environments).

6 Conclusion

In order to simply fast discovering multi-dimensional frequent spatiotemporal association patterns, in this paper, firstly, we construct a star association model based on event, the method of forming association patterns for the

model is very flexible, which can show more spatio-temporal information; and then propose an algorithm of discovering frequent spatiotemporal association patterns based on granular computing, which has two advantages; one is updating the mixed radix numeral to generate candidate; the method improves the speed of generating candidate. The other is adopting granular computing to discover frequent spatiotemporal association patterns to avoid repeatedly reading database. These experimental results indicate that the two key technologies improve the efficiency of algorithm. When $|M| \leq \mu$ (μ is a parameter with the computing environments), the algorithm is suitable for mining frequent patterns on the type of dataset ($\rho \leq 1$), and mining frequent association patterns with the lower support on the type of dataset ($\rho > 1$), but it is unsuitable for mining frequent association patterns with the higher support on the type of dataset ($\rho > 1$). Hence, we need to study the disadvantage in the future.

Acknowledgement

The authors would like to thank the anonymous reviewers for the constructive comment. This work was supported by the Chongqing education commission of science and technology research projects (#KJ121107, #KJ121111, KJ131108) in China.

References

[1] Cucchiara R., Piccardi M., Mello P. (2000). Image analysis and rule-based reasoning for a traffic monitoring system. IEEE Transactions on Intelligent Transportation Systems, IEEE Press, vol. 1, no. 2, pp.119-130.

[2] Pandey G., Atluri G., Steinbach M., et al (2009). An association analysis approach to biclustering. In Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Paris, France, pp. 677-686.

[3] Lee I., Phillips P. (2008). Urban crime analysis through areal categorized multivariate associations mining. Applied Artificial Intelligence, vol. 22, no. 5, pp.483-499.

[4] Huang Y., Kao L., Sandnes F. (2007). Predicting ocean salinity and temperature variations using data mining and fuzzy inference. International Journal of Fuzzy Systems, vol. 9, no. 3, pp. 143-151.

[5] Chang C., Shyue S. (2009). Association rules mining with GIS: An application to Taiwan census 2000. In Proceedings of the 6th international conference on Fuzzy systems and knowledge discovery, Tianjin, China, pp. 65-69.

[6] Zeitouni K., Yeh L., Aufaure M. (2000). Join indices as a tool for spatio data mining. In Proceedings of International Workshop on Temporal, Spatio and Spatiotemporal Data Mining (Berlin, Springer), pp. 102-114.

[7] Mennis J., Liu J. (2005). Mining association rules in spatio-temporal data: An analysis of urban socioeconomic and land cover change. Transactions in GIS, vol. 9, no. 1, pp, 5-17.

[8] Yang H., Parthasarathy S. (2006). Mining spatio and spatio-temporal patterns in scientific data. In Proceedings of 22nd International Conference on Data Engineering Workshops, Atlanta, GA, USA, pp. x146.

[9] Lee I. (2004). Mining multivariate associations within GIS environments. In Proceedings of 17th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, Ottawa, Canada, pp. 1062-1071.

[10] Ding W., Eick C., Wang J., et al. (2006). A framework for regional association rule mining in spatio datasets. In Proceedings of the Sixth IEEE International Conference on Data Mining, IEEE Press, Hong Kong, pp. 851-856.

[11] Yang H., Parthasarathy S., Mehta S. (2005). Mining spatio object associations for scientific data. In Proceedings of the 19th International Joint Conference on Artificial Intelligence, Edinburgh, UK, pp.902-907.

[12] Jong S.P., Chen M.S., and Yu P.S. (1997). Using a hash-based method with transaction trimming for mining association rules. IEEE Transactions on Knowledge and Data Engineering, IEEE Press, vol. 9, no. 5, pp. 813-825.

[13] Han J.W., Pei J., and Yin Y.W. et al.(2004). Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Mining and Knowledge Discovery, vol. 8, no. 1, pp.53-87.

