# Table of Contents

1. **THE APPLICATION OF AGENT TECHNOLOGY IN STOCK PRICE PREDICTION** ................................................................. T1-1  
   *Itaza Afiani Mohtar, Zulaiha Ali Othman*

2. **AUTOMATIC FACIAL FEATURE POINTS DETECTION IN BESPECTACLED FACES** ...................................................... T1-8  
   *R.Saravanan, Usha Sridhar*

3. **DATA MINING IN RESERVOIR OPERATION AND FLOOD CONTROL USING ARTIFICIAL NEURAL NETWORKS** ........ T1-14  
   *Ramani Bai. V, Woo Chaw Seng, Faridha Othman, Gopinath Ramadas*

4. **DEFINING AGENT SYSTEM FOR KNOWLEDGE ACQUISITION IN PBL DISCUSSIONS** .............................................. T1-21  
   *Akcell Chiang, Mohd Sapiyan Baba*

5. **DETECTION OF OPTIC DISC IN COLOR FUNDUS IMAGES USING GREEN CHANNEL COMPONENT** .................... T1-27  
   *Ahmed wasif Reza, C. Eswaran, Subhas Hati, Sithi Vinayakam, Ahmed Fauzi Md Sharif*

6. **DISTRIBUTED ALGORITHM FOR CLUSTERING LARGE DATASETS** ............ T1-34  
   *Radhika M. Pai, Ananthanarayana V.S.*

7. **DNA SEQUENCE DATABASE CLASSIFICATION AND REDUCTION: ROUGH SETS THEORY APPROACH** ..................... T1-41  
   *M Nordin A Rahman, M Yazid M Saman, Aziz Ahmad, A Osman M Tap*

8. **A GENETIC-BASED FUZZY REGRESSION ANALYSIS: AN APPROACH FOR ESTIMATION OF SOFTWARE SOURCE CODE SIZE** ........ T1-49  
   *M. Hadi Mashinchi, Siti Mariyam HJ. Shamsuddin*

9. **HEART DISEASE DECISION SUPPORT SYSTEM USING DATA MINING CLASSIFICATION MODELING TECHNIQUES** ........ T1-54  
   * Sellappan Palaniappan, Rafiah Awang*

10. **INVESTIGATING MULTIOBJECTIVE OPTIMIZATION USING AN AUGMENTED COEVOLUTIONARY SPEA2 ALGORITHM** ............ T1-65  
    *Tse Guan Tan, Jason Teo, Hui Keng Lau*
<table>
<thead>
<tr>
<th>No.</th>
<th>Title</th>
<th>Authors</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>AN INVESTIGATION OF AN INTEGRATION OF INDIVIDUAL AND SOCIAL LEARNING IN AN EVOLUTIONARY APPROACH TO THE GAME OF TIC-TAC-TOE</td>
<td>Razali Yaakob, Graham Kendall</td>
<td>T1-72</td>
</tr>
<tr>
<td>13</td>
<td>PRECISE FINGERPRINT ENROLMENT THROUGH PROJECTION INCORPORATED SUBSPACE BASED ON PRINCIPAL COMPONENT ANALYSIS (PCA)</td>
<td>Md. Rajibul Islam, Md. Shohel Sayeed, Andrews Samraj</td>
<td>T1-85</td>
</tr>
<tr>
<td>14</td>
<td>PROBABILISTIC MODEL FOR RESPONSE GENERATION WITH APPLICATION OF BAYESIAN NETWORKS</td>
<td>Aida Mustapha, Md. Nasir Sulaiman, Ramlan Mahmod, Hasan Selamat</td>
<td>T1-92</td>
</tr>
<tr>
<td>15</td>
<td>A SET OF SCALAR FEATURES REPRESENTATION FOR 3D FACE RECOGNITION</td>
<td>Fatimah Khalid, Tengku Mohd. Tengku Sembok, Khairuddin Omar</td>
<td>T1-99</td>
</tr>
<tr>
<td>16</td>
<td>A STUDY ON NORMALIZATION AND FEATURE EXTRACTION TECHNIQUES FOR THE RECOGNITION OF ISOLATED KANNADA HANDWRITTEN NUMERALS</td>
<td>Surekha Patil, Meenakshi Patil</td>
<td>T1-105</td>
</tr>
<tr>
<td>17</td>
<td>TOOL SUPPORT FOR COMPETENCY EVALUATION OF WEB-ONTOLOGIES</td>
<td>Muthukkaruppan Annamalai, Noor Hasimah IbrahIm Teo</td>
<td>T1-110</td>
</tr>
<tr>
<td>18</td>
<td>THE USE OF A DECISION SUPPORT FRAMEWORK IN A MULTI-CRITERIA PURCHASE DECISIONS</td>
<td>P.D.D. Dominic, Shakirah Mohd Taib, Emy Elyanee Mustapha</td>
<td>T1-118</td>
</tr>
<tr>
<td>19</td>
<td>UTILIZING TOP-DOWN PARSING TECHNIQUE IN THE DEVELOPMENT OF A MALAY LANGUAGE SENTENCE PARSER</td>
<td>Ahmad I. Z. Abidin, S. P. Yong, Rozana Kasbon, Hazreen Azman</td>
<td>T1-125</td>
</tr>
</tbody>
</table>
THE APPLICATION OF AGENT TECHNOLOGY IN STOCK PRICE PREDICTION

Itaza Afiani Mohtar 1, Zulaiha Ali Othman2
1Faculty of Information Technology and Quantitative Science
Mara University of Technology, Seri Iskandar Campus, 32600 Bota, Perak
Email: itaza328@perak.uitm.edu.my
2Faculty of Technology and Information Science
Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor
Email: zao@ftsm.ukm.my

ABSTRACT

Stock price prediction which is based on the movements of the stocks is used by investors to make the right decision (buy or sell) at the right time. Artificial Neural Networks (ANN) is one of the tools used for prediction. Accuracy of prediction depends heavily on the data learned by the ANN. Therefore the more data learned, the better the accuracy. This paper discusses the use of agents in ANN for stock price prediction. The architecture of the multiagent system and the tasks of the three agents are also discussed. The agent system was compared with the traditional ANN system based on the time required to complete all the processes, the cost involved and the effect of additional data on the accuracy performance of the ANN. It was found that the agent system recorded completion of all the processes earlier than the traditional ANN system. This consequently contributed to lesser cost. The accuracy performance of the ANN increased with additional data collected daily by the agents. Conclusively, the use of agents in stock price prediction reduced time and cost and increased the accuracy of the ANN’s prediction.

Keywords: Multiagent system, Stock price prediction, Artificial neural network.

1.0 INTRODUCTION

Stock price prediction is a popular method used by investors to determine the best time to buy or sell their stocks. Artificial Neural Networks (ANN) is one of the most common techniques used for predicting stocks. ANN has been proven to produce relatively good predictions but it could not maintain its accuracy level if the knowledge learnt is not the latest. Therefore, to overcome this, data used for training must be updated with the current stock data. Updating a neural network’s knowledge is complex and time consuming. The new data must first be collected, preprocessed, transformed and included into the datasets. Then, the neural network will be retrained with the newly updated datasets before it can be used for predictions.

In data collection, more recent stock prices relative to the archived data must first be collected. This can be done by collecting current archived data available in the market. After the current data is obtained, it will undergo preprocessing. Data preprocessing involves ensuring the completeness of the data, representing the data and normalizing the data [1]. Data can be preprocessed manually or by using data mining software. After completing the data preprocessing stage, the data is then included into the training datasets. The ANN will be trained with the newly updated datasets. Figure 1 shows the common processes a neural network system has to undergo before it is able to predict the stock prices.

The aforementioned processes are complex, time consuming and user-initiated. Conclusively, a system that can execute all the processes autonomously would greatly reduce cost and at the same time ensure the level of prediction accuracy is maintained. This brings to the possibility of using the agent approach in designing the system. But what can agents offer? This is further explained in the next subsection.

1.1 What Can Agents Offer?

An agent is defined as “a computer system that is situated in some environment and that is capable of autonomous actions in this environment in order to meet its design objectives” [2]. The advantages offered by the agent paradigm to enable the neural network to maintain the level of prediction accuracy are:

- Enables automatic updates and preprocessing of data
- Increased performance
- Flexible design
- Scalability
- Easy to maintain

Data updating and preprocessing without user initiation can be achieved by assigning the agents the relevant tasks to accomplish the goals. Each agent in the system will be trusted to execute their tasks correctly and responsibly. With
autonomy, they have control over their actions which directly reduces user intervention. This ensures current data can be automatically added into the system. Figure 2 shows the processes that can be executed by the agents.

![Fig. 1: The Traditional ANN system’s processes](image)

![Figure 2: The agent-based ANN system](image)

In the multiagent system, all the processes; data updating and preprocessing, training and prediction are executed parallel by the agents. Therefore, user initiation is not needed. This ensures predictions based on the latest knowledge are obtainable without user having to execute the complex processes as shown in Figure 1.

Deciding when to train the data is crucial when resource is an issue. This is because training the datasets consumes a lot of processing power and takes a long time. By giving the agent the ability to decide when to do and whether ‘to do or not to do’ training will ensure optimal use of resources. This will increase the system’s performance.

The agent paradigm also offers flexible design approach when designing the system. If a task is seen as complex, it can be subdivided and allocated to agents in the system. This flexible design also enables easy system expansion in the future. For example, if after system execution, an agent is overloaded with tasks then an additional agent(s) can be added into the system to take on some of the tasks.

Agents are independent of each other but they collaborate to reach the system’s goal. This enables system scalability whereby if the performance of the system becomes slow, other than including additional hardware to upgrade the performance, some of the agents can be relocated into another machine or combined into the same machine. Although distributed, the agent will still have the ability to communicate with each other through agent communication language (ACL) thus ensuring that the system goal can still be achieved.

Over time, an agent might be facing a problem or needs to be upgraded. As independent entities, maintenance will only need to focus on the particular agent without having to troubleshoot the whole system. This means that other agents in the system can still carry on with their tasks without being distracted by the situation.

Based on the discussions above, there are many advantages of using the agent approach to develop a neural network based stock price prediction system. In the next section, the design of the multiagent system is explained.
2.0 The Multiagent System Architecture

The stock price prediction system using the multiagent approach consists of three agents; DownloadAgent, NeuralNetAgent and UserAgent. DownloadAgent is responsible for collecting, parsing and preprocessing daily closing stock prices from a bourse’s website. It is also responsible for including the current stock prices into the existing training datasets.

NeuralNetAgent is responsible for doing the training of the datasets and predicting future stock prices. UserAgent on the other hand, acts as an intermediary between the user and the system by enabling the user to access the predictions. It uses the service of NeuralNetAgent to obtain predictions. Figure 3 illustrates the system architecture following an agent-based system development methodology by Othman [3]. CU represents the user-agent interaction while agent-agent interactions are represented by CA.

At 5.30 pm local time daily, DownloadAgent will go to the bourse’s website to collect closing stock prices by downloading and parsing the webpage for relevant data. The parsed data is then cleaned, represented and normalized before it is included into existing training datasets. DownloadAgent will also create a training schedule. It then informs NeuralNetAgent the datasets and the training schedules are available (CA-1.1).

NeuralNetAgent will first do the predictions. It will send a request for the input data for predictions (CA-1.2) from DownloadAgent. DownloadAgent will send the data (CA-1.3) and NeuralNetAgent will then proceed with the predictions.

After NeuralNetAgent has completed the predictions, it sends the results to UserAgent (CA-2.1). The system is now equipped with the latest precise predictions of future stock prices easily accessible by request from users through UserAgent (CU-1). Next, NeuralNetAgent will proceed with the training by first requesting for the training schedule from DownloadAgent (CA-1.4). NeuralNetAgent will perform the training process based on the schedule sent by DownloadAgent (CA-1.5).

2.1 DownloadAgent

DownloadAgent is responsible for collecting daily closing stock prices from the bourse’s website. After downloading the page, all HTML tags and stop words will be removed using a parser. Relevant data; open, high, low and close prices, identified by corresponding counter names will be extracted. The prices are first checked for completeness. In cases of null (N/A) data, it will be left null. After going through the cleaning process, the data then undergo transformation process. The transformation process involves the following:

- Derive the Volatility Indicator (VI) using the following equation
  \[
  VI = \frac{P_H - P_L \times (P_C - P_O)}{(P_H + P_L + P_C + P_O) \times \frac{1}{4}}
  \]
  where,
  - \(P_H\) = highest traded price during the day,
  - \(P_L\) = lowest traded price during the day,
  - \(P_O\) = opening price,
  - \(P_C\) = closing price.
- Transform using Moving Average for \(n = 5\) and \(n = 10\) using this equation

![Fig. 3: The Multiagent System Architecture](image-url)
where,

\[ MA = \sum_{i}^{n} \frac{P_i}{n} \]

\( n \) = number of time periods in the moving average.

After the transformation process is completed, the data is then updated into the training datasets.

DownloadAgent is also responsible for creating a simple schedule for the training session. The schedule is created based on the following rule:

1. If closing price \( t \neq \text{closing price } t-1 \)
   then send for training
2. else if high \( t \neq \text{high } t-1 \)
   then send for training
3. else if low \( t \neq \text{low } t-1 \)
   then send for training
4. else do not train

Other than updating and transforming the data, DownloadAgent is also responsible for granting access to the datasets when requested by NeuralNetAgent.

2.2 NeuralNetAgent

This agent is equipped with an ANN for predicting prices. The ANN modeled by Mohamad Faidzul [4] with the following properties:

- The input variables are:
  \( P_t \) = closing price for the day
  \( MA_5 \) = moving average for day 5.
  \( MA_{10} \) = moving average for day 10
  \( V_{t-1} \) = volatility indicator for day \(-1\)

- Output variable is:
  \( P_{t+1} \) = predicted price for tomorrow

- Training algorithm: Backpropagation
  - Momentum rate: 0.9
  - Learning rate: 0.3
  - Activation function: Symmetrical Sigmoid, \( S(x) = \frac{1 - e^{-x}}{1 + e^{-x}} \)

This ANN model produced an accuracy of up to 91%.

Before data is fed into the ANN, it has to be transformed using the following:

- Normal Relative Variable Difference function with the following equation
  \[ R_{t} = \frac{h_t - h_{t-1}}{q} \]
  where,
  \( h \) = closing stock price for day \( t \)
  \( q = h_{t-1} \)

- Normalization using min-max normalization
  \[ v' = \frac{v - i_{\text{min}}}{i_{\text{max}} - i_{\text{min}}} (b - a) + a \]
  where,
  \( v' \) = normalized data,
  \( v \) = data to be normalized,
  \( i_{\text{min}} \) = minimum value of input data,
  \( i_{\text{max}} \) = maximum value of input data,
  \( a \) = new minimum value,
  \( b \) = new maximum value,
2.3 UserAgent

UserAgent acts as the intermediary between the system and the user. A user can view the predictions based on his choices by interacting with UserAgent. If UserAgent receives message from NeuralNetAgent latest predictions are ready, it will request for access to the predictions. After granted permission from NeuralNetAgent it will display the predictions to the user.

3.0 Discussion of Findings

Applying the multiagent system in ANN used for stock price prediction has two benefits:
1. Reduces the workload in data preprocessing which is traditionally conducted by user before it can be utilized by the ANN.
2. Increases the accuracy of the prediction.

The multiagent system was able to reduce the workload and decreased the time and cost in data preprocessing. In a 10 days trial, the multiagent system was compared to a traditional neural network system based on time and cost required to complete all the processes; data preprocessing, training and prediction. Figure 4 shows the time recorded by the two systems.

![Fig. 4: Time Comparison between the Agent system and the ANN system](image)

The agent system was able to complete all the tasks on average one hour earlier compared to the ANN system. This is because in the ANN system, data preprocessing which is a time consuming effort has to be done manually. User initiation is also needed during the training and prediction process which further increases the time used to complete all the processes. This differs in the agent system. All the processes are executed autonomously by the agents. This ensures smooth transition from one process to the other which conclusively ensures faster completion.

Using the following equation, the cost for the two systems was computed.

\[
\text{Cost} = t \times (\text{CL} + \text{CM})
\]

where,
- \(t\) = time in hours,
- \(\text{CL}\) = labour cost per hour,
- \(\text{CM}\) = miscellaneous cost per hour.

Assuming that \(\text{CL} = 15\) and \(\text{CM} = 10\), the following graph was obtained.

![Fig. 5: Cost comparison between the Agent system and the ANN system](image)
The result shows that the agent system is approximately three times cheaper in cost compared to the ANN system. For example, for day 1, the cost recorded by the agent system was 52.01 whereas the ANN was 158.59.