[14] Lee H., Han J., Miller H., et al. (2007). Temporal and spatiotemporal data mining. IGI Publishing, New York.

[15] Tanbeer S., Ahmed C., Jeong B., et al. (2009). Efficient single-pass frequent pattern mining using a prefix-tree, Information Sciences, vol.179, no.5, pp.559-583.

[16] Lee A.J.T., Liu Y.H., Tsai H.M., et al. (2008). Mining frequent patterns in image databases with 9D-SPA representation. The Journal of Systems and Software, vol.82, no.4, pp. 603-618.

[17] Hobbs J. R. (1985). Granularity. In Proceedings of the 9th International Joint Conference on Artificial Intelligence, San Francisco, USA, pp. 432-435.

[18] Giunchglia F., Walsh T. (1992). A theory of abstraction. Artificial Intelligence, vol. 57, no. 2-3. pp. 323-389.

[19] Yao Y.Y. (2004). A partition model of granular computing. Lecture Notes in Computer Science Transactions on Rough Sets, vol. 3100, pp.232–253.

[20] Pawlak Z. (1998). Granularity of knowledge, indiscernibility and rough sets. In Proceedings of IEEE Int Conf on Fuzzy Systems, IEEE Press, Anchorage, AK, pp.106–110.

[21] Zhang L., Zhang B. (2003). The quotient space theory of problem solving. Lecture Notes in Computer Science, vol. 2639, pp. 11–15.

A Fast Implementation of Rules Based Machine Translation Systems for Similar Natural Languages

Jernej Vičič
Faculty of Mathematics, Natural Sciences and Information Technologies
University of Primorska
E-mail: jernej.vicic@upr.si
http://www.jt.upr.si/doktoratjernej/thesis/final/

Thesis Summary

Keywords: machine translation, machine translation of related languages, shallow transfer RBMT, RBMT

Received: March 25, 2013

This paper is and extended abstract of the doctoral thesis [1]. It presents an overview of the systems and methods for the natural language machine translation. It focuses primarily on systems and methods for shallow transfer rule based machine translation which are better suited for the translation of related languages. The major problem of the rule-based translation systems is costly manual production of dictionaries and translation rules in the case of a classical approach to building such systems. The work provides an overview over the collection of selected and new methods designed for automatic production of materials for the installation of systems based on translation rules.

Povzetek: Pričujoče delo je razširjen povzetek doktorske disertacije [1]. Predstavlja pregled strojnega prevajanja naravnih jezikov, osredotoča se predvsem na sisteme in metode za prevajanje na osnovi pravil plitkega prenosa, ki so najprimernejše za sorodne naravne jezike. Največja težava sistemov, ki temeljijo na pravilih, je dolgotrajna in draga ročna izdelava slovarjev ter prevajalnih pravil v primeru klasičnega pristopa h gradnji prevajalnih sistemov na osnovi pravil. Delo ponuja pregled zbirke izbranih in na novo zasnovanih metod samodejne izdelave gradiv za postavitev prevajalnih sistemov na osnovi pravil.

1 Introduction and problem statement

The paper presents an attempt to automate all data creation processes of a rule-based shallow-transfer machine translation system and its background. Several methods that automate some parts of the shallow transfer Rule Based Machine Translation (RBMT) system construction have been presented and are even used as part of the construction toolkits like Apertium [2], which is a widely used open source toolkit for creating machine translation systems between related languages.

Parts of the creation process have been addressed by several authors, some of these technologies have been used in our experiments along with newly developed methods. All methods and materials discussed in this paper were tested on a fully functional machine translation system based on Apertium. The system uses an architecture similar to the one presented in Figure 1.

Although it seems that Statistical Machine Translation (SMT) would be a perfect choice as some of the best performing machine translation systems are based on the SMT technologies, the stochastic approach has a couple of drawbacks that cannot be ignored; the SMT systems, to be successful, require huge amounts of parallel texts.

Another reason for choosing the RBMT approach is the nature of the languages involved in our experiments (Slovenian paired with Serbian, Czech, English and Estonian language). These are languages with rich inflectional morphology and as such they present a big problem for SMT.