The use of the multiagent system in ANN increases the accuracy of the prediction. To demonstrate this, more current closing stock price data is added into the datasets and the RMS and POCID was recorded. Four datasets from three stocks containing price data ranging from 1/10/93 – 30/9/99, 1/10/93 – 4/1/06, 1/10/93 – 27/1/06 and 1/10/93 – 28/2/06 have been tested. The effects of additional data on the performance of the ANN were observed. Figure 6 shows the result based on the RMS.

The figure shows after additional data is added, there is a decrease in the RMS. For Stock 1, when the ANN was trained with 72 months of data, the RMS was 3.4301E-7. It further decreased to 1.6737E-7 when 146 months of data were trained, 1.6698E-7 for 147 months and 6.5163E-8 for 148 months of data. The decrease is also recorded for the other two stocks. This indicates that the ANN is able to generalize better when more data is learned. When compared to previous research, the RMS recorded in our experiment was far better. The decrease in RMS is also followed by the increase in POCID as in Figure 7.

In Figure 7, there is a slight increase in POCID each time additional data is learned by the ANN. Stock 1 showed an initial POCID of 86.11% for 72 months of data. It increased by 2.78% when another two months of data was added. It increased further by 2.22% when additional one month of data was learned.

The decrease in RMS and the increase in POCID indicate that the more data learned, the better the generalization and convergence of the ANN. This also signifies that there are no significant effects of the Returns function used as was initially stated in [5]. We also predict that the RMS will further decrease and POCID will increase if new data is learnt. Therefore we can conclude that the more data learned by the ANN, the better the prediction.

**Fig. 6: The RMS of three stocks**

**Fig. 7: The POCID of three stocks**

### 4.0 Conclusion

The increase recorded in the POCID and RMS shows the agent approach significantly increase the performance of the ANN through autonomous data preprocessing and updating. The increase in the accuracy would help investors make better decisions on the best time to sell or buy stocks. In our research, adding 10 – 25 data has successfully increased the POCID and decreased the RMS. Therefore we can predict that with more data, there is a possibility that POCID could reach 100%. This demonstrates the significant roles played by agents to increase ANN capabilities thus making it worthwhile to consider this approach when developing a prediction system based on ANN.

**REFERENCES**


©FCSIT, UM 2007


BIOGRAPHY

Itaza Afiani Mohtar obtained her Master of Information Technology (Computer Science) from Universiti Kebangsaan Malaysia in 2006. Currently, she is a lecturer at the Faculty of Information Technology and Quantitative Science, Mara University of Technology Perak. Her research areas include artificial intelligence and computer science and education.

Zulaiha Ali Othman is a lecturer at the Faculty of Technology and Information Science, Universiti Kebangsaan Malaysia. Her research areas include agents and parallel processing.
AUTOMATIC FACIAL FEATURE POINTS DETECTION IN BESPECTACLED FACES

R. Saravanan, Usha Sridhar
Foundations & Futures Group
Cranes Software International Limited
Bangalore, India
Saravanan.R@cranessoftware.com, Usha.Sridhar@cranessoftware.com

Abstract

Detection of eye-centers, a set of key facial feature points, in a bespectacled face has been a challenge in automated face recognition systems because glasses and frames can significantly alter the intensity profile in the region around the eyes. The altered profile usually is caused not only by the spectacle frames, but also by reflections from glass surfaces, which are dependent on illumination conditions and the tints in the glass, leading to poor accuracy in feature point detection. The problem of robust feature point detection in the presence of spectacles in combination of pose and illumination variations has not been fully addressed. This paper describes a new algorithm to extract the eye-centers accurately in the presence of spectacles. The algorithm uses image analysis in combination with a robust rule engine derived from elements of geometrical structure of a face. The novelty of the algorithm stems from the use of local adaptive thresholding to find the lip region, and then filtering out infeasible combinations of shape objects in the eye region using the lip as a reference. Further localization of eye-pupil center in the algorithm delivers eye-center locations with high accuracy. The method handles both variations in illumination and in-depth rotations up to 20 degrees. The paper presents results from the application of the algorithm on a corpus of 939 bespectacled faces obtained from many different public databases. The results show that in 96% of the images the algorithm locates eye-centers within 3 pixels of the pupil found with manual effort.

Keywords: Automatic Feature Detection, Image Processing, Face Recognition, Adaptive Thresholding

1.0 INTRODUCTION

Face recognition systems have been an active area of research in the recent years. This specific biometric has gained prominence due to a new diversity of potential applications in the areas of secure access, surveillance, forensics, and the entertainment industry. Among the various methods [1, 2], landmark or fiducial feature point driven face recognition systems have shown more promise because they have shown the capability to handle many practical situations involving in-depth rotations, facial expression, and illumination variations in the faces under examination.

Landmark-driven face recognition systems [3] require co-ordinates of key fiducial points, such as those of the eye-centers, nose-tip, lip etc. It has been shown that the bulk of the required additional fiducial locations can be estimated, within certain in-depth rotation restrictions, if a few essential points, viz., eye-centers and lip are obtained accurately. While semi-automated methods, wherein, a human end-user can provide eye-center locations very accurately via visual inspection are straightforward to implement, automated detection of even these basic location co-ordinates is critical to the wider applicability of face recognition systems from a usability perspective.

Given a face image, accurate automated detection of facial feature points in the image is hence an important prerequisite for this class of face recognition systems. Methods reported in the literature to perform this task may be broadly classified into four groups, viz., a) Geometric Feature-based; b) Template based; c) Color segmentation based; and d) Appearance based approaches.

While there are many promising alternative methods reported in recent literature to automatically detect features [5, 6, 7, 8, 9] in frontal images with limited in-depth rotations, even with substantial illumination variations, the presence of spectacles has posed serious difficulties. Reflections from glasses result in a number of irregularly shaped objects to appear in the intensity profile around the region of the eyes. This profile may vary hugely with the illumination conditions. Also, should there be a light tint in the glasses, the eye region when visible through glasses, exhibits an intensity profile quite different from the surrounding region. These factors can produce a suffusion of shape objects that can complicate the process of eye-center detection, at times, even producing false positives. Filtering out these ‘infeasible’ shapes is often difficult since they may resemble real eyes in shape, intensity gradients and relative distances. Results related to work with images of faces with spectacles have reported poor to medium accuracy [4, 7, 8].

In this paper we describe a novel method to detect eye-centers in bespectacled faces using the lip as a reference. By detecting the lip region using adaptive thresholding, our algorithm uses knowledge of geometrical structure of the face to search and localize the eye-center in the presence of spectacles. Although the overall structure of our face feature detection methodology relies on an iterative technique that attempts to find eye-centers and lips in a parallel lock-step approach, the methodology includes a robust rule engine that makes it applicable even in the presence of spectacles.
‘best-fit’ manner such that it can handle faces both with and without spectacles, our focus in this paper is only on the method to handle bespectacled faces, using lip finding as a start point.

We use image analysis techniques combined with elements of a known geometric structure [6] in the form of a robust rule engine, on gray face images. All the images used in the experimental results in this paper are faces with spectacles. They have been extracted from a number public domain sources such as Yale, ORL, Georgia Tech, BioID, University of Bern and University of Essex.

We demonstrate that our method handles: a) Variations in illumination conditions and reflections on spectacles; b) Variations in spectacle frames ranging from thick black rims to rimless glasses c) Presence or absence of structural components like beards, mustaches; d) In-depth rotations up to 20 degrees e) Facial expression variation like smiles, opened lip, closed eyes, etc.

The novelty of our method lies in the manner in which geometric structural properties of the human face have been combined with results of image analysis using adaptive thresholding to yield a robust solution to the problem of handling variations that occur in typical input faces through pose, rotation, and artifacts of extraneous elements; given reflections from spectacle surfaces [10].

It is assumed for the purpose of this paper that an image is delivered as input, where a face is positively the primary content of the image. The organization of the paper is as follows. The second section describes our method of lip object detection from thresholding followed by a new method for eye center detection and pupil localization in the presence of variation of intensity profile due to spectacles. The next section presents an integrated view of the results as applied to several standard publicly available databases. Concluding remarks appear in the last section.

2.0 ALGORITHM TO OBTAIN FEATURE POINTS

The overall algorithm comprises two main tasks. The first is to find the lip as described in the pseudo code of Fig. 1. The algorithm uses gray scale images as input. Color images are presented to the algorithm after converting them to this format.

2.1 Lip Finding

Initially, local adaptive thresholding is applied to handle the non-distinctive intensity distribution. For the segmented image, objects are labeled with attributes for position and area. All black and white objects smaller than 25 are filtered out. The image is then finally dilated with a structuring element to improve pixel connectivity. The resultant object list is the input to the calculations mentioned in the later steps.

| Step 1 | Get Input Image(P) |
| Step 2 | Adaptive Thresholding |
| 2.1 | Image(A) = Mean (Image(P)) – (Image(P)) – C, C = 0.02 |
| 2.2 | Image(BW) = Image(A)0 |
| Step 3 | Label and Filter |
| 3.1 | Using Image(BW), Remove white and black noise: Delete objects < 25 |
| 3.2 | BW = Dilate(BW, SE), SE=Structuring Element |
| 3.3 | Find Attributes {h, w, s, cx, cy}, for objects l = 1,..N where h=height, w=width, s=area, cx, cy = x,y co-ordinates of object center |
| Step 4 | Lip_Find Rule Set |
| 4.1 | Consider the objects in lower half of the image, and away from corners |
| 4.2 | Find the object satisfying h < w AND w > 50 |
| 4.3 | From the objects satisfying the above criteria, find the object whose width is largest |
| 4.4 | Extract mw =lip width, ml = lip left, mr = lip right, and mc = lip center |

Fig.1 : Algorithm to Find Lip
A set of rules is applied to validate the lip region. The lower half of the image away from the corners of the image is considered to obtain objects satisfying a geometric heuristic for capturing the lip object, viz., an object whose width is greater than height; while the width is greater than a preset number. The object should have the largest width from the obtained set. This filtering process delivers the dimensions and center location of the lip.

The transformations applied on a sample image are depicted in Fig. 2. The first image is the result of adaptive thresholding, and the second image is produced after noise removal. Note that the third image offers many candidate lip objects, marked in red. The application of geometrical structure constraints in Step 4.3 of the algorithm yields one lip object as required.

### 2.2 Finding Eye-Centers

The second task is to use the lip found as above as a reference to locate the eye region, and then localize the eye-center. This latter task is described in pseudo code in Fig. 3.

![Fig. 2: Progressive Transformations for finding Lip](image)

In order to locate the eye region window, which is 1.5 times the width of the lip and 30 pixels in height, is selected, using the lip as a reference. This window is moved vertically in relation to the lip position, while counting the white pixels. The highest peak of a histogram generated with this data provides the mechanism to set the window or bounding box around the eye region as shown in Fig. 4.

![Fig. 3: Algorithm to find Eye Centers](image)
In the bounding box found in the region of the eyes, the vertical projection histogram, $V(x)$, delivers potential x-coordinates of the eyes as peaks from the sum of squares of inverted pixel values projected on the x-axis. The horizontal projection histogram, $HL(y)$, similarly delivers potential y-coordinates of eye-centers from peaks of the projected pixel values along the y-axis. The 2D coordinates of the eyes are obtained from a conjunction of these x-axis and y-axis peaks.

Peaks and valleys are found for the bounded eye region to localize the eye pupil locations. Peaks give the x-coordinates of the eyes. The y-positions of the eyes are obtained by performing a similar set of operations on a set of horizontal project histograms in a window around the x-positions obtained. The dimension of the window is set on the basis of the geometrical observations on pupil radius. Shown above in Fig. 5 is an example of a horizontal projection histogram.

2.3 Validation of Eye-Center Coordinates

The eye-center and lip positions are validated on the basis of a set of constraints, which they need to satisfy in terms of their relative distances in the generic face. These constraints relate to:

- The edge connecting the pupils lie on a straight line or within an angle of 20° (in our case)
- The edges connecting the two eye pupils and the center of the lip closely forms an isosceles triangle

3.0 RESULTS

Below is a discussion of the results and observations from the application of the algorithm to images in certain special situations.

Glass Reflections

One of key benefits the algorithm delivers is the ability to handle the problems posed by reflections from glasses, which cause a suffusion of irregular light patches in the eye-region. This ability stems from the manner in which the peaks and valleys in the combination of vertical and horizontal projection histograms turn neutral to the reflections. The reflections only cause more valleys, without impacting the selection of the nose bridge location and the eye locations. As an illustration, consider the series of operations as described in the algorithm on an image below where the profusion of reflections nearly occludes the eye. The algorithm produces accurate eye centers even in this case.
This property of the algorithms leads to successful capture of eye-centers and lips for a number of other similar images with no difficulty, as shown in the set of pictures below.

Fig. 7: Images with Reflection on Spectacles

**Different Frame Types**

The algorithm has also demonstrated the ability to handle different types of spectacle frames, especially those that add irregular reflection patches from the edges. As mentioned the 2D projection histogram helps to identify the frame edges through peaks, leaving the pupil localization method to fine tune pupil centers despite the frame edge intensity profile. Below is a collection of images with different types of frames that the algorithm was able to successfully process.

Fig. 8: Images with different Spectacle Frames

**Rotational Variations**

The algorithm successfully handles limited levels of in-depth rotations. Some sample images shown from the result set are shown below.

Fig. 9: Spectacle images with Rotational Variations

In order to estimate the accuracy of the algorithm, 939 images of bespectacled faces were extracted from six databases, and the eye-centers for left and right eye \((L_x^m, L_y^m, R_x^m, R_y^m)\) were obtained for each image \(j\) by manual inspection. The Euclidean distance \(d_j\) between these points from the centers obtained by the automated algorithm \((L_x^a, L_y^a, R_x^a, R_y^a)\) for the image \(j\) is calculated and used as an estimate of error. The counts as a proportion of the total number of images (939) in which the algorithm produced eye-center locations within the pupil, 1 pixel away, 2 pixels away, 3 pixels away, and so on, are plotted separately for the left and right eye (Fig.10). It is observed that in 96% of the images, the algorithm delivered eye-center locations, which were found to be within 3 pixels of pupil, as determined by manual inspection.

**4.0 CONCLUSION**

A new algorithm to detect two important face feature points in images of bespectacled faces was presented in this paper. Shape objects created by frames and reflections from glasses present additional difficulties in the detection process. The algorithm presented here uses the lip region as a reference to identify the eye region using geometric considerations. The eye centers are detected by combining information from 2D projection histograms with a set of geometric constraints to filter out irregular shape objects from the eye region, to deliver eye centers unambiguously. Using 939 images of bespectacled faces taken from public domain, the results show that in 96% of the images the algorithm locates eye-centers within 3 pixels of the pupil found with manual effort. Further work to improve the pupil localization is underway.
Fig. 10: Pixel Localization Error

REFERENCES


BIOGRAPHY

R Saravanan is a member of the technical staff at the Foundations & Futures Group of Cranes Software International Limited (CSIL), Bangalore, India. His current research interests are in the application of image processing techniques for face and object recognition. Saravanan obtained a bachelor's degree in computer science from the Madras University in 2000 and an M.Phil in computer science, specializing in image processing, in 2006. He has also worked as a researcher with the Central Electronics Engineering Research Institute, CSIR at Chennai on image analysis applications.

Usha Sridhar is Executive VP and head of the Foundations and Futures Group at Cranes Software International Limited (CSIL), Bangalore, India. She directs research efforts in the Analytix strategic business unit on areas related to learning systems and their applications to image, text and business analytics. Previously, her research interests are in pattern learning and recognition applications to unstructured data and images. Usha holds bachelor's and masters degrees in Economics and Econometrics, from the Osmania University, Hyderabad and obtained her Ph.D in Operations Research from the Indian Institute of Science, Bangalore in 1989.
DATA MINING IN RESERVOIR OPERATION AND FLOOD CONTROL USING ARTIFICIAL NEURAL NETWORKS

Ramani Bai. V\textsuperscript{1}, Woo Chaw Seng\textsuperscript{2}, Faridha Othman\textsuperscript{1} and Gopinath Ramadas\textsuperscript{1},
\textsuperscript{1}Department of Civil and Environmental Engg., Faculty of Engineering, University of Malaya, 50603 Kuala Lumpur, Malaysia. Email: ramanibai@um.edu.my
\textsuperscript{2}Dept. of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia. Email: woocs@fsktm.um.edu.my

ABSTRACT

Artificial neural networks have strong data fitting capability. In domains where explaining rules are critical, such as release of water from dam, denying loan applications etc., classical neural networks are not the tools of choice. ‘Neural Networks Cannot Explain Results’. This is the biggest criticism directed at neural networks or a challenge directed at using neural networks in water resources engineering. The main goal of this research is through processing of data (records from the past) to describe the underlying dynamics of the complex systems and predict its future. One of the solutions is data mining that is sorting through data to identify patterns and establish relationships. Using the best represented data from several previous time steps, a more complex data-driven model on artificial neural network can be built. A problem related to operation of water reservoir is selected to provide a better data representation to the network to evolve better results and continuity of the system performance compared to earlier works. This paper illustrates an artificial developmental system that is a computationally efficient technique for the automatic generation of complex Artificial Neural Network (ANN), which can better represent all the features of the data. MLFFNN models of operation of dam are developed by representing the training set to the network in following four different forms viz. 1)Historical data without optimization given as monthly (HANN), 2)Classical method of optimized results given as monthly (ANN), 3)Implicit method with optimized results given as yearly (IANN), and 4) Explicit method with optimized results given as monthly (EANN). Results have shown that neural network estimates are sensitive to sample representation, but are robust in terms of network architecture. Also the comparison to conventional statistical models, show the superiority of this approach of using ANN. In addition, this research offers an effective and reliable approach that can point out the best direction for maintaining continuity course of operation and hence with significant benefit to the decision makers on water release from the dam. The presented approach to model approximation may be used in various schemes of water resources optimization.

Keywords: Optimization, reservoir operation, multi-layer feed forward network, data mining, training, validation, training set and flood control.

1.0 INTRODUCTION

The major sources of uncertainties are natural variability and data availability. The uncertainty increases with decrease in data availabilities and increase in natural variabilities. Hydrological information is indispensable for reservoir management (Simonovic, 2000). Though the number of hydrological stations in operation worldwide as reported by WMO (1995) and domestic data centers are very impressive, their distribution is not uniform, being scarce over large areas. Further, financial constraints of governments and their services in the countries have resulted in reduction in the data collection programmes all over the world. Rule curve and storage allocation zones have been the first operating policies used to manage multi-purpose reservoirs. Operating policies associated to rule curves define the ideal storage, pool level and discharges at different times of the year for each reservoir. The rule curve is based on historical operating practice.

In operating policies based on multiple zones, the total storage volume of the reservoir is divided into several zones as in Fig.2. The inactive zone or dead storage zone represents the lower part of the reservoir that is not normally used. The buffer zone is above the inactive zone. Only essential needs are satisfied when the storage volumes are within this zone, usually as a result of a dry period. The conservation or active zone represents the volume of water that can be used to satisfy various beneficial uses including recreational and environmental needs. The flood control zone is above the conservation zone and it is reserved for flood detention especially during periods of abnormally high runoff. The spill zone is the upper portion of the pool, in which the downstream flows are at or near their maximum.
The purpose of operation and control is to influence the behavior of a system by changing an input(s) to that system according to a rule or set of rules that model how the system operates. The aim of this research is to determine the best structure of neural networks for predicting the reservoir release from a dam for flood operation and control.

Figure 1: Storage allocation zones within a Reservoir

2.0 ARTIFICIAL NEURAL NETWORKS IN WATER RESOURCES ENGINEERING

Parallel distributed processing (PDP) models are a class of neurally inspired information processing models that attempt to model information processing the way it actually takes place in the brain (Rumelhart et al. 1986). Neural networks, a new class of paradigm are modeled based on the structure of human nerve systems. These artificial networks, which learn from examples, have proved parallel processing, tolerant to noise, self-organizing and able to approximate functional relationships between cause and effect (Mays and Tung 1992). Multi-layered backpropagation algorithm proposed by Rumelhart et al. (1986) is one of the most popular neural network models. Earlier works on peak flow forecasting include conceptual and empirical models. Traditional methods are adequately summarized in hydrology texts such as Viessman et al. (1989), Ponce (1989) and McCuen and Snyder (1986). Many of these methods assume a uniform distribution in space or a coarse sub-division of the watershed area into sub-area within which the peak flow can be assumed to be uniform in a piecewise fashion. To overcome such difficulties and to model the reality, soft computing in the field of hydrological forecasting is gaining ground (See and Openshaw 1999).

Artificial Neural Networks (ANN) is one among the intelligent methods that appears to be extremely effective in handling dynamic, non-linear and noisy data especially when the underlying physical relationships are not fully understood (Zealand 1999). Ramani Bai, V (2005) proved that intra neuronal feedback network (RNN) as a powerful operation tool that is drawn on the most recent developments in artificial intelligence research with a capability of not requiring assumptions about the underlying population. The data set resulted in a better representation when an RNN model replaces the ANN model after optimization for water control problems. Typical applications of neural networks include pattern recognition, classification, forecasting, etc. Because of the inherent parallel and distributed computation, neural network is a promising technique for optimization problems especially when a real time decision is required. During the past decade, many researchers in many disciplines have extensively explored the applications of neural networks to optimization problems such as engineering and operations management (Mays, 2000; Cunha and Sousa 1999; Himmelblau 1972 ; Duren et. al. 1972). Typically, the problem of deriving operation policies for the reservoir using ANN was carried out after a deterministic optimization model in the present work. Karunanithi et al (1994) stated that several studies have been dedicated to the prediction of river flows with no exogenous inputs but with the only use of past flow observations. Ramani Bai et. al. (2007) showed that the BPLM algorithm obtains the best results more quickly than the BPM algorithm. Also, BP using Levenberg-Marquardt algorithm needs no additional executable program modules in the source code. BPM has taken epoch number of 5000. On the contrary, the number of epochs to find the optimal solution at different tests is significantly reducing (Mohan and Ramani Bai 2003). Furthermore this has taken relatively smaller running time to find the optimal solution when an ANN replaces the linear programming. The
data set resulted in a better representation when an ANN model replaces the optimization model using linear programming.

3.0 DATA MINING PROTOCOL FOR DAM OPERATION MODELING

The developed model on neural network using Levenberg Marquardt backpropagation algorithm (Ramani Bai et al. 2007) for reservoir operation and controlling was chosen for this study. The fact was, given a set of training examples with known outcomes; the idea was to find a way to correctly classify further instances that were provided later. It means mapping input to output. For example, one may wish to automatically diagnose how much amount of water to release when it had rained on that day’s morning using any classifier developed. This needs lot of information to be properly stored in the model developed and which could be understood and retrieved by the network. There are numerous ways to provide solution using neural networks. One such approach is “data mining” and the development of this field is needed. Data mining means knowledge extraction or elucidation process. It is sorting through data to find out the relationships and characteristics that exists among them. Thus some of the “black box” mapping of the ANN was refined by this way, which had led to an increased safe use by water and other industries modeling, and eventually in control. This type of study is to overcome the limitations of this algorithm such as time consuming and training tasks to some extent. The effect of the patterns of data presented to the network of dam operation is addressed in this work. The time period pool elevation, inflow and storage are taken as input and the outflow (release) of the dam is considered as output from the network.

The knowledge base necessary for the explicit network shown in Fig. 3 is trained with 30 water year flow data (1975-2004) of Batu dam located in north east of Kuala Lumpur. All the input patterns with variables such as time period, inflow, initial storage, pool elevation and release were also normalized and presented to the network. The model was developed using Neural Network toolbox in Matlab (Matlab 2001) ver 6. software. Four different multi layer feed forward neural network (MLFFNN) models for dam operation and control operation of reservoirs were developed by representing the training set to the network in following four different data forms.

3.1 Historical data without optimization given as monthly data (HANN): The knowledge extraction in data mining was performed by presenting monthly historical data (4 inputs and 1 output) to the network as input-output pattern for training without optimizing them.

3.2 Classical method of optimized results given as monthly data (ANN): In this method, the historical data were optimized using LP and the results from LP were presented as input pattern (4 inputs and 1 output) to the network.

3.3 Implicit method with optimized results given as yearly data (IANN): Data mining in this case was conducted with yearly data (36 inputs and 12 outputs) as input-output pattern for training. The time period input was removed in this case as yearly data are used.

3.4 Explicit method with optimized results given as monthly data (EANN): First, the knowledge elucidation process was done by constructing monthly models with monthly flow data (4 inputs and 1 output) so as to give the network more specific pattern it should represent on the problem of dam operation.
Fig. 2. Flow chart of development of database for EANN

Fig. 4 EANN Architecture of Flood Operation
Fig. 3 Training and Validation Results of EANN model for flood operation

In all the above cases, the initial values of the weighting factors and bias of the neural networks were chosen randomly, starting at the same point and adopting the range for the different parameters. A set of input-output responses representing a variety of flow scenarios was presented to the artificial neural network and flow chart of this simulation is shown in Fig.4. These monthly models were linked to a final model so that when a particular month’s data is presented the corresponding network will be activated to represent the pattern for that particular month’s operation. This new database, which was checked for continuity of the output from ANN, was then subjected to training by a final explicit network. This was named as explicit artificial neural network and the flow diagram of this framework.

4.0 VALIDATION OF ANN MODEL

The database was represented to the system explicitly and in four different ways as mentioned above so that all the features of the data were better understood and learned by network. The results have shown better improvement by EANN in derivation of rules as compared with optimization and other three ANN models. This final model that was trained and tested representing the explicit artificial neural network model for flood operation is shown in Fig.4. The error plots and calibration plots (target vectors and calibrated vectors) were drawn and are shown in Fig.3.

The final network designed gave $R^2$ value of 0.95 on validation. This has shown the better choice among all the four types of networks explained earlier. In addition this type of explicit modelling had taken quick convergence time. It had taken epoch number 70 with SSE 0.070 during training. This was mainly due to the explicit way of representing the data pattern to the network and careful modelling with continuity of flow consideration in the problem. This framework is checked for its performance with the independent data (2005-'06) of Batu dam situated in Klang river basin. The results have shown an improvement in better prediction of dam releases as compared to the actual historic values. The validation results are plotted as shown in Fig.3. The parameters designed for the EANN are shown in Table 1.
Table 1. Parameters designed for EANN

<table>
<thead>
<tr>
<th>No.</th>
<th>Parameters designed</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Number of hidden layers</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>Number of Neurons: 1st layer</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>Number of Neurons: 2nd layer</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>Error goal</td>
<td>0.1(normalized)</td>
</tr>
<tr>
<td>5</td>
<td>Epoch Number</td>
<td>70</td>
</tr>
</tbody>
</table>

The EANN network designed for flood operation has produced minimum epoch number with SSE of 0.07 (normalized) in short running time. Hence the data mining framework as developed in EANN model is recommended for reservoir operation and flood control.

5.0 CONCLUSIONS

Applicability of certain ANN concepts and guidelines related to pattern of data, the architecture, sampling, training and multiple function approximation were assessed in this research work. The model was developed with proper extraction of knowledge from data through data mining process. The data mining applications using ANN have revealed the following facts when compared to models developed with Optimization only. Among four different types of data representation to the network, explicit method of using optimized values (EANN) as input data for training have resulted in better operation than the other three types for reservoir operation and control models owing to its low connectivity between each layer, minimized connectivity within each layer, rapid training, efficient memory utilization. This is mainly due to reduction in complexity of the model conceptually and by making proper continuity between consequent months. Hence it is also recommended to use heuristic approach such as artificial neural network with some study of data mining rather than using optimization only.

6.0 ACKNOWLEDGMENT:

The authors acknowledge the research grant provided by the GOVERNMENT OF MALAYSIA, represented by Ministry of Science, Technology and Innovation (MOSTI) for their support and publication of the paper that has resulted in this article. The authors also thank the Institute of Research Management and consultancy (IPPP) for their fine co-ordination between the researchers and the MOSTI.

7.0 REFERENCES:


BIOGRAPHY

Dr. Ramani Bai Varadharajan, Ph.D (Environmental & Water Resources Engineering): Servicing as Lecturer in Department of Civil Engineering in University of Malaya. She has 15 years of significant service in Environmental and Water resources systems analysis, Hydrology, Water quality and treatment, Optimization techniques, Computer modeling, Hydrology and Water quality modeling using Artificial Intelligence methods as Fuzzy logic, Artificial Neural Networks and Neuro-fuzzy logic systems and Geographical Information systems.
ABSTRACT

We develop an algorithm for an agent system, namely MALESAbrain, to help educators acquire students’ learning knowledge in a problem-based learning discussion. The principles of the PBL discipline and the advantages of using PBL pedagogy are explored. In developing the PBL support tools, existing chat room and forum technologies are adopted to support the PBL discussions that are used in our study. On the downside, the authors found that without a chairperson or a tutor present in the chat room, the discussion becomes inefficient and ineffective. Consequently, our study defines the PBL functions for the chairperson and tutor agents. The authors then formulate the mathematical description for the agent which is translated into an algorithm for MALESAbrain.

Keywords: Problem-based Learning (PBL), knowledge acquisition

1.0 INTRODUCTION

Problem-based Learning (PBL) has been widely used in medical and nursing education since the 1960s and is currently a major learning strategy for a variety of programs ranging from kindergarten to university. Organizations such as the World Health Organization (WHO) and World Bank recommend the use of PBL in education curricula [1].

In the PBL pedagogy, students play the main role in driving the learning process [2]. The five-step-PBL discussion, as shown in Fig 1, illustrates how new knowledge is fed back to the main stream of the discussion through student effort and contribution: The first-step provides a well designed problem scenario to initiate the students’ interests and experiences to motivate the construction of a plan or solution to solve the problem. The students go onto the second-step where they brainstorm and generate useful ideas for the problems and solutions. Through the debates, arguments, and discussions, students in the third-step will develop their group understanding and identify important issues that they have agreed upon, namely facts. However, in the fourth-step, some of the identified facts are vague, imprecise, inaccurate and inconsistent with the discussion. They will be pointed out, as learning issues that require further study. At the end of the fifth-step, the learning issues are distributed to each student, as actions. The actions are designated as assignment for research. Students may use the library or the Internet to find the answer, or go to the laboratory to conduct an experiment, or even visit an expert to ask for advice to help solve the problem.
The student-centered PBL pedagogy has been suggested to have the following contributions to learning:

1. The discussion of the problem-scenario in the group promotes the use of communication skills and cooperative learning.
2. The brainstorming of ideas helps cross existing learning boundaries. This promotes creative learning, and knowledge integration.
3. The identification of the learning issues for research promotes active learning and critical thinking.
4. Research is needed to construct the action plans. This promotes new knowledge development.
5. The research findings are reported to the group. This promotes peer-to-peer learning.
6. By completing final products, such as a project, a research report or a presentation, it promotes independent study and self-directed working skills amidst the teamwork [3-7]

In the PBL supporting technologies, Cheng (2004) [8] analyzes the use of communication tools in the distributed Problem-based learning (dPBL) environment. They investigated synchronous and asynchronous communication tools, including ISDN video conference, WebCam video/audio, Microsoft NetMeeting® Chat and threaded discussion forum. In the PBL experiment, students are asked to rate the effectiveness of the available communication tools for each of the PBL activities. The rating is measured on a scale from 0 to 5 with 0 being not effective and 5 being very effective. The results show that the students’ rating for the synchronous chat room and the discussion forum is highly effective for clarifying and understanding the problem statement as well as the identification and prioritization of learning issues.

In the review, our study discovered that there were many researches which adopted or developed these ICT technologies to support PBL, for instance: Guzdial (1997) [9] developed Web-SMILE which integrated whiteboards and threaded discussions to organize facts, ideas and learning issues within the framework of PBL. Beer (2003) [10] created a virtual college OTIS, which included a virtual meeting room, a lecture room, a library, a help desk, a patient consultation room, etc., to allow students to communicate with both the staff and their peers in the virtual environment. Valaitis (2005) [11] successfully used the computer-mediated communication (CMC) software - FirstClass© Conference - to implement the PBL process, as suggested by McMaster University, online. However, on the down side, our study also discovered the weakness of the meeting functions in the Messengers. They are inefficient and ineffective in aiding the students to come up with their learning issues during their online discussions. This is because with the instant messengers students are not monitored hence the tendency to waste time on irrelevant chats are common [12].

Our study hypothesizes that this weakness is due to the design of the messengers. The messaging programs have neither a chairperson nor a tutor to help regulate the participants towards meaningful discussions. The chairperson and the tutor are two important roles that help provide an efficient and effective discussion environment in PBL meetings. In the messengers’ chat room, without an efficient chairperson, students cannot resolve their conflicts by themselves and end up wasting their time debating among themselves. In the issue-based forum, without an effective PBL tutor, students may unconsciously focus on the unimportant and meaningless issues which go against the discussion topic. Consequently, in this paper, we report a cognitive chat forum system MALESAbrain with two intelligent agents to assist students in PBL discussions. The first is the GDSS agent (or Group Decision Support System agent) which plays the chairperson’s role in handling the meeting tasks. The second is the tutor agent, which will guide the students, based on the meeting schedule, in PBL discussions.