Last but not least reason for using an RBMT machine translation system is the chance for the linguistic experts to further refine the results of the automatically produced data and thus to be able to improve the system in a controlled way.

2 Methodology

The modules presented in Figure 1 and numbered with numbers 1 through 5 require linguistic data (monolingual dictionaries, bilingual dictionaries, translation rules, etc.). Each module was examined and a method for linguistic data creation was designed.

The following types of data are needed for all modules of the system: the monolingual source dictionary with morphological information for source language parsing, monolingual target dictionary with morphological information for target language generation, bilingual translation dictionary, finite-state rules for shallow transfer and local agree-

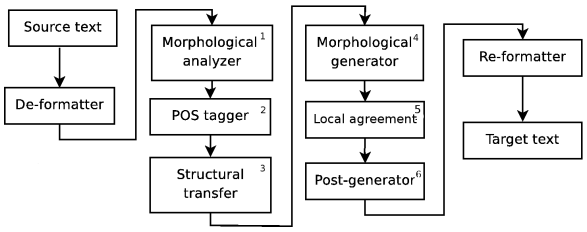


Figure 1: The modules of a typical shallow transfer translation system. The system [2] follows this design. An addition of the original architecture is the local agreement module tagged as number 6.

ment, statistical target language model, modeled source language tags.

3 Evaluation methodology and results

The evaluation focused only on the translation quality; the translation speed and responsiveness of the system, user-friendliness and other features of the translation systems are not presented. Were used the following methods: the automatic objective evaluation using the METEOR [3] metric, the non-automatic evaluation using weighted Levenshtein edit-distance [4] on a human corrected output of the translation system, the non-automatic subjective evaluation following [5] guidelines. The translation system was constructed according to the methodology presented in Section 2 using the selected training set. The evaluated values in each fold and the average final values are presented.

4 Discussion and further work

The agreement among all three evaluation methods is quite high, which shows that the results of the evaluation process are valid. The translation quality of the Slovenian-Serbian translation system is higher than the systems for distant language pairs. This can be attributed to the fact that the similarity of the first language pair is bigger.

The automatically generated linguistic data is far from perfect and additional manual labor will have to be executed in order to obtain better translation quality.

References

[1] J. Vičič, "Hitra postavitev prevajalnih sistemov na osnovi pravil za sorodne naravne jezike," Ph.D. dissertation, Univerza v Ljubljani, 2012. [Online]. Available: http://eprints.fri.uni-lj.si/1778/

[2] S. A. M. Corbi-Bellot, M. L. Forcada, Ortiz-Rojas, "An open-source shallow-transfer machine translation engine for the Romance languages of Spain," in *EAMT*, 2005, pp. 79–86.

[3] A. Lavie and M. J. Denkowski, "The Meteor metric for automatic evaluation of machine translation," *Machine Translation*, vol. 23, no. 2-3, pp. 105–115, Sep. 2009.

[4] K. S. Fu, *Syntactic Pattern Recognition and Applications*. Prentice-Hall, Englewood Cliffs, NJ, 1982.

[5] LDC, "Linguistic data annotation specification: Assessment of fluency and adequacy in translations," LDC, Tech. Rep., 2005.

CONTENTS OF *Informatica* Volume 37 (2013) pp. 1–459

Papers

Editorials

JOŽEF STEFAN INSTITUTE

Jožef Stefan (1835-1893) was one of the most prominent physicists of the 19th century. Born to Slovene parents, he obtained his Ph.D. at Vienna University, where he was later Director of the Physics Institute, Vice-President of the Vienna Academy of Sciences and a member of several scientific institutions in Europe. Stefan explored many areas in hydrodynamics, optics, acoustics, electricity, magnetism and the kinetic theory of gases. Among other things, he originated the law that the total radiation from a black body is proportional to the 4th power of its absolute temperature, known as the Stefan–Boltzmann law.