2.0 GDSS AGENT

Generally speaking, the meeting chairperson has to conduct two important tasks:
1. Pointing out the important or common issues of interest to the participants in order to focus the discussion
2. Coming up with the consensus issues based on the meeting decision votes

Based on these tasks, this paper provides two definitions to help the GDSS (or chairperson) agent conduct the meeting in the MALESAbrain system. The first definition is a data structure - called AK-cell which combines knowledge-content and knowledge-weight for cooperative learning. It allows for the discussion-knowledge to become calculable in MALESAbrain’s threshold system. The knowledge pieces (or AK-cells) become mobile because of the combination of knowledge-weight and knowledge-content in the data structure. Knowledge-weight ranks AK-cell into different areas of importance, based on the participants’ judgments and the thresholds that are set up. This mechanism helps learners pay more attention to the consensus knowledge-contents for further discussion; and to think about why certain issues accumulate higher scores than others.
The second definition is the learning thresholds, which enables the GDSS agent to recognize the importance of the knowledge-weights among individual AK-cells. The intelligent learning tool focuses its attention on the significant AK-cells and organizes the knowledge structure whenever the thresholds are reached. The hierarchical structure of AK-cells will be automatically arranged and ordered according to their weights based on the judgments of thresholds and the promotion/relegation competition. Subsequently, some of the useful knowledge will be highlighted and moved to a higher level to generate more discussions and the less meaningful knowledge will be deleted. At the end of the meeting, the knowledge pieces with the highest weights will remain at the top of MALESAbrain system to become the consensus issues based upon decisions made during the meeting.

**Definition 2.** The learning threshold $\theta$ is defined as $\theta = \{\theta_kq, \theta_kr, \theta_km\}$ for decision making and for comparing the retained AK-cells $k_i$ (see definition 1), where

- $\theta_kq$ is an AK-cell qualification threshold; when $w_i \geq \theta_kq$ then $k_i$ becomes a qualified AK-cell, which is the minimum requirement to join the competition for promotion to a higher order of discussion position
- $\theta_kr$ is an AK-cell rejection threshold; when $w_i < \theta_kr$ then delete the AK-cell $k_i$
- $\theta_km$ is an AK-cell maturity threshold; when $w_i \geq \theta_km$ then the AK-cell $k_i$ has matured, and the learning group agrees that the discussion-issue $Issue_i$ has achieved consensus in the PBL discussion.

Note: Whenever any of the thresholds are reached, the GDSS agent triggers the MALESAbrain forum to reorganize the knowledge structure.

### 3.0 PBL TUTOR AGENT

The tutor agent is created to provide student guidance during the PBL meeting. The tutor agent is not constructed to answer all questions that students may ask, but instead it is installed to help direct students during their PBL discussions. The agent is installed with a scheduled PBL timetable to facilitate the meeting’s advancement with respect to the educator’s or lecturer’s expectations.

In order to help facilitate and advance the participants in achieving the learning targets, the PBL timetable is an important cognitive device for the tutor agent. It schedules the timeline of meeting achievements from posting general issues to concluding specific results. Based on the educator’s learning expectations for the meeting groups, the PBL timetable normally schedules the meeting into four quarters, as described below:

1. In the beginning of the first quarter, the educator assigns the tutor agent to encourage the students to study their issues against the learning topic setup. The tutor agent allows for students to post their own problems or solutions under the learning-subject domain.
**Definition 3.** The *PBL-meeting function* $M$ is an allocation function to assign a particular tutor agent to guide a PBL group, which is defined as

$$M(\tau) \subseteq \Omega,$$

where

- $\tau$ is a discussion-topic, and
- $\Omega$ is the learning-subject domain.

2. In order to advance the meeting discussion in the second quarter, the tutor agent advises the participants to diversify their discussions. The tutor agent directs students to read the *lower weight issues* that have been posted on the MALESAbrain forum.

3. To narrow down the students’ discussion to the specific issues in the third quarter, the tutor agent guides the participants towards studying the consensus issues. The intensified discussions assist the participants in identifying the key learning issues based on their level of the importance. The tutor agent advises students to review the *high weight knowledge pieces* to formulate some conclusions for the meeting.

4. Before the end of the *PBL meeting* in the fourth quarter, the tutor agent provides the undetermined issues list to all the participants. This list includes the *lowest weight knowledge pieces* prior to deletion. This is done to make sure that none of the issues suggested by the students are ignored during the meeting.

**Definition 4.** *due-time* $D$ is defined as function to set up the meeting due in the *PBL-meeting*.

$$D = \text{due-time(“hh:mm:ss DD/MM/YY”)}$$

The number of discussion sessions in a meeting is flexible and depending on the remaining time. The following definition provides an amended schedule setup to the tutor agent for different PBL requirements.

**Definition 5.** The *time-remaining* $\nabla$ is a countdown clock of the real number domain between 0 to 1 to measure the remaining time in the *PBL-meeting*, which is defined as

$$\nabla = \frac{(D - \text{systemTime})}{{(D - \text{beginTime})}},$$

where

- $D$ (see definition 4) is the *due-time* of a *PBL-meeting*
- $\text{systemTime}$ is the system time in the MALESAbrain system, and
- $\text{beginTime}$ is the beginning time for the *PBL-meeting* in the MALESAbrain system

4.0 **ALGORITHM**

```plaintext
1 /**********************************************************************************
2 2 The mathematical symbols used in the algorithm may found in the definitions
3 /**********************************************************************************/
```

©FCSIT, UM 2007
/* The Agent Algorithm */

SELECT CASE /* the tutor agent based on the time remaining in the PBL meeting to advise the
requesting participant (definition 5) */

CASE (\( \tau \geq 3/4 \)) /* the system meeting time is within the First Quarter at the PBL timetable */

ADVISE(M(\( \tau \subseteq \Omega \)) /* the tutor agent encourages the participants to post their discussion-
issues or response-comments based on the learning-topic within the learning-subject domain
(definition 3) */

CASE (\( \tau \geq 1/2 \)) /* the system meeting time is within the Second Quarter */

ADVISE(k_i = <\( \phi_i, w_i >\)) WHERE \( w_i \geq \theta_{kq} \) /* the GDSS agent asks the tutor agent to
encourage participants to focus their discussions on the general qualified AK-cells for
collecting a wide angle of the discussion-issues (definition 2) */

CASE (\( \tau \geq 1/4 \)) /* the system meeting time is within the Third Quarter */

ADVISE(k_i = <\( \phi_i, w_i >\)) WHERE \( w_i \geq \theta_{km} \) /* the GDSS agent asks the tutor agent to
encourage participants to focus their discussions on the specific matured AK-cells for
deciding the highest score (or consensus) issues in the meeting (definition 2) */

CASE (\( \tau \geq 0 \)) /* the system meeting time is finally within the Fourth Quarter */

ADVISE(M(\( \tau \subseteq \Omega \)) /* the GDSS agent asks the tutor agent to encourage participants to
think carefully about whether they have properly answered all the problems in the learning-
topic or not (definition 3) */

REVIEW(k_i = <\( \phi_i, w_i >\)) WHERE \( \theta_{kq} \geq w_i \geq \theta_{kr} \) /* in order to avoid ignoring any possible
breaking issues in the meeting, the GDSS agent asks the tutor agent to provide a list of low-
weight AK-cells for the final meeting review (definition 2) */

CASE (\( \tau < 0 \)) /* End Discussion */

MSN CHAT (participants, EXTEND (M (\( \tau, \theta_{kq}, \theta_{kr}, \theta_{km}, D)\))) /* after due, the participants
can discuss with their educator to decide whether to extend the PBL discussion or not.
(definition 2,3,4) */

END SELECT

Algorithm 1: A brief version of MALESAbrain agent system algorithm

5.0 DISCUSSION AND CONCLUSION

This paper reports MALESAbrain as a new knowledge acquisition methodology, which provides an agent system to
reduce the workload of the human chairperson that is mandatory for an online PBL meeting discussion. The kernel
design of MALESAbrain system is a methodology geared towards acquiring and retaining the knowledge determined
during PBL meetings. It helps participants integrate their learning knowledge by making them think critically about a
problem from different perspectives throughout the PBL discussion. Generally speaking, MALESAbrain has
contributed two notions to make the PBL meeting more effective and efficient for knowledge acquisition:

1. The first notion is the created data structure – coined as “AK-cell” (see definition 1) – for PBL meetings, which
combines knowledge-content and knowledge-weight. It allows the discussion-knowledge to become calculable
through a threshold system. The knowledge pieces become mobile because of the combination of knowledge-
weight and knowledge-content in the data structure. Knowledge-weight ranks AK-cells into different areas based
on their importance, the participants’ judgments and the thresholds set up.

2. The second notion is the autonomous decision-making mechanism – defined as the “learning threshold” (see
definition 2). It helps the GDSS agent automatically arrange and order the construction of a hierarchical knowledge
base. It helps the participants distinguish the importance of the issues for solving the meeting problem via
consensus.
MALESAbraın system been applied to IT courses for two-year PBL experiments. The future work will carried out on the library science foundation course at the University of Malaya. On the other hand, to face the new challenges of the e-Learning generation, IT education ought to integrate the original teaching methods with the PBL pedagogy in order to formulate a new and improved bi-directional pedagogy - when technologies are ready to support more problem/solution discussions online. The authors expect that the publication of this paper will motivate the cooperation between IT & PBL instructors.

REFERENCES


BIOGRAPHY

Akcell Chiang obtained his Master of Applied Mathematics from the University of Massachusetts Lowell in 1990. Currently, he is an assistant professor at Tong Nan University of Science and Technology and a PhD student at the Faculty of Computer Science and Information Technology at the University of Malaya. His research areas include problem-based learning and knowledge acquisition. He has published a number of papers related to these areas. He is also a newly invited candidate for "Who's Who in Science and Engineering” in 2007 – 2008.

Mohd Sapiyan Baba is a professor at Department of Artificial Intelligence at the University of Malaya. His area of research is in the field of Artificial Intelligence and his current research interest is AI in Education.
DETECTION OF OPTIC DISC IN COLOR FUNDUS IMAGES USING GREEN CHANNEL COMPONENT

Ahmed wasif Reza¹, C. Eswaran¹, Subhas Hati¹, Sithi Vinayakam², Ahmed Fauzi Md Sharif³

¹ Faculty of Information Technology, Multimedia University, 63100 Cyberjaya, Malaysia. Email: awreza98@yahoo.com
² Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia. Email: msithi@um.edu.my
³ Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia. Email: drafauzi@gmail.com

ABSTRACT

Detection of optic disc in fundus images forms an important step in the automatic screening process for eye diseases. This paper presents a novel technique for tracing of optic disc boundary using green channel component in the color fundus images. The brightness, color, and contrast attributes of optic disc are similar to those of other components such as exudates and cotton wools which are also present in the fundus images. The identification of these components helps in detecting the abnormal signs of eye diseases such as diabetic retinopathy and glaucoma. The proposed algorithm employs average filtering, contrast stretching transformation, thresholding, morphological gradient and boundary tracing techniques to detect the optic disc in the retinal image. The preprocessing steps are used for enhancing the image to make it better fit for the subsequent analysis. The morphological gradient is useful for determining the edge points. The tracing phase then tracks the border pixels along the top of the ridges and returns the accurate row and column coordinates of all the pixels lying on the border of the optic disc. The proposed method is tested on images of DRIVE database with different brightness and contrast conditions using a fixed threshold value. The results obtained confirm the accuracy and robustness of the proposed algorithm.

Keywords: Vision and image processing, diabetic retinopathy, optic disc, boundary tracing and detection.

1.0 INTRODUCTION

Eye diseases such as diabetic retinopathy and glaucoma affect the retina and cause blindness. This condition tends to occur in patients who have had diabetes for 5 or more years. It is reported that more than half of all newly registered blindness is caused by retinal diseases, and diabetic retinopathy is one of the main contributors [1]. Automatic screening for diabetic eye disease has been shown to be very effective in preventing loss of sight. Manual analysis and diagnosis requires a great deal of time and energy to review photographs which are obtained by fundus camera. Therefore, automated analysis and diagnosis translates to a huge amount of savings in terms of the number of retinal images that need to be manually reviewed by the medical professionals each year [2]. With this motivation in mind, this paper presents an automatic tracing technique for the optic disc in the human retina.

The optic disc is the entrance of the vessels and the optic nerve into the retina. It appears in color fundus images as a bright yellowish or white region. Its shape is more or less circular, interrupted by outgoing vessels. Sometimes the optic disc has the form of an ellipse because of a non-negligible angle between image plane and object plane. The size varies from patient to patient; its diameter lies between 40 and 60 pixels in 640×480 color photographs [3]. The disc can be seen as a landmark and it is indispensable for the detection of exudates, because the optic disc has similar attributes in terms of brightness, color, and contrast. The diameter delivers a calibration of the measurements that are done [4], and it determines approximately the localization of macula [5], the center of vision, which is of great importance as lesions in the macular region affect vision immediately.

Various methods have been reported for the detection of optic disc. In [3], optic disc is detected by means of morphological filtering techniques and watershed transformation. In [6], the optic disc is localized exploiting its high gray level variation. This approach has been shown to work well, if there are no or only a few pathologies like exudates that also appear very bright and are also well contrasted. In [7], an area threshold is used to localize the optic disc. The contours are detected by means of the Hough transform, i.e., the gradient of the image is calculated, and the best fitting circle is determined. This approach is quite time consuming and it relies on conditions about the shape of the optic disc that are not always met. Sometimes, the optic disc is even not visible entirely in the image plane, and so the shape is far from being circular or even elliptic. Also, in [8], the Hough transform is used to detect the contours of the optic disc in infrared and argon-blue image. Despite some improvements, problems have been reported in cases where the optic disc does not meet the shape condition or the contrast is very low. In [5], the optic disc is localized by backtracing the
vessels to their origin. This is certainly one of the safest ways to localize the optic disc, but it has to rely on vessel
detection. It is desirable to separate segmentation tasks in order to avoid an accumulation of segmentation errors and to
save computational time. In [9], the morphological filtering techniques and active contours are used to find the
boundary of the optic disc. In [10] an area threshold is used to localize the optic disc and watershed transformation to
find its contours. In summary, the knowledge of optic disc location is very essential for automatic analysis of retinal
fundus images. Besides the methods discussed above, there exist other techniques such as principal component analysis

The method proposed in this paper for the segmentation of optic disc uses the following three main phases: (1) image
processing and enhancement phase, (2) morphological gradient and reconstruction phase, and (3) boundary tracing
and detection phase. The first phase involves filtering, contrast stretching, and thresholding to detect optic disc as a hole in
the binary image. To select a threshold value for separating the optic disc, the filtration technique blends the objects of
the retinal image with low intensity variation into the background, while leaving the optic disc relatively unchanged.
After that, a suitable contrast stretching transformation is applied to the filtered image to make the optic disc feature
more distinguishable from the background. The aim of the next thresholding step in the first phase is to localize the
optic disc perfectly in a binary image. The intermediate phase, namely, morphological gradient and reconstruction
phase computes edge points, performs edge linking operation, and finally implements the filling hole operation along
the edges of the optic disc. The third phase, namely boundary tracing and detection phase, tracks the border pixels
along the edges of the optic disc and returns the row and column coordinates of the border pixels. This phase requires
the identification of the row and column coordinates of a pixel on the boundary as a starting point. The performance of
the proposed method is evaluated using the images of publicly available DRIVE (Digital Retinal Images for Vessel
Extraction) database [11]. The experimental results obtained with images of different brightness and contrast conditions
confirm the accuracy and robustness of the proposed method. The organization of the remaining part of this paper is as
follows: In Chapter 2, the methodology for tracing the optic disc boundary in the color fundus images is presented.
Chapter 3 presents the test results obtained using images of DRIVE database. It also presents the conclusion of this
research work.

2.0 PROPOSED METHODOLOGY

Fig.1 shows the steps involved in the proposed algorithm for the detection of optic disc.

```
Fig. 1: Steps involved in the proposed optic disc tracing algorithm
```

Each of these steps is described in detail in the following sections.

2.1 Average Filtering

The aim of this step is to enhance the image to make it better fit for the subsequent analysis stage. In the original fundus
image [see Fig. 2 (a)] the intensity variation between the optic disc and the blood vessels is relatively high and the
vessels usually have poor local contrast with respect to the background. Selecting a threshold value for isolating only
the optic disc is a tedious task and hence enhancement or preprocessing of the image for subsequent analysis becomes
indispensable. An averaging filter of size 25x35 pixels containing equal weights of one is applied to the image in order
to blend the small objects with low intensity variations into the background, while leaving the object of interest i.e. optic disc relatively unchanged. Fig. 2(b) shows the average filtered image.

2.2 Green Channel Extraction

As the green channel contains good contrast between the background and the optic disc, it is reliable to work on the green channel of the RGB color space in order to localize the optic disc. Initially, the green component is extracted from the RGB color image and then the image is cropped before applying the subsequent steps. This is done for decreasing the computational time and for focusing on the object of interest, viz the retina. After cropping, the image size is reduced from 584×565 to 512×512. Fig. 3 (a) shows the cropped green component of the filtered RGB image of Fig. 2 (b).

![Fig. 2: (a) Original RGB color image and (b) average filtered RGB image](image)

2.3 Contrast Adjustment

The image shown in Fig. 3 (a) is enhanced by applying the contrast stretching transformation to make the optic disc feature more distinguishable from the background. In this transformation, only the darker regions have their intensity values enhanced slightly while the brighter regions of the image remain more or less unchanged. The resulting image is shown in Fig. 3 (b) using the contrast stretching transformation plotted in Fig. 3 (c).

![Fig. 3: (a) Cropped green component, (b) image after contrast adjustment, and (c) contrast stretching transformation](image)
2.4 Thresholding Operation

A threshold value $T$ is chosen to obtain a binary image that isolates the optic disc from the background. This is done using a trial and error procedure on several images. This thresholding operation results in two groups of pixels, $G_1(x, y)$ and $G_2(x, y)$ as illustrated in equation (1). The binary image $G(x, y)$ obtained after applying equation (1) with $T = 245$ is shown in Fig. 4 (a).

$$G(x, y) = \begin{cases} 1 & \text{if } G_1(x, y) \geq T \\ 0 & \text{if } G_2(x, y) < T \end{cases} \quad (1)$$

Fig. 4: (a) Binary image showing the optic disc after thresholding and (b) edge detection using morphological gradient

Fig. 5: (a) Filling holes using morphological reconstruction and (b) resultant optic disc with boundary tracing
2.5 Morphological Gradient and Reconstruction

Morphological gradient [13] is applied to the thresholded image obtained from the previous phase to detect the optic disc as edges. After applying the dilation process [13] on the edges, we get the image shown in Fig. 4 (b). This step ensures accurate pixels detection along the boundary. In the next step, morphological reconstruction [13], namely, filling holes is applied to fill the holes along the border of the optic disc in the binary image. The resultant image is shown in Figure 5 (a).

2.6 Boundary Tracing and Detection

This phase traces the border pixels along the edges of the optic disc and returns the row and column coordinates of the border pixels. This technique uses the row and column coordinates of a border pixel as the starting point for the boundary tracing. Here, the nonzero pixels in the binary image belong to the optic disc and pixels with values 0 (zero) constitute the background. The obtained optic disc has a filled border and the selected starting point is on a thin part of the optic disc. Initial row coordinate is calculated by dividing y-dimension (it is seen that optic disc is placed approximately in the middle position along the row). The procedure used in the phase locates all the nonzero (one) elements of the binary array which match with the row coordinate and assign them to column coordinate. After specifying the row and column coordinates of the starting point, this phase thus stores the row and column coordinates of all the pixels of the border of the optic disc in a cell array and these values are returned as linear indices [13]. This tracing technique traces the outside border of the object or the inside border of the hole depending on the direction specified. For filled objects as in the proposed approach, the direction parameter (north, south, east, and west) is not important and this is the justification for filling up of holes using morphological reconstruction in the previous phase. The optic disc obtained after the boundary tracing operation is plotted using the row and column coordinates and it is superimposed on the enhanced green channel image. The resultant image at the end of boundary tracing and detection phase is shown in Fig. 5(b).

3.0 RESULTS AND CONCLUSIONS

The proposed method described in the previous section is tested on images of publicly available DRIVE database. The DRIVE database contains 20 color images of the retina with 565×584 pixels and 8 bits per color channel. The software chosen to carry out this experiment is MATLAB [13], which is a high-performance language for technical computing. The software module that is used extensively in this project is the Image Processing Toolbox. Fig. 6 shows the experimental results obtained using the proposed method on some of the test images. An important advantage of the proposed algorithm is to trace the boundary of the optic disc sharply. Further it may be noted that the successful results are obtained for different test images with varying brightness and contrast conditions using the same threshold value, namely, $T=245$. The proposed algorithm has given satisfactory results for 18 images in tracing the boundary of the optic disc with $T=245$. Out of the remaining two test images with a changed threshold value of $T=250$, the optic disc could be detected successfully. For the last image, it was not possible to select a suitable threshold value due to very low contrast between the optic disc and the background.

The results presented in this paper for optic disc detection will be very useful in the development of a computer based automatic screening system for the early diagnosis of eye diseases such as diabetic retinopathy and glaucoma.
ACKNOWLEDGMENT

This research work is supported by E-Science Project (No: 01-02-01-SF0025) sponsored by Ministry of Science, Technology and Innovation (MOSTI).

REFERENCES


DISTRIBUTED ALGORITHM FOR CLUSTERING LARGE DATASETS

Radhika M. Pai¹, Ananthanarayana V.S.²
¹Dept. of Computer Science &Engg., Manipal Institute of Technology, Manipal-576104, INDIA
radhikampai@rediffmail.com
²Dept. of Information Technology, National Institute of Technology, Karnataka, Surathkal-575025, INDIA
ananthvs1967@gmail.com

ABSTRACT
In this paper, we propose a novel scheme called Prefix-Suffix trees which is a compact and complete representation of the database that requires only a single database scan for its construction. We propose a clustering algorithm based on this scheme and a parallel implementation of the algorithm based on the message passing model. The proposed algorithm exploits the inherent parallelism in the algorithm. The algorithm is implemented on a cluster computer with 3 nodes running LINUX. We compare the execution times experimentally and show that the parallel algorithm is efficient than its sequential counterpart.

Keywords: Distributed algorithm, Data mining, Clustering, Prefix-Suffix Trees, PC-Trees.

1.0 INTRODUCTION
Clustering is an exploratory data analysis task that requires very huge amounts of data [1, 2]. This huge amount of data cannot be stored in main memory. Either the data has to be stored using abstractions which compacts the data size or the data has to be stored on secondary storage and the algorithm performs the disk scans to read the data. The first method requires less space and also consumes less time as it doesn’t access the secondary storage device. But the time consumed depends on the number of disk scans required for generating the abstractions. One scan is the best. In this paper we propose a novel scheme called Prefix-Suffix Trees which is an abstraction of the database which can be constructed from the database in a single scan. This scheme is a compact representation of the database. We compare our scheme with an existing abstraction called Pattern-Count(PC)-tree[7] and show that this scheme is more compact than the PC-tree. We further propose a clustering algorithm based on this scheme and show that it gives better accuracy than the PC-tree based algorithm. We also propose a parallel algorithm based on Message Passing Model [9,10] to reduce the execution time of the algorithm and which exploits the inherent parallelism present in the algorithm. We compare the experimental execution times of both the parallel and the sequential algorithms and show that the execution time of the parallel algorithm is less than its sequential counterpart. The paper is organized as follows. In section 2, the Prefix-Suffix trees and the clustering algorithm based on this is explained. Section 3 gives the parallel model and the parallel algorithm. Section 4 describes the experimental setup and the datasets used. It also gives a comparison of the parallel and the sequential algorithms run on the same machine and the results. Section 5 gives the conclusion.

2.0 PREFIX-SUFFIX TREES
The Prefix-Suffix Trees based scheme which we propose is an abstract and compact representation of the transaction database. It is also a complete representation, order independent and incremental. The Prefix-Suffix Trees are made up of nodes forming trees. Each node consists of four fields. They are ‘Feature’ specifies the feature value of a pattern. The feature field of the last node indicates the transaction-id of the transaction which helps in retrieving the original transactions. ‘Count’ The count value specifies the number of patterns represented by a portion of the path reaching this node. ‘Child-pointer’ represents the pointer to the following path. 'Sibling-pointer’ points to the node which indicates the subsequent other paths from the node under consideration. Fig. 1 shows the node structure of Prefix-Suffix trees.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Count</th>
<th>Sibling-pointer</th>
<th>Child-pointer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1 Node structure of Prefix-Suffix Trees

2.1 Construction of the Prefix-Suffix Trees
The algorithm for the construction of the Prefix-Suffix Trees is as follows.
Let Tᵢ be the transaction database.
Partition the transaction database Tᵢ into 2 equal parts.
Let the 2 parts be $T_1$ and $T_2$ respectively.
For $T_1$, construct the Tree called the Prefix Tree as follows.
Let the root of the Prefix-tree be $TR_1$.
For each pattern, $t_i \in T_1$
Let $m_i$ be the set of positions of non-zero values in $t_i$.
If no sub-pattern starting from $TR_1$ exists corresponding to $m_i$,
THEN
Create a new branch with nodes having 'Feature' fields as values of $m_i$ and 'Count' fields with values set to 1.
ELSE
Put values of $m_i$ in an existing branch $e_b$ by incrementing the corresponding 'count' field values of $m_i$ by appending additional nodes with 'count' field values set to 1 to the branch $e_b$.
For $T_2$, reverse $T_2$ to get $T_{2r}$.
Construct the tree for $T_{2r}$ as described earlier which is called the Suffix Tree.
Let $TR_1$ be the root of Prefix Tree corresponding to $T_1$ and let $TR_2R$ be the root of the Suffix Tree corresponding to $T_{2r}$

An example-Consider the following set of transactions as shown in Table1.

<table>
<thead>
<tr>
<th>TransNo</th>
<th>Features</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,2,3,4,5,8,9,10,11,12,14,15,16</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1,2,3,4,7,10,11,12,14,15,16</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2,3,4,5,6,12,14,15,16</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2,4,5,7,9,12,13,14</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>2,4,5,6,8,12,13,14</td>
<td>3</td>
</tr>
</tbody>
</table>

The first column gives the transaction number, the second column gives the set of features and the last column gives the label. The partitioned set of transactions Part1 and Part2 are given in Table2, Table3 and the set of reversed transactions of part2 are given in Table4 respectively. The Prefix tree, The Suffix-tree and the PC-tree for the set of transactions in Table 1 is given in Fig.2(a), Fig.2(b) and Fig.2(c) respectively. In the figures, the nodes are indicated by circles, the right arrow is the child pointer, the downward pointer is the sibling pointer. The first number inside the circle is the feature value and the number after the colon is the count. The last node in all branches is the label node.

---

**Table 1: Sample set of Transactions**

<table>
<thead>
<tr>
<th>TransNo</th>
<th>Features</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,2,3,4,5,8,9,10,11,12,14,15,16</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1,2,3,4,7,10,11,12,14,15,16</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2,3,4,5,6,12,14,15,16</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2,4,5,7,9,12,13,14</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>2,4,5,6,8,12,13,14</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TransNo</th>
<th>Features</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9,10,11,12,14,15,16</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>10,11,12,14,15,16</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>12,14,15,16</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>9,12,13,14</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>12,13,14</td>
<td>3</td>
</tr>
</tbody>
</table>

**Table 2: Set of Transactions in Part1**

<table>
<thead>
<tr>
<th>TransNo</th>
<th>Features</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,2,3,4,5,8</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1,2,3,4</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2,3,4,5,6</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2,4,5,7,3</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>2,4,5,6,8</td>
<td>3</td>
</tr>
</tbody>
</table>

**Table 3: Set of Transactions in Part2**

<table>
<thead>
<tr>
<th>TransNo</th>
<th>Features</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16,15,14,12,11,10,9</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>16,15,14,12,11,10</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>16,15,14,12</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>14,13,12,9</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>14,13,12</td>
<td>3</td>
</tr>
</tbody>
</table>

**Table 4: Reversed Part2 set of Transactions**

<table>
<thead>
<tr>
<th>TransNo</th>
<th>Features</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16,15,14,12,11,10,9</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>16,15,14,12,11,10</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>16,15,14,12</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>14,13,12,9</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>14,13,12</td>
<td>3</td>
</tr>
</tbody>
</table>
2.2 Clustering Algorithm

In order to cluster the test pattern, the test pattern is also partitioned into 2-partitions using the same partitioning criteria as used for the training patterns.

Let \( c \) be the number of classes, \( k \), the number of nearest neighbours.

The algorithm proceeds as follows.

For each branch \( b_l \) in the Prefix-Tree \( TR_1 \).

Find the matches between the test pattern and the branch \( b_l \), let it be \( C_{1}^{l} \).

Find \( k \) largest counts in decreasing order. Let them be \( C_{1}^{l}, C_{2}^{l}, \ldots, C_{k}^{l} \) and the corresponding labels be \( O_{1}^{l}, O_{2}^{l}, \ldots, O_{k}^{l} \).

Similarly, for each branch \( b_m \) in Suffix-Tree \( TR_2 \).

Find the matches between the test pattern and the branch \( b_m \), let it be \( C_{2}^{m} \).

Find \( k \) largest counts in decreasing order. Let them be \( C_{2}^{m}, C_{2}^{m}, \ldots, C_{k}^{m} \) and the corresponding labels be \( O_{2}^{m}, O_{2}^{m}, \ldots, O_{k}^{m} \).

For \( i = 1 \) to \( k \)

For \( j = 1 \) to \( k \)

Find \( C_{p} = C_{i} + C_{j} \) if \( O_{i} = O_{j} \) where \( 0 \leq p \leq k - 1 \).

Find \( k \) largest counts in decreasing order among all \( C_{p} \) where \( 0 \leq p \leq k \).

Compute the weight \( W_{p} = 1 - (C_{k} - C_{p})/(C_{p} - C_{1}) \)

For \( n = 1 \) to \( c \)

\[ \text{Sum}_{n} = \sum_{i=1}^{p}[W_{m}] \text{ where } (O_{m} = O_{n}) \]

Output (label = \( O_{x} \)) for which \( \text{Sum}_{x} \) is maximum for \( x \in 1, 2, \ldots, c \)

3.0 DISTRIBUTED ALGORITHM BASED ON PREFIX-SUFFIX TREES

3.1 Parallel model

Our parallel algorithm design is based on the Single Program Multiple Data (SPMD) model using Message Passing interface (MPI) [9, 10] which is currently the most prevalent model for computing on distributed memory multiprocessors (DMMs). We assume a shared nothing machine with \( P \) processors each with a local memory which are connected by a high speed gigabit switch. The communication between these processors is captured by MPI, the message passing interface, which is standardized, portable and widely available message passing system designed by a
group of researchers from academia and industry. From a programmer’s perspective, parallel computing using MPI appears as follows. The programmer writes a single program in C or C++ or FORTRAN 77, compiles it and links it using the MPI library. The resulting object code is loaded in the local memory of every processor taking part in the computation; thus creating P "parallel" processes. Each process is assigned a unique identifier between 0 and P-1. Depending on its processor identifier, each process may follow a distinct execution path through the same code. These processes may communicate with each other by calling appropriate routines in the MPI library which encapsulates the details of communications between various processors.

<table>
<thead>
<tr>
<th>ROUTINES</th>
<th>MEANING</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPI_Comm_size()</td>
<td>Returns the number of processes</td>
</tr>
<tr>
<td>MPI_Comm_rank()</td>
<td>Returns the process identifier for the calling process</td>
</tr>
<tr>
<td>MPI_Send()</td>
<td>Sends the data to the numbered process</td>
</tr>
<tr>
<td>MPI_Recv()</td>
<td>Receives the data from the numbered process</td>
</tr>
<tr>
<td>MPI_Finalize()</td>
<td>Finalizes the routine</td>
</tr>
</tbody>
</table>

3.2 Parallel Algorithm

Table 5 gives a glossary of various routines which we use in our parallel version of the algorithm which is given below. Let T be the transaction database.

Partition the transaction database T into 2 equal parts.

Let the 2 parts be Tr1 and Tr2 respectively.

Let c be the number of classes, k, the number of nearest neighbours.

Call MPI_Comm_size()

Get the process id of the process by calling µ = MPI_Comm_rank()

if µ = 0

For each test pattern

Receive the k nearest neighbours of the Prefix-tree for the test pattern from process 1

Receive the k nearest neighbours of the Suffix-tree for the test pattern from process 2

For i= 1 to k

For j= 1 to k

Find Cp = Ci + Cj if Oi == Oj where 0 <= p <= k - 1.

Find k largest counts in decreasing order among all Cp where 0 <= p <= k.

Compute the weight , Wp = 1 - (Ck - Cp)/(Cp - C1)

For n = 1 to c

Sumn = \sum_{m} Wm where (O_m = n)

Output (label = Ox) for which Sumx is maximum for x \in 1, 2, . . . , c

if µ = 1

For Tr1 construct the tree as follows which is called Prefix Tree.