The Jožef Stefan Institute (JSI) is the leading independent scientific research institution in Slovenia, covering a broad spectrum of fundamental and applied research in the fields of physics, chemistry and biochemistry, electronics and information science, nuclear science technology, energy research and environmental science.

The Jožef Stefan Institute (JSI) is a research organisation for pure and applied research in the natural sciences and technology. Both are closely interconnected in research departments composed of different task teams. Emphasis in basic research is given to the development and education of young scientists, while applied research and development serve for the transfer of advanced knowledge, contributing to the development of the national economy and society in general.

At present the Institute, with a total of about 900 staff, has 700 researchers, about 250 of whom are postgraduates, around 500 of whom have doctorates (Ph.D.), and around 200 of whom have permanent professorships or temporary teaching assignments at the Universities.

In view of its activities and status, the JSI plays the role of a national institute, complementing the role of the universities and bridging the gap between basic science and applications.

Research at the JSI includes the following major fields: physics; chemistry; electronics, informatics and computer sciences; biochemistry; ecology; reactor technology; applied mathematics. Most of the activities are more or less closely connected to information sciences, in particular computer sciences, artificial intelligence, language and speech technologies, computer-aided design, computer architectures, biocybernetics and robotics, computer automation and control, professional electronics, digital communications and networks, and applied mathematics.

The Institute is located in Ljubljana, the capital of the independent state of S**love**nia (or S♡nia). The capital today is considered a crossroad between East, West and Mediterranean Europe, offering excellent productive capabilities and solid business opportunities, with strong international connections. Ljubljana is connected to important centers such as Prague, Budapest, Vienna, Zagreb, Milan, Rome, Monaco, Nice, Bern and Munich, all within a radius of 600 km.

From the Jožef Stefan Institute, the Technology park "Ljubljana" has been proposed as part of the national strategy for technological development to foster synergies between research and industry, to promote joint ventures between university bodies, research institutes and innovative industry, to act as an incubator for high-tech initiatives and to accelerate the development cycle of innovative products.

Part of the Institute was reorganized into several high-tech units supported by and connected within the Technology park at the Jožef Stefan Institute, established as the beginning of a regional Technology park "Ljubljana". The project was developed at a particularly historical moment, characterized by the process of state reorganisation, privatisation and private initiative. The national Technology Park is a shareholding company hosting an independent venture-capital institution.

The promoters and operational entities of the project are the Republic of Slovenia, Ministry of Higher Education, Science and Technology and the Jožef Stefan Institute. The framework of the operation also includes the University of Ljubljana, the National Institute of Chemistry, the Institute for Electronics and Vacuum Technology and the Institute for Materials and Construction Research among others. In addition, the project is supported by the Ministry of the Economy, the National Chamber of Economy and the City of Ljubljana.

Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel.:+386 1 4773 900, Fax.:+386 1 251 93 85
WWW: http://www.ijs.si
E-mail: matjaz.gams@ijs.si
Public relations: Polona Strnad

INFORMATICA

AN INTERNATIONAL JOURNAL OF COMPUTING AND INFORMATICS

INVITATION, COOPERATION

Submissions and Refereeing

Please submit a manuscript at: http://www.informatica.si/Editors/PaperUpload.asp. At least two referees outside the author's country will examine it, and they are invited to make as many remarks as possible from typing errors to global philosophical disagreements. The chosen editor will send the author the obtained reviews. If the paper is accepted, the editor will also send an email to the managing editor. The executive board will inform the author that the paper has been accepted, and the author will send the paper to the managing editor. The paper will be published within one year of receipt of email with the text in Informatica MS Word format or Informatica LaTeX format and figures in .eps format. Style and examples of papers can be obtained from http://www.informatica.si. Opinions, news, calls for conferences, calls for papers, etc. should be sent directly to the managing editor.