Let the root of the Prefix-tree be TR1.

For each pattern, ti \in Tr1

Let mi be the set of positions of non-zero values in ti.

If no sub-pattern starting from TR1 exists corresponding to mi,

THEN

Create a new branch with nodes having 'Feature' fields as values of mi and 'Count' fields with values set to 1.

ELSE

Put values of mi in an existing branch eb by incrementing the corresponding 'count' field values of mi by appending additional nodes with 'count' field values set to 1 to the branch eb.

For each part of the test pattern

For each branch bj in the Prefix-Tree TR1.

Find the matches between the test pattern and the branch bj, let it be Ci.

Find k largest counts in decreasing order. Let them be C_i^1, C_i^2, ... C_i^k and the corresponding labels be O_i^1, O_i^2, ... O_i^k

Send the k largest counts table to process 0

if µ = 2

For Tr2, reverse Tr2 to get Tr2r.
For Tr2r construct the tree as follows which is called Suffix Tree.  
Let the root of the Suffix-tree be TR2R.  
For each pattern, ti 2 Tr1  
Let mi be the set of positions of non-zero values in ti.  
If no sub-pattern starting from TR1 exists corresponding to m,  
THEN  
Create a new branch with nodes having 'Feature' fields as values of m and 'Count' fields with values set to 1.  
ELSE  
Put values of m in an existing branch eb by incrementing the corresponding 'count' field values of m by appending additional nodes with 'count' field values set to 1 to the branch eb.  
For each test pattern,  
for each branch bm in Suffix-Tree TR2R,  
Find the matches between the test pattern and the branch bm.  
Let it be C21.  
Find k largest counts in decreasing order.  
Let them be C21, C22, … C2k and the corresponding labels be O21, O22, … O2k.  
Send the k largest counts to process 0.  

The algorithm is implemented on a cluster computer with three nodes, each with a Xeon processor having 512 MB of RAM with a clock frequency of 2.66 GHz. running LINUX. The processes are numbered 0, 1 and 2. The process 1 creates the Prefix Tree and for each prefix part of the test pattern finds the k-nearest neighbours and sends it to process 0. Similarly process 2 creates the Suffix Tree and for each suffix part of the test pattern sends the k-nearest neighbours to the process 0. The process 0 receives the neighbours from process 1 and process 2 and processes it to output the label of the unknown test pattern. Meanwhile both processes 1 and 2 will be processing the next pattern. In this manner, the parallelism is achieved which cuts down the execution time.

4.0 EXPERIMENTS AND RESULTS

We have used several datasets to test our algorithm. The experiments were carried out on a 3-node cluster computer, each with a Xeon processor, 512 MB RAM using the message passing model. The programs were coded in C++ with MPI calls embedded in the program.

4.1 Dataset 1,2,3: OCR data

This is a handwritten digit dataset[7,8] . There are 6670 patterns in the training set, 3333 patterns in the test set and 10 classes. Each class has approximately 670 training patterns and 333 test patterns. We conducted experiments separately with 2000(200 patterns from each class), 4000(400 patterns from each class) and 6670 training patterns(667 patterns from each class) which we call as Dataset 1, 2, and 3 respectively and compare our results for accuracy and space, the results of which are given in Fig. 3A and Fig. 3B respectively. We also compare our Prefix-Suffix Trees based Parallel Algorithm with the same algorithm executed sequentially on the same machine. The comparison based on execution times are summarized in Fig.3C.

4.2 Dataset4: USPS data

This is a handwritten digit dataset which is a collection of handwritten digits scanned from the U.S. postal services [5]. There are 7291 patterns in the training set and 2007 patterns in the testing set and 10 classes. The Fig. 3B shows the memory size requirements for the original database, PC-tree and the Prefix-Suffix Trees. The results based on accuracy are summarized in Fig. 3A. The time for execution of the algorithm is compared with its sequential counterpart and is given in Fig. 3C.

4.3 Dataset5: MNIST data

This is a data which is a mixture of the NIST(National Institute of standards and technology) special database 3 and 1 [4]. This is collection of handwritten digits written by census bureau employees and high school students. This is a large data set having 60000 patterns in the training set and 10000 patterns in the test set and 10 classes. The comparison of accuracy and space with respect to the PC-Tree based algorithm is given in Fig. 3A and Fig. 3B respectively. The results of the comparison of sequential algorithm and its parallel version are shown in Fig. 3C. From the results we see that the time consumed by our Parallel algorithm is less compared to its sequential counterpart.
5.0 CONCLUSION

In this paper, a novel scheme called Prefix-Suffix Trees for compact storage of patterns is proposed and the clustering algorithm based on this scheme is described. The scheme is compact compared to an abstraction called PC-tree and the algorithm is found to give good results in terms of accuracy. This has been established by our experiments. We have also implemented the parallel version of the same algorithm and the execution times of the parallel and sequential algorithms are compared by executing on the same machine.

References


DNA SEQUENCE DATABASE CLASSIFICATION AND REDUCTION: ROUGH SETS THEORY APPROACH

M Nordin A Rahman1, M Yazid M Saman2, Aziz Ahmad2 and A Osman M Tap1
1Faculty of Informatics
University of Darul Iman Malaysia, 21300 K Terengganu, Malaysia.
Email: mohdnabd@udm.edu.my
2Faculty of Science and Technology
University of Malaysia Terengganu, 21030 K Terengganu, Malaysia.
Email: yazid@umt.edu.my, aaziz@umt.edu.my
Email: amantap@yahoo.com

ABSTRACT
Data classification is a vital task in large scale data mining application. DNA sequences are the basis of life and they encode all the necessary information needed to reproduce life. The size of public DNA sequence databases are growing doubling every year. This situation makes automatic classification and reduction of DNA sequences has become important for effective sequence similarity search problem. A challenge in DNA sequence similarity search is that the sequence record structure does not have any attribute that can be used for implementing classification process. In this paper, by means of filtering process an automaton based exact string matching is employed to generate a special attribute used for DNA sequence database classification and reduction. Rough sets theory provides an indiscernibility relation technique which can be used to classify and reduct the database based on some definition of 'equivalence'. The generated attribute is used for constructing indiscernibility relation among sequences. With computational implementation, the experiments are executed to investigate the effectiveness of rough sets theory on generating DNA sequence database classification and reduction. Moreover, the experiments will demonstrate that the DNA sequence similarity search performance is significantly improved by using this approach.

Keywords: DNA sequence, filtering, rough sets theory, database classification, optimal alignment

1.0 INTRODUCTION
The rough sets theory has been introduced by [8]. It is a mathematical tool that can be used for processing and analyzing of inexact, uncertain, and vague datasets [1][3][6][7][14]. Practically, rough set theory has been applied to the number of application domains such as medical diagnosis, engineering reliability, expert systems, empirical study of materials data, machine diagnosis, business failure prediction, activity-based travel modeling, travel demand analysis, solving linear programming and data mining [11]. Classification of objects in the databases or information systems sources based on rough set theory has been done in many applications [1][11][14]. The indiscernibility relation technique in rough set theory has the ability to search objects in a multi-dimensional data space and determine the relationship between objects and its attributes. Therefore, rough sets theory method is useful for exploring DNA sequence patterns and ease to construct genomic sequence classification. Classification will help to improve DNA sequence similarity search performance.

One of the most vital applications in studying of molecular biology is to construct the evolutionary history of genes. Since discovery activities in new genomic projects have rapidly active, the number of DNA sequences are expected to be increased dramatically every year. Therefore, the need of sequence similarity search tool that performs efficiently for producing optimal result is highly demanded. An automatic classification and reduction of DNA sequences in a database has become important for effective sequence similarity search problem. Clearly that a reduct set of a DNA sequence database can extract the sequence features that affect the similarity search result. The ‘redundant’ DNA sequences can be eliminated from such class of sequences. However, the sequence record structure in a database does not have any specific attribute that can be used for implementing classification process. The important attributes associated with each registered sequence can be classified as sequence identification, definition, the number of each base, authors name, date submitted, keywords and sequence. Word stemming, removing common stop words or stripping least useful keywords cannot be applied in sequence homology searching problem. There are no specific attributes that can be used by biologist to differentiate sequences precisely based on its homology functions.

The growing advancements in engineering computational biology requires for better sequence searching algorithms to deal with the complex and extensive sequence data. A new practical filtering mechanism has been developed to discard irrelevant database sequence from being executed for optimal alignment. The generalization of this filtering process is described in Section 1.1. In order to improve the developed filtering capabilities, the rough sets theory is applied in the architecture. Rough sets theory is a technique to manage the uncertainty data. It provides equivalence relations and
reduction concepts that can reduce the size of datasets without losing essential classificatory information. The generated equivalence relations give an important implication to the granulation of the sequence database. Based on this equivalence relation, the lower and upper approximations of subsets of DNA sequence database can be defined. To evaluate the effectiveness of the proposed sequence searching model, we conduct experiments to compare the performances of developed technique with and without rough sets object classification and reduction. Generally the points of using rough sets theory in DNA sequence similarity search are as follows:

- Clustering and minimizing the size of DNA sequence databases before an optimal alignment is executed and;
- Synthesizing new symbolic rules in order to select DNA sequences in database to be executed for comparisons.

1.1 Filtering DNA Sequences: Automaton based Exact String Matching Approach

Generally, the basic problem domain to be solved can be define as: Let $T = \{x_1, x_2, ..., x_n\}$ be a DNA sequence database and $q$ is a query sequence; let $\theta$ be a fixed score threshold and $F$ be an alignment scoring function. By means of an optimal local alignment, $A$, of $(q, x_i)$, find $R \subseteq T$ where $\forall r_i \in R$ has score $F(A) \geq \theta$. Optimal local alignment can be done by utilizing Smith-Waterman algorithm [12] which running in quadratic time and space complexity. By retaining the optimal result generated from Smith-Waterman algorithm, this paper suggests that the DNA sequence database needs to be filtered and clustered. Filtering can reduce the size of database and simplify the searching process. From the same point of view, clustering database will partition database and therefore efficiency and effectiveness of retrieval can be improved.

The two popular heuristics methods for sequence similarity search: BLAST[4] and FASTA[10] are based on filtering mechanism. The methods try to identify small exact matches of fixed size and extend these identical segments according to a threshold value. Those methods intend to remove the expected non-homology segments in both query and targeted sequences before alignment process is executed. Instead of removing the expected non-homology segments in the compared sequences, this paper presents the filtering technique for discarding the expected irrelevant database sequences to the query. The filtering process generates matching score value and only the sequences that have a high score value are selected for rigorous optimal alignment process. The developed filtering technique used an automaton based exact string matching proposed by [2][5]. The automaton based exact string matching is also known as Aho-Corasick algorithm. The algorithm emphasizes for locating multiple patterns in a given text. The algorithm is very fast in practical with the time complexity $\Theta(n)$ for pre-processing and $\Theta(m+k)$ for searching phase, where $n$ is total length of all patterns, $m$ is targeted text length and $k$ is the number of patterns occurrences in the targeted text. The Aho-Corasick algorithm is developed based on three functions: goto function ($g$), failure function ($f$) and output function ($h$).

Figure 1 shows the states series in the proposed filtering technique. A set of patterns with a fix length is randomly generated from the query sequence. All those patterns will be inserted into automaton goto function. Based on this goto function, failure function is constructed in order to speed up matching process. Failure function creates a fundamental form to continue the matching process when mismatch comparison is occurred. DNA sequences from database are scanned to this automaton for locating the query patterns. Using BLOSUM62 scoring matrix, total exact matching score is calculated. The scanning process is iterated until all sequences in database are processed. Based on total exact matching score, the sequences are ranked by using Quick-Sort algorithm. Precise is the most general measurement of data retrieval performance. Therefore, the quality of the proposed filtering technique is evaluated by precise measurement and can be formulated as $\alpha = \frac{\text{relevance sequence}}{\text{returned sequence}} \times 100\%$. 

©FCSIT, UM 2007
2.0 ROUGH SETS THEORY

Rough sets theory has been proposed by Polish mathematician, Professor Pawlak for knowledge discovery, machine learning and decision analysis [8][9]. Rough sets theory has attracted much attention in modeling uncertainty and vague datasets. An upper and a lower approximation of a set, the approximation space and models of sets are the fundamental concepts in rough sets theory [14]. General elements involve in rough sets theory can be described as follows:

- The universe is a non-empty finite set of objects named training set, \( U = \{x_1, x_2, \ldots, x_n\} \);
- The attributes is a non-empty finite set of attributes, \( A = \{a_1, a_2, \ldots, a_k\} \);
- An information system (also called decision table) is a pair of the universe and attributes, \( IS = <U, A> \);
- The indiscernibility relation (also known as equivalence class) defines a partition in \( U \). Let \( a \in A \) and \( B \subseteq A \), the indiscernibility relation is defined as, \( R(B) = \{(x, y) \in U \times U : \text{for all } a \in B, a(x) = a(y)\} \);
- Let \( C \) be a set of condition attributes and \( R(C) \) be an indiscernibility relation on \( U \), an approximate spaces could be define as a pair, \( AS = <U, R(C)> \);
- \( [x]_B \) denotes the equivalence class of \( B \) containing \( x \), for any element \( x \) of \( U \);
- Based on a singleton \( x \), for a given \( B \subseteq A \) and \( X \subseteq U \), the lower approximation (\( B X \)) of the set \( X \) in IS and the upper approximation of the set \( X \) in IS (\( \overline{B}X \)) are defined as follows:
  - \( B X = \{x \in U : [x]_B \subseteq X\} \);
  - \( \overline{B}X = \{x \in U : [x]_B \cap X \neq \varnothing\} \);
- For a given \( B \subseteq A \) and \( X \subseteq U \), the boundary of \( X \) in IS can be defined as, \( \text{BND}(X) = \overline{B}X - B X \). \( \text{BND}(X) \) consists of objects that do not certainly belong to \( X \) on the basis of \( A \).

The attributes in \( A \) can be further classified into two disjoint subsets, condition attributes (C) and decision attributes (D) such that \( A = C \cup D \) and \( C \cap D = \varnothing \). The decision attribute can have some values though quite often it is binary [6]. Let \( B \subseteq A \), \( U/R(B) \) denotes a family of all equivalence classes of the relation \( R(B) \), called elementary sets. The elements of \( B X \) are all and only those objects \( x \in U \) which belong to the equivalence class generated by the indiscernibility relation contained in \( X \). Meanwhile, the elements of \( \overline{B}X \) are all those objects \( x \in U \) which belong to the equivalence classes generated by the indiscernibility relation containing at least on object \( x \) belonging to \( X \). The \( \text{BND}(X) \) indicates that objects in IS are inconsistent or vague. If upper and lower approximations of \( X \) are equal then \( X \) is an ordinary set. Clearly that, rough sets theory mainly resolves to the problem how \( X \subseteq U \) can be covered with a set of equivalence classes according to indiscernibility relation.
3.0 DNA SEQUENCE DATASET CLASSIFICATION AND REDUCTION

In recent years mathematical methods have played increasingly important role in computational biology. In particular, analyzing the enormous amount of data involved in DNA or protein sequences requires for advance numerical techniques. In the proposed model, an indiscernibility relation presented in rough sets theory is used for calculating the classification and reduction rules of processed DNA sequence database. Rough sets theory does not impose any static statistical parameters, therefore assumption requirements can be minimized and allowing the data to represent itself [1].

A unique attribute generated from automaton filtering process is used as a parameter in constructing indiscernibility relation among the sequences. Based on this indiscernibility relation, DNA sequence database clustering can be made. From rough sets theory point of view, DNA sequences are considered indiscernible if they share the same value for such defined attribute. By applying indiscernibility relation method, the DNA sequence database will be partitioned into several parts for improving sequence similarity searching performance. Once classification of sequences is completed, reduction process will be executed. The generated reduct set of a concept would keep essential information of the original class. Only sequences in the reduct set will consider for optimal local alignment process. The optimal local alignment process is implemented using Smith-Waterman algorithm [12] which is based on dynamic programming methods. The inspiration of this algorithm is to build up the result by using previous results for smaller subsequences. Rough sets theory is applied after ranking state in filtering process. Figure 2 illustrates the methodology for applying rough sets theory in DNA sequence database classification and reduction. There are three main rough sets operations involve in the methodology: data classification, defining for set approximation and data reduction. The following subsections will discussed the details of those operations.