QUESTIONNAIRE

☐ Send Informatica free of charge

☐ Yes, we subscribe

Please, complete the order form and send it to Dr. Drago Torkar, Informatica, Institut Jožef Stefan, Jamova 39, 1000 Ljubljana, Slovenia. E-mail: drago.torkar@ijs.si

Since 1977, Informatica has been a major Slovenian scientific journal of computing and informatics, including telecommunications, automation and other related areas. In its 16th year (more than nineteen years ago) it became truly international, although it still remains connected to Central Europe. The basic aim of Informatica is to impose intellectual values (science, engineering) in a distributed organisation.

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the Refereeing Board.

Informatica is free of charge for major scientific, educational and governmental institutions. Others should subscribe (see the last page of Informatica).

ORDER FORM – INFORMATICA

Name: .....

Title and Profession (optional): .....

.....

Home Address and Telephone (optional): .....

.....

Office Address and Telephone (optional): .....

.....

E-mail Address (optional): .....

Signature and Date: .....

Informatica WWW:

http://www.informatica.si/

Referees from 2008 on:

A. Abraham, S. Abraham, R. Accornero, A. Adhikari, R. Ahmad, G. Alvarez, N. Anciaux, R. Arora, I. Awan, J. Azimi, C. Badica, Z. Balogh, S. Banerjee, G. Barbier, A. Baruzzo, B. Batagelj, T. Beaubouef, N. Beaulieu, M. ter Beek, P. Bellavista, K. Bilal, S. Bishop, J. Bodlaj, M. Bohanec, D. Bolme, Z. Bonikowski, B. Bošković, M. Botta, P. Brazdil, J. Brest, J. Brichau, A. Brodnik, D. Brown, I. Bruha, M. Bruynooghe, W. Buntine, D.D. Burdescu, J. Buys, X. Cai, Y. Cai, J.C. Cano, T. Cao, J.-V. Capella-Hernández, N. Carver, M. Cavazza, R. Ceylan, A. Chebotko, I. Chekalov, J. Chen, L.-M. Cheng, G. Chiola, Y.-C. Chiou, I. Chorbev, S.R. Choudhary, S.S.M. Chow, K.R. Chowdhury, V. Christlein, W. Chu, L. Chung, M. Ciglarič, J.-N. Colin, V. Cortellessa, J. Cui, P. Cui, Z. Cui, D. Cutting, A. Cuzzocrea, V. Cvjetkovic, J. Cypryjanski, L. Čehovin, D. Čerepnalkoski, I. Čosić, G. Daniele, G. Danoy, M. Dash, S. Datt, A. Datta, M.-Y. Day, F. Debili, C.J. Debono, J. Dedič, P. Degano, A. Dekdouk, H. Demirel, B. Demoen, S. Dendamrongvit, T. Deng, A. Derezinska, J. Dezert, G. Dias, I. Dimitrovski, S. Dobrišek, Q. Dou, J. Doumen, E. Dovgan, B. Dragovich, D. Drajic, O. Drbohlav, M. Drole, J. Dujmović, O. Ebers, J. Eder, S. Elaluf-Calderwood, E. Engström, U. riza Erturk, A. Farago, C. Fei, L. Feng, Y.X. Feng, B. Filipič, I. Fister, I. Fister Jr., D. Fišer, A. Flores, V.A. Fomichov, S. Forli, A. Freitas, J. Fridrich, S. Friedman, C. Fu, X. Fu, T. Fujimoto, G. Fung, S. Gabrielli, D. Galindo, A. Gambarara, M. Gams, M. Ganzha, J. Garbajosa, R. Gennari, G. Georgeson, N. Gligorić, S. Goel, G.H. Gonnet, D.S. Goodsell, S. Gordillo, J. Gore, M. Grčar, M. Grgurović, D. Grosse, Z.-H. Guan, D. Gubiani, M. Guid, C. Guo, B. Gupta, M. Gusev, M. Hahsler, Z. Haiping, A. Hameed, C. Hamzaçebi, Q.-L. Han, H. Hanping, T. Härder, J.N. Hatzopoulos, S. Hazelhurst, K. Hempstalk, J.M.G. Hidalgo, J. Hodgson, M. Holbl, M.P. Hong, G. Howells, M. Hu, J. Hyvärinen, D. Ienco, B. Ionescu, R. Irfan, N. Jaisankar, D. Jakobović, K. Jassem, I. Jawhar, Y. Jia, T. Jin, I. Jureta, Đ. Juričić, S. K, S. Kalajdziski, Y. Kalantidis, B. Kaluža, D. Kanellopoulos, R. Kapoor, D. Karapetyan, A. Kassler, D.S. Katz, A. Kaveh, S.U. Khan, M. Khattak, V. Khomenko, E.S. Khorasani, I. Kitanovski, D. Kocev, J. Kocijan, J. Kollár, A. Kontostathis, P. Korošec, A. Koschmider, D. Košir, J. Kovač, A. Krajnc, M. Krevs, J. Krogstie, P. Krsek, M. Kubat, M. Kukar, A. Kulis, A.P.S. Kumar, H. Kwaśnicka, W.K. Lai, C.-S. Laih, K.-Y. Lam, N. Landwehr, J. Lanir, A. Lavrov, M. Layouni, G. Leban, A. Lee, Y.-C. Lee, U. Legat, A. Leonardis, G. Li, G.-Z. Li, J. Li, X. Li, X. Li, Y. Li, Y. Li, S. Lian, L. Liao, C. Lim, J.-C. Lin, H. Liu, J. Liu, P. Liu, X. Liu, X. Liu, F. Logist, S. Loskovska, H. Lu, Z. Lu, X. Luo, M. Luštrek, I.V. Lyustig, S.A. Madani, M. Mahoney, S.U.R. Malik, Y. Marinakis, D. Marinčič, J. Marques-Silva, A. Martin, D. Marwede, M. Matijašević, T. Matsui, L. McMillan, A. McPherson, A. McPherson, Z. Meng, M.C. Mihaescu, V. Milea, N. Min-Allah, E. Minisci, V. Mišić, A.-H. Mogos, P. Mohapatra, D.D. Monica, A. Montanari, A. Moroni, J. Mosegaard, M. Moškon, L. de M. Mourelle, H. Moustafa, M. Možina, M. Mrak, Y. Mu, J. Mula, D. Nagamalai, M. Di Natale, A. Navarra, P. Navrat, N. Nedjah, R. Nejabati, W. Ng, Z. Ni, E.S. Nielsen, O. Nouali, F. Novak, B. Novikov, P. Nurmi, D. Obrul, B. Oliboni, X. Pan, M. Pančur, W. Pang, G. Papa, M. Paprzycki, M. Paralič, B.-K. Park, P. Patel, T.B. Pedersen, Z. Peng, R.G. Pensa, J. Perš, D. Petcu, B. Petelin, M. Petkovšek, D. Pevec, M. Pičulin, R. Piltaver, E. Pirogova, V. Podpečan, M. Polo, V. Pomponiu, E. Popescu, D. Poshyvanyk, B. Potočnik, R.J. Povinelli, S.R.M. Prasanna, K. Pripužić, G. Puppis, H. Qian, Y. Qian, L. Qiao, C. Qin, J. Que, J.-J. Quisquater, C. Rafe, S. Rahimi, V. Rajkovič, D. Raković, J. Ramaekers, J. Ramon, R. Ravnik, Y. Reddy, W. Reimche, H. Rezankova, D. Rispoli, B. Ristevski, B. Robič, J.A. Rodriguez-Aguilar, P. Rohatgi, W. Rossak, I. Rožanc, J. Rupnik, S.B. Sadkhan, K. Saeed, M. Saeki, K.S.M. Sahari, C. Sakharwade, E. Sakkopoulos, P. Sala, M.H. Samadzadeh, J.S. Sandhu, P. Scaglioso, V. Schau, W. Schempp, J. Seberry, A. Senanayake, M. Senobari, T.C. Seong, S. Shamala, c. shi, Z. Shi, L. Shiguo, N. Shilov, Z.-E.H. Slimane, F. Smith, H. Sneed, P. Sokolowski, T. Song, A. Soppera, A. Sorniotti, M. Stajdohar, L. Stanescu, D. Strnad, X. Sun, L. Šajn, R. Šenkeřík, M.R. Šikonja, J. Šilc, I. Škrjanc, T. Štajner, B. Šter, V. Štruc, H. Takizawa, C. Talcott, N. Tomasev, D. Torkar, S. Torrente, M. Trampuš, C. Tranoris, K. Trojacanec, M. Tschierschke, F. De Turck, J. Twycross, N. Tziritas, W. Vanhoof, P. Vateekul, L.A. Vese, A. Visconti, B. Vlaovič, V. Vojisavljević, M. Vozalis, P. Vračar, V. Vranić, C.-H. Wang, H. Wang, H. Wang, H. Wang, S. Wang, X.-F. Wang, X. Wang, Y. Wang, A. Wasilewska, S. Wenzel, V. Wickramasinghe, J. Wong, S. Wrobel, K. Wrona, B. Wu, L. Xiang, Y. Xiang, D. Xiao, F. Xie, L. Xie, Z. Xing, H. Yang, X. Yang, N.Y. Yen, C. Yong-Sheng, J.J. You, G. Yu, X. Zabulis, A. Zainal, A. Zamuda, M. Zand, Z. Zhang, Z. Zhao, D. Zheng, J. Zheng, X. Zheng, Z.-H. Zhou, F. Zhuang, A. Zimmermann, M.J. Zuo, B. Zupan, M. Zuqiang, B. Žalik, J. Žižka,