Fig. 2: Rough sets theory operations in DNA sequence database classification and reduction

3.1 DNA Sequence Database Classification

All objects in the universe set are associated with some knowledge. In this case, the universe is the processed DNA sequence database and the generated exact matching score is considered as a unique knowledge to each object in the universe. This processed database expresses the first biologist perception to the targeted sequences objectively and the classification rules may induce from their related attributes. The model uses the indiscernibility relation to group the sequences based on some definition of equivalence which relates to the problem domain application. All sequences in database with similar knowledge are indiscernible and construct the clusters, which can be considered as elementary granules. Let say, T is a set of sequences to be performed for optimal alignment (T is also considered as the universe) and $B_i \subseteq T$ where $i = 1, 2, \ldots, n$ and $n$ is the number of equivalence classes to be constructed. Therefore, we can say that each $B_i$ may form a partition on $T$ where for each $B_i \neq \{\emptyset\}$ and can be denoted as, \(\text{CLASS} = \{B_1, B_2, \ldots, B_n\}\). From the perspective of information systems, $B_i$ is also called as knowledge base. Those generated equivalence classes correspond to the elementary classes of rough sets theory. Rough sets theory determines a degree of sequence’s dependency and their significance. The analysis of uncertain relationship among sequences acquired from exact matching results between the patterns from query and the targeted sequences. In other words, two sequences, $x_i, x_j \in T$ are indiscernible when they are equivalent with regards to their total exact matching score value, $\varepsilon$. For any $B \subseteq T$, the formal indiscernibility relation can be define as, $R(B) = \{(x_i, x_j) \in T \times T : \varepsilon(x_i) = \varepsilon(x_j)\}$. The sequences in $B_i$ are labeled with a tag representing their class. Figure 3 depicts an example of a set of equivalence classes generated from a processed DNA sequence database.
Fig. 3: Construction of equivalence class

Although classification has been made to database, the ranking position of sequences is retained. From Figure 3 shot, based on indiscernibility of sequence exact matching score there are nine classes are generated. The partitioning of the universe $T$ can be signed as a set of equivalence class, $CLASS = \{B_1, B_2, \ldots, B_9\}$ where $B_i \subseteq T$. For each $B_i \subseteq T$ we can denote as, $B_1 = \{\{gbvrl2.seq, 102, \ldots\}\}$, $B_2 = \{\{gbvrl2.seq, 36800, \ldots\}\}$, $B_3 = \{\{gbvrl2.seq, 37605, \ldots\}\}$, $B_4 = \{\{gbvrl1.seq, 12, \ldots\}\}$, $B_5 = \{\{gbvrl3.seq, 44, \ldots\}\}$, $B_6 = \{\{gbvrl1.seq, 18, \ldots\}\}$, $B_7 = \{\{gbvrl1.seq, 22, \ldots\}\}$, $B_8 = \{\{gbvrl2.seq, 333, \ldots\}\}$, $B_9 = \{\{gbvrl2.seq, 348, \ldots\}\}$.

Each class may contain the set of relevance sequences to the query or the set of irrelevance sequences to the query. However, it is well known that these two types class cannot be exactly determined until the optimal alignment process is executed. The classification (or partitioning) of universe does not give any significant impacts to the whole process of the proposed DNA sequence similarity search model and therefore the process is continued for reduction method.

3.2 DNA Sequence Database Reduction

The model needs to reduce the number of targeted sequences to be executed for rigorous optimal alignment process. Further selection of targeted DNA sequences is important. A feature selection method based on indiscernibility relation in rough sets theory is applied to remove 'redundant' sequences from the database. These redundant sequences give impact on the optimal local alignment processing performance. The reduct set is the core concept in rough set theory [11][14]. Generating the reduct set of sequences is an essential in this model. The model will automatically indicate which classes and sequences as well that must be included the reduct set. The reduct set reduces the dimensionality of original data set, ensuring no useful information is lost. In other words, reduction of sequences will reduce elementary set members and consequently improve the efficiency of optimal alignment process. In reduction process, two kinds of rules are encouraged; certain rules and possible rules. During dataset reduction, those rules are utilized in determining the approximation set of sequences to be covered in the optimal alignment. The certain rules are induced from lower approximation and the possible data set solicited by the upper approximation produces possible rules.

Let $T^*$ is a set of selected DNA sequences from database (or a set of relevance sequences after filtering) that to be executed for optimal alignment process. From the perspective of the rough sets theory, the set $T^*$ is regularly called a concept. Upon classification process of the ranked sequences has been completed, the model will have upper approximation of $T^*$ contains all sequences which can possibly be classified as belonging to the set $T^*$. The model also will have lower approximation of $T^*$ contains of the sequences that are definitely known to belong to $T^*$. For a comprehensive DNA sequence database sequence, $T$, and if $R$ is the indiscernibility relation and $T^* \subseteq T$, the lower approximation of set $T^*$, $\bar{R}T^* = \bigcup \{x_R | x \in T : [x]_R \subseteq T^*\}$, the upper approximation of set $T^*$, $\overline{R}T^* = \bigcup \{x_R | x \in T : [x]_R \cap T^* \neq \emptyset\}$ and boundary of set $T^*$, $\text{BN}(T^*) = \overline{R}T^* - \bar{R}T^*$. In other words, the lower approximation of set $T^*$ is the union of all those elementary sets each of which is contained by $T^*$ and the upper approximation of set $X$ is the union of those.
elementary sets each of which has a non empty intersection with T*. Furthermore, the relation of these lower and upper approximations can be made as, \( R^*_T \subseteq T^* \subseteq R_T^* \).

Figure 4 shows the relationship of the lower and upper approximations of set T*. Region S1 contains DNA sequences that have a very high exact matching score value and certainly will be executed for optimal alignment process. Specifically, the lower approximation of set T* is considered as a reduct set of T*. Those sequences in the reduct set are selected from the most top equivalence classes generated in the previous database classification stage. The next paragraph will discuss the details how this reduct set is generated. DNA sequences that have an enough high exact matching score value but not selected for optimal alignment process are considered in upper approximation of set T*. However, the user can select these sequences for further experiments and analysis. This upper approximation set of T* is denoted in region S2. Region S3 includes DNA sequences that are classified into the same class with sequences in region S1. These sequences will not execute for optimal alignment process. Region S4 contains DNA sequences that have a very low exact matching score value, and those sequences are certainly not included in set T*.

If \( P \subseteq B \), P is independent and \( R(P) = R(B) \), then P is a reduct of B. Superfluous or dispensable DNA sequences can be removed from the generated classes. A sequence \( x \in B \) is dispensable \( B \subseteq T \) if \( R(B) = R(B - \{x\}) \). Otherwise, the sequence is indispensable in B. If B and C are a collection of sets and \( C \subseteq B \) where the elements \( x \in B - C \) are dispensable, then C is a reduct of B and called as REDUCT(B) if \( R(C) = R(B) \). However, the model uses this reduct set to represent the entire members of the concept for executing in rigorous optimal alignment process. A reduct set should have a minimal set occurring in classification and suffices to define basic concepts or descriptions of the sequences in all \( B_i \). If \( P \subseteq B \), P is independent and \( R(P) = R(B) \), then P is a reduct of B. As mentioned early, two sequences \( x_i, x_j \in T \) are indiscernible when they are equivalence with regards to their total exact matching score, \( e \). P is a singleton set \( \{p\} \). p will represent all the sequences in B. In other words, only p will be chosen for optimal alignment process. The produced alignment result is accepted for representing similarity search of the entire class. Since all sequences in a class are considered equivalence, any sequence in a class can represent the class. However, the model chooses the first sequence in the class \( B_i \) to be inserted into the reduct set. The model also can randomly select one sequence from among those sequences belonging to the class to be included into the reduct set. If REDUCT represents the set of reduct sequences from all sequence classes, Figure 5 shows the generalization of the reduction algorithm proposed for the model. A reduct set is constructed from an empty set by inserting the selected sequences until all classes in the concept are processed.

<table>
<thead>
<tr>
<th>Input</th>
<th>Sequence classes, CLASS = {B_1, B_2, ..., B_n}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output</td>
<td>a reduct set, REDUCT</td>
</tr>
</tbody>
</table>

- \( \text{REDUCT} \leftarrow \{\emptyset\} \)
- For \( i \leftarrow 1 \) to \( n \) // number of classes considered
  - \( p \leftarrow \) the first sequence in \( B_i \)
  - Add \( p \) into REDUCT
- Return REDUCT

Fig. 5: The algorithm for computing a reduct set

### 3.3 The Experiments

In order to evaluate the proposed model, a prototype computer program has been designed and implemented. The prototype is developed using Java language, running under the Microsoft Windows XP Professional operating systems. The processor used was 1.86GHz Intel \( \circledast \) Core\textsuperscript{TM} 2 Duo processor with 1GB RAM. We have performed several queries for a set of 30,000 real DNA sequences downloaded from bio-mirror portal [15]. The query sequences are randomly
selected from those 30,000 sequences. Table 1 shows the experiment results for ten queries. The table has tabulated the results based on the process of retrieving 10% DNA sequences from the database that have similarity to the query. If \( T^* \) is the number of relevance sequences after automaton based filtering then the dependency of approximation of classification \( B \) (or quality of classification) by an indiscernibility relation \( R \) is defined as,

\[
\gamma_R(B) = \frac{\sum_i |RB_i\bigcap B|}{|B|}
\]

where \(|Y|\) indicates the cardinality of a set \( Y \). The dependency of the set \( B_{i+1} \) to degree \( k \) to the set \( B_i \) is denoted as, \( \gamma_k(B) = k \) where \( 0 \leq k \leq 1 \).

Table 1: Experiment results

<table>
<thead>
<tr>
<th>Query #</th>
<th>Automaton based filtering</th>
<th>Rough sets theory classification and reduction</th>
<th>Total Smith-Waterman skipped from T</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Relevance sequence (( T^* ))</td>
<td>Smith-Waterman skipped from T</td>
<td>Filtering precision (%)</td>
</tr>
<tr>
<td>Query 1</td>
<td>3,247</td>
<td>26,753</td>
<td>92.39</td>
</tr>
<tr>
<td></td>
<td>(89.18%)</td>
<td></td>
<td>(50.88%)</td>
</tr>
<tr>
<td>Query 2</td>
<td>3,359</td>
<td>26,641</td>
<td>89.31</td>
</tr>
<tr>
<td></td>
<td>(88.80%)</td>
<td></td>
<td>(60.67%)</td>
</tr>
<tr>
<td>Query 3</td>
<td>3,159</td>
<td>26,841</td>
<td>94.97</td>
</tr>
<tr>
<td></td>
<td>(89.47%)</td>
<td></td>
<td>(53.75%)</td>
</tr>
<tr>
<td>Query 4</td>
<td>4,055</td>
<td>25,945</td>
<td>73.98</td>
</tr>
<tr>
<td></td>
<td>(86.48%)</td>
<td></td>
<td>(61.68%)</td>
</tr>
<tr>
<td>Query 5</td>
<td>3,138</td>
<td>26,862</td>
<td>95.60</td>
</tr>
<tr>
<td></td>
<td>(89.54%)</td>
<td></td>
<td>(48.82%)</td>
</tr>
<tr>
<td>Query 6</td>
<td>3,511</td>
<td>26,489</td>
<td>85.45</td>
</tr>
<tr>
<td></td>
<td>(88.30%)</td>
<td></td>
<td>(60.04%)</td>
</tr>
<tr>
<td>Query 7</td>
<td>3,190</td>
<td>26,810</td>
<td>94.04</td>
</tr>
<tr>
<td></td>
<td>(89.37%)</td>
<td></td>
<td>(51.94%)</td>
</tr>
<tr>
<td>Query 8</td>
<td>3,393</td>
<td>26,607</td>
<td>88.42</td>
</tr>
<tr>
<td></td>
<td>(88.69%)</td>
<td></td>
<td>(45.03%)</td>
</tr>
<tr>
<td>Query 9</td>
<td>3,290</td>
<td>26,710</td>
<td>91.19</td>
</tr>
<tr>
<td></td>
<td>(89.03%)</td>
<td></td>
<td>(53.22%)</td>
</tr>
<tr>
<td>Query 10</td>
<td>3,084</td>
<td>26,916</td>
<td>97.28</td>
</tr>
<tr>
<td></td>
<td>(89.72%)</td>
<td></td>
<td>(48.44%)</td>
</tr>
</tbody>
</table>

The experiment results show that the automaton based filtering technique discards a huge number of irrelevant DNA sequences from being executed for rigorous optimal alignment. The filtering process successfully generates a group of sequences from database that have a highly potential similar to the query. By using the rough sets theory method, the relevance sequences have been classified and a reduce dataset of relevance sequence is produced. This classification and reduction process significantly improves the efficiency of the similarity search process. For instance in Query 2, automaton based filtering successfully groups the 3,359 relevance sequences to the query. However, after the classification and reduction process only 1,321 sequences are selected to be performed for rigorous optimal local alignment. These 1,321 sequences represent the characteristics of the entire relevance sequences in the generated classes. Overall, the combination of automaton based filtering and rough sets theory technique effectively remove 28,679 database sequences (95.60%) from being executed for rigorous optimal alignment. Obviously, the proposed model gives high efficiency of DNA sequence similarity search process with low computational numbers for \( \Theta(n \times m) \) time complexity Smith-Waterman algorithms. Hence, the time for retrieving a set of similar sequences from a database to a query is minimized.

4.0 CONCLUSIONS

Discovering for new biological sequences such as DNA is an energetic activity. The paper presents an application of rough sets theory to DNA sequence database classification and reduction. Similarity distance is a fundamental concept to define a data classification. The concept of indiscernibility relation provided by rough sets theory is potentially for similarity measure. Since indiscernibility relation approach does not use any distance function, a very less tasks are
expected for clustering the objects. Using the automaton based exact string matching, DNA sequence database has been processed and a total exact matching score is generated. This matching score is used in constructing indiscernibility relation among the sequences in database. Indiscernibility of processed DNA sequences is accomplished through the partitioning of the domain of dataset into equivalence classes. Sequences in all classes are then managed into approximation regions of rough sets; lower and upper approximations. Based on these approximation regions a reduct set of DNA sequence dataset is generated. Only sequences in the reduct set are considered for optimal alignment process and the produced alignment results represent the whole DNA sequences in all equivalence classes of the concept. Consequently, the number of iterations for rigorous optimal alignment algorithm will be minimized and improves the efficiency of sequence similarity search.

REFERENCES


A GENETIC-BASED FUZZY REGRESSION ANALYSIS: AN APPROACH FOR ESTIMATION OF SOFTWARE SOURCE CODE SIZE

M. Hadi Mashinchi, Siti Mariyam HJ. Shamsuddin
Faculty of Computer Science and Information System,
Universiti Teknologi Malaysia, Johor, Malaysia.
Email: h_mashinchi@yahoo.com, mariyam@utm.my

ABSTRACT

Approximating the software size code before developing is a very important issue in software planning and development. A correct approximation leads to proper scheduling and cost approximation. In this paper, a genetic-based fuzzy regression analysis is applied to learn from the training data and then the size approximation of software in advance of development is estimated. The simulation results show that our proposed method can estimate code size efficiently.

Keywords: Software Code Size, Fuzzy Regression, Genetic Algorithms

1.0 INTRODUCTION

Many efforts have been done to predict the size of the software in advance of software development. Although there are other factors that can have effect on the project scheduling but software sizing is considered one of the most crucial issues. This is due to the great influence of software sizing on the duration of project, number of team members and also cost approximation. It has been defined that the major issue in software cost estimation is to compute the accurate size of the software prior to the software development [1]. Indeed, if exact approximation of the software sizing is available then the project scheduling can be defined precisely in advance of starting the software development.

Many methods have been proposed for software sizing prediction. Some approaches and applications for software sizing are discussed in [2]. A general software size prediction is proposed in [3] with the aim of independency of size prediction with program type and characteristics. Software sizing for object-oriented systems are investigated in [4]. Also, a fuzzy logic model for software source code sizing estimation is proposed recently [5].

In this research, we proposed a new approach for source code estimation. For estimation purpose, a genetic-based fuzzy regression analysis is proposed. A fuzzy regression can find a fuzzy function in form of \( \tilde{y} = f(\tilde{x}) \) that can interpolate the fuzzy number data \( (\tilde{x}_i, f(\tilde{x}_i)) \) and \( i \) is the index for the fuzzy data that are going to be interpolated. Recently, a genetic algorithm is applied for the regression purpose of crisp data [6]. This method has shown better approximation comparing to the other approaches such as ones using neural networks. In addition, genetic algorithms are used to find a polynomial function to fit the fuzzy data [7], [8] as an evolutionary approach for fuzzy regression. In this paper, this regression approach is applied to estimate code sizing of software in advance of software development.

The rest of the paper is as following; in Section 2 fuzzy regression with genetic algorithms is discussed. In Section 3 is shown how to apply the genetic-based fuzzy regression for code sizing estimation. Section 4 gives a simulation result for proposed method. Finally, in Section 5 the conclusion and future works are discussed.