Informatica

An International Journal of Computing and Informatics

Web edition of Informatica may be accessed at: http://www.informatica.si.

Subscription Information Informatica (ISSN 0350-5596) is published four times a year in Spring, Summer, Autumn, and Winter (4 issues per year) by the Slovene Society Informatika, Litostrojska cesta 54, 1000 Ljubljana, Slovenia.
The subscription rate for 2013 (Volume 37) is
– 60 EUR for institutions,
– 30 EUR for individuals, and
– 15 EUR for students
Claims for missing issues will be honored free of charge within six months after the publication date of the issue.

Typesetting: Borut Žnidar.
Printing: ABO grafika d.o.o., Ob železnici 16, 1000 Ljubljana.

Orders may be placed by email (drago.torkar@ijs.si), telephone (+386 1 477 3900) or fax (+386 1 251 93 85). The payment should be made to our bank account no.: 02083-0013014662 at NLB d.d., 1520 Ljubljana, Trg republike 2, Slovenija, IBAN no.: SI56020830013014662, SWIFT Code: LJBASI2X.

Informatica is published by Slovene Society Informatika (president Niko Schlamberger) in cooperation with the following societies (and contact persons):
Robotics Society of Slovenia (Jadran Lenarčič)
Slovene Society for Pattern Recognition (Janez Perš)
Slovenian Artificial Intelligence Society (Dunja Mladenić)
Cognitive Science Society (Urban Kordeš)
Slovenian Society of Mathematicians, Physicists and Astronomers (Andrej Likar)
Automatic Control Society of Slovenia (Sašo Blažič)
Slovenian Association of Technical and Natural Sciences / Engineering Academy of Slovenia (Vojteh Leskovšek)
ACM Slovenia (Andrej Brodnik)

Informatica is surveyed by: ACM Digital Library, Citeseer, COBISS, Compendex, Computer & Information Systems Abstracts, Computer Database, Computer Science Index, Current Mathematical Publications, DBLP Computer Science Bibliography, Directory of Open Access Journals, InfoTrac OneFile, Inspec, Linguistic and Language Behaviour Abstracts, Mathematical Reviews, MatSciNet, MatSci on SilverPlatter, Scopus, Zentralblatt Math

Volume 37 Number 4 December 2013 **ISSN 0350-5596**

Informatica

An International Journal of Computing and Informatics

Informatica **37** (2013) Number 4, pp. 359–459