2.0 GENETIC-BASED FUZZY REGRESSION ANALYSIS

Genetic algorithms inspired from natural evolutionary processes. Three operations are involved in genetic algorithms; selection, crossover and mutation. In the first stage, random population of solutions is initiated. Then, a fitness function which returns the goodness of each solution is applied to select the best solutions among the current generation for crossover and mutation process. Selected solutions merge together to make new solutions. Then a mutation operation is applied to change some portion of the produced solutions for next generation. These steps are repeated until a solution with a required fitness value is captured.

Many approaches are proposed for fuzzy regression. Among them, evolutionary approaches have a great performance especially when the number of the variables is increased for optimization. Also, recently a genetic-based fuzzy BP method is proposed by authors in [9] which the regression process takes less time comparing to the one without genetic
algorithms in [10]. Another proposed regression method with evolutionary algorithms is genetic-based fuzzy regression method in [6], [7].

In the proposed genetic-based fuzzy regression method in [6], [7], a linear fuzzy function is proposed which is able to interpolate the set of data. Then if the assumed linear function was not able to fit the fuzzy data, another function will be used. The linear fuzzy function is in form of

$$f(\tilde{x}) = \tilde{a} \tilde{x} + \tilde{b}$$

where \(\tilde{x}, \tilde{a}\) and \(\tilde{b}\) are fuzzy values. A fuzzy set \(\tilde{A}\) on \(\mathbb{R}\), the set of real numbers, is defined in (1).

$$\tilde{A} = \{(x, \mu_\tilde{A}(x)) | x \in \mathbb{R}, \mu_\tilde{A} : \mathbb{R} \rightarrow [0,1]) \}.$$ (1)

From now, by \(\alpha^{th}\) level set \((\alpha - cut)\) and strong \(\alpha - cut\) for fuzzy set \(\tilde{A}\), we mean the set in (2) and (3) respectively.

$$\tilde{A}_\alpha = \{x \in \mathbb{R} | \mu(x) \geq \alpha; \alpha \in [0,1]\},$$ (2) $$\tilde{A}_\alpha = \{x \in \mathbb{R} | \mu(x) > \alpha; \alpha \in [0,1]\}. $$ (3)

Having pair set of fuzzy input-output values in form of \(\left((\tilde{x}_1, \tilde{y}_1), (\tilde{x}_2, \tilde{y}_2), ..., (\tilde{x}_n, \tilde{y}_n)\right)\), where \(\tilde{x}_j\) and \(\tilde{y}_j\) are fuzzy inputs and desired fuzzy outputs respectively, the aim is to find almost the best fuzzy values \(\tilde{a}\) and \(\tilde{b}\) for \(f(\tilde{x}) = \tilde{a}\tilde{x} + \tilde{b}\). The optimized \(\tilde{a}\) and \(\tilde{b}\) make the linear function able to fit the fuzzy input-output values with a good accuracy which is computed with fitness function in genetic algorithms. The fitness function is defined in (4).

$$f(\tilde{x}) = \int_{\alpha=0}^{1} \left( L^L_{\alpha}(x^L, x^U) - R^L_{\alpha} \right) + \left( L^U_{\alpha}(x^L, x^U) - R^U_{\alpha} \right) d\alpha,$$ (4)

where \(L^L_{\alpha}(x^L, x^U)\) and \(R^L_{\alpha}\) are the lower bands of close interval for left side of the equation subsituted with \(\alpha^{th}\) level sets of the \(\tilde{x}\) and fuzzy constant in the right side of the fuzzy equation respectively. Similarly, \(L^U_{\alpha}(x^L, x^U)\) and \(R^U_{\alpha}\) are the upper bands of close interval for left side of the equation subsituted with \(\alpha^{th}\) level sets of the \(\tilde{x}\) and fuzzy constant in the right side of the fuzzy equation respectively.

3.0 SOFTWARE SIZING PREDICTION WITH GENETIC_BASED FUZZY REGRESSION

Regarding to [1] size estimation for software is a very difficult and subject to wide uncertainties. Therefore, a fuzzy regression method can be helpful. For size estimation two variables have more influence; first is the number of non-menu processes and second is the number of attributes in system’s data model [5]. According to an experiment in [5] twelve fuzzy If-Then rules are generated for software code size in following form:

**IF (number of attribute is A) and (number of non-menu is B) THEN (software size is C).**

Thus, fuzzy values of \(\tilde{a}, \tilde{b}\), and \(\tilde{c}\) for a fuzzy linear function in form of \(\tilde{z} = \tilde{a}\tilde{x} + \tilde{b}\) will be estimated by genetic-based fuzzy regression to fit the fuzzy If-Then rules.

The ultimate aim is to find fuzzy values of \(\tilde{a}, \tilde{b}\), and \(\tilde{c}\) in which after substituting inputs of fuzzy If-Then rules the outputs of linear function become exactly as the desired fuzzy outputs. Therefore, the fitness value is zero since the actual and the desired output are same. So, a good approximation for \(\tilde{a}, \tilde{b}\), and \(\tilde{c}\) occurs when the fitness value is very close to zero.
4.0 SIMULATION AND RESULT

For the simulation, four fuzzy If-Then rules from [5] as following are used to approximate the fuzzy linear function and the rests are used for validating and testing the estimated function.

\[
\text{IF (} x \text{ is small) and (} y \text{ is small) THEN (} z \text{ is small),}
\text{IF (} x \text{ is large) and (} y \text{ is large) THEN (} z \text{ is large),}
\text{IF (} x \text{ is small) and (} y \text{ is large) THEN (} z \text{ is medium),}
\text{IF (} x \text{ is large) and (} y \text{ is small) THEN (} z \text{ is medium),}
\]

where \( x \), \( y \) and \( z \) are represent number of non-menu process, attributes and software code size, repectively. Figure 1 shows the fuzzy membership function for small, medium and large.

![Figure 1: fuzzy membership functions of small, medium and large](image)

For optimization, 1000 generations of genetic algorithms with 50 chromosomes is applied to find the best values of the \( \tilde{a} \), \( \tilde{b} \), and \( \tilde{c} \) in function \( \tilde{z} = \tilde{a} \tilde{x} + \tilde{b} \tilde{y} + \tilde{c} \). Also, for simplicity, a genetic algorithm is applied for only level set one and strong level set zero. The estimated \( \tilde{a} \), \( \tilde{b} \), and \( \tilde{c} \) are shown in Table 1.

<table>
<thead>
<tr>
<th>Strong level set 0</th>
<th>Level set 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tilde{a} )</td>
<td>[0.4569,0.4732]</td>
</tr>
<tr>
<td>( \tilde{b} )</td>
<td>[0.4586,0.4676]</td>
</tr>
<tr>
<td>( \tilde{c} )</td>
<td>[0.1695,0.3025]</td>
</tr>
</tbody>
</table>

The estimated \( \tilde{a} \), \( \tilde{b} \), and \( \tilde{c} \) after 1000 generations, generate error of 1.1573 for four fuzzy If-Then rules. Thus, the error for each level set of each fuzzy rule is 0.1447. Lower error can be reached if more generations are applied.

For the testing purpose following fuzzy If-Then rules are used from [5]. Figure 2 shows their fuzzy membership function.

R1: IF (\( x \) is small-medium) and (\( y \) is medium) THEN (\( z \) is small-medium),
R2: IF (\( x \) is medium) and (\( y \) is medium-large) THEN (\( z \) is medium-large),

The results from the testing stage are shown in Figure 3 and Table 2.

![Figure 2: fuzzy membership functions of small-medium and medium-large](image)
### Table 2: Level set intervals of tested fuzzy If-Then rules, R1 and R2.

<table>
<thead>
<tr>
<th></th>
<th>Strong level set</th>
<th>Level set 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>[0.6376, 3.5444]</td>
<td>[1.5993, 2.6316]</td>
</tr>
<tr>
<td>R2</td>
<td>[2.5047, 5.4206]</td>
<td>[3.4319, 4.4923]</td>
</tr>
</tbody>
</table>

The results from the testing stage are shown in Figure 3 and table 2.

![Figure 3: approximated fuzzy membership functions of small-medium and medium-large.](image)

The actual fuzzy outputs are good representations for small-medium and medium-large in R1 and R2 respectively. Thus, it is clear from Figure 3 that the estimated function is able to predict linguistic verbal of software code size of fuzzy If-Then rules with a good approximation.

#### 5.0 CONCLUSION AND FUTURE WORK

In this paper a genetic-based fuzzy linear regression approach is used for software source code size estimation. In the proposed method, the fuzzy coefficients, $\hat{a}$, $\hat{b}$, and $\hat{c}$ can be approximated with any degree of accuracy by genetic algorithms.

Three future researches are going to be held, the first stage is to apply a proper defuzzification method to find the exact software code size. Second, in case of more complicated fuzzy data, we are going to apply genetic algorithm to find the best type of fuzzy function to fit the data. And third, after finding the optimized type of function and its coefficients, it is possible to find the roots of the estimated function [11] to reach desired software source code size.

#### Acknowledgement

Authors would like to thank Ministry of Higher Education (MOHE), and Research Management Center, Universiti Teknologi Malaysia for their support in terms of research grant and research activities.

#### REFERENCES


HEART DISEASE DECISION SUPPORT SYSTEM USING DATA MINING CLASSIFICATION MODELING TECHNIQUES

Sellappan Palaniappan¹, Rafiah Awang²
¹,²Department of Information Technology, Malaysia University of Science and Technology Malaysia, Kelana Square, 47301 Petaling Jaya, Malaysia.
Email: {sell, mitm5rb1}@must.edu.my

ABSTRACT

The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not “mined” to discover hidden information for effective decision making by healthcare practitioners. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining modeling techniques can help remedy this situation. This research has developed a prototype Heart Disease Decision Support System (HDDSS) using Data Mining Classification Modeling Techniques, namely, Decision Trees, Naïve Bayes and Neural Network. Results show that each technique has its own strength in realizing the objectives of the defined mining goals. HDDSS can answer complex “what if” queries, which traditional decision support systems, cannot. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting heart disease. It enables significant knowledge, e.g., patterns, relationships between medical factors related to heart disease, to be established. HDDSS is Web-based, user-friendly, scalable, reliable and expandable. It is implemented on the.NET platform.

Keywords: healthcare, heart disease, decision support system, data mining, decisions trees, neural network, Naive Bayes

1.1 BACKGROUND AND MOTIVATION

A major challenge for healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable prices. Quality service implies diagnosing patients correctly and administering treatments that are effective. Making wrong clinical decisions is unacceptable and can be disastrous. Hospitals must also strive to minimize the cost of clinical tests as well as provide preventive care. To achieve these objectives, they need intelligent computer-based information and/or decision support systems. Such systems must provide decision support to assist healthcare practitioners (doctors, clinicians) to make intelligent clinical decisions.

Many hospitals today use computer-based information systems to manage their healthcare information, e.g., Hospital Information Systems, Total Hospital Information Systems, Laboratory Information Systems, Patient Information Systems, Pharmaceutical Information Systems and Basic Hospital Information System [17].

All these systems generate huge amounts of healthcare data, which take the form of numbers, structured and unstructured text, and charts and images. Unfortunately, most of these data are rarely used to support clinical decision-making. There is a wealth of hidden information in the data that is largely untapped. This raises an important question: “How do we turn the data into useful information to support decision making by healthcare practitioners?” This is the main motivation for this research.

2.0 PROBLEM STATEMENT

Many hospital information systems are designed to support patient billing, inventory management and generation of simple statistics. Where there are decision support systems, they are largely limited. They can only answer simple queries like “What is the average age of patients who have heart disease?”, “How many surgeries had resulted in hospital stays longer than 10 days?” and “Identify the female patients whose are single, above 30 years old and who have been treated for cancer?” They cannot answer complex queries like “Identify the important preoperative predictors that increase the length of hospital stay?”, “Given patient records on cancer, should treatment include chemotherapy alone, radiation alone, or both chemotherapy and radiation?” and “Given patient records, predict the probability of patients who will have heart disease?”
Clinical decisions are often made based on doctors’ intuition rather than on the knowledge-rich data stored in the database. This practice leads to unwanted biases, errors and excessive medical cost which affect the quality of service provided to patients.

Wu, et al suggested that integration of clinical decision support with computer-based patient records has the potential to reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome [24]. This is promising as modeling and data analysis tools such as data mining have the potential to create a knowledge-rich environment that can significantly improve the quality of service.

3.0 OBJECTIVES OF RESEARCH

The main objective of this research is to develop a prototype Heart Disease Decision Support System (HDDSS) using data mining. Specifically, it will use three data mining classification modeling techniques, namely, Decision Trees, Naïve Bayes and Neural Network. HDDSS can discover and extract hidden knowledge (patterns and relationships) associated with heart disease from a historical heart disease database (declassified). It can answer complex queries for diagnosing heart disease. To enhance visualization and ease of interpretation, it will display the result in tabular as well as graphical formats.

HDDSS will provide decision support to assist healthcare practitioners to make intelligent clinical decisions related to heart disease that traditional decision support systems cannot. It will also enable them to provide more effective treatments and reduce medical costs. Thus it will improve the quality of service provided to patients.

HDDSS will be implemented on the .NET platform. It will be Web-based, user-friendly, scalable, reliable and expandable.

4.0 DATA MINING REVIEW

Although data mining has been around for around two decades its potential is only being realized now. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. Examples of data mining applications include studying air travel patterns to maximize seat utilization, buying patterns and associations between products in supermarkets to increase sales, and using X-ray results to analyze and detect abnormal pattern [21].

Fayyad defines data mining as “a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database” [7]. Giudici defines it as “a process of selection, exploration and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of database” [9].

Data mining uses two strategies: supervised and unsupervised learning. Supervised learning uses values of input variables to predict a target variable with a known value while unsupervised learning works in a similar way, but more frequently it predicts a target variable where a known value does not exist [17].

4.1 Data Mining Techniques

There are several data mining techniques such as Decision Trees, Neural Network, Naïve Bayes, Logistic Regression, Association Rules and Clustering. Each technique serves a different purpose depending on the modeling objective. The two most common modeling objectives are classification and prediction. Classification models predict categorical labels (discrete, unordered) while prediction models predict continuous-valued functions [10]. Classification models also predict class labels of unknown records. Decision Trees and Neural Networks use classification algorithms while Regression, Association Rules and Clustering use prediction algorithms [4]. A brief survey of the three data mining techniques used in this research is given below.

4.1.1 Decision Trees

Decision Trees split data recursively into mutually exclusive and exhaustive subgroups. All input attributes from the training set are evaluated at each split for their impact on the predictable attribute. The rule evaluation is first started at the root of the decision tree. The root has no incoming edges but has outgoing edges (branches). Then, at each internal node (nonleaf), the evaluation is repeated recursively. Each internal node has only one incoming edge and two or more
outgoing edges. The evaluation stops when there is no more split or the record arrives at a leaf node. Each leaf node holds a class label with one incoming and no outgoing edge.

Recursive algorithms employ a greedy strategy that grows a decision tree by making a series of locally optimum decisions about which attribute to use for partitioning the data [20]. Three popular decision tree algorithms are CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and C4.5. The algorithms differ in selection of splits, when to stop a node from splitting, and assignment of class to a non-split node [11]. Gini index is used in CART to measure the impurity of a partition or set of training tuples [10]. ID3 and C4.5 algorithms use a criterion called gain to select attributes based on the entropy concept used in information theory [14].

Decision Trees is the most popular data mining technique. It is used in both predictive and classification modeling objectives. It handles high dimensional, categorical and continuous data. It shows a series of conditional choices and the impact of time on decisions and produces numerical results. Its output is easy to read and interpret. But it has one drawback: it is not effective for predicting continuous values.

4.1.2 Naives Bayes

Bayes’ Rule, due to Reverend Thomas Bayes, is the basis for many machine-learning and data mining methods [20]. The algorithm (rule) can be used to create models with predictive capabilities. It provides new ways of exploring and understanding data. The algorithm learns from the “evidence” by calculating the correlation between the target variable and the other variables.

Bayes Rule uses a combination of conditional and unconditional probabilities. It states that if you have a hypothesis H and evidence about hypothesis E, then you can calculate the probability of H using the formula P(H|E) = P(E|H) * P(H)/P(E) [20]. This simply states that the probability of a hypothesis given the evidence is equal to the probability of the evidence given the hypothesis multiplied by the probability of the hypothesis. Naive Bayes (Bayes Rule) is suitable for classification modeling objectives. It executes fast and its output is easy to read and understand.

4.1.3 Neural Networks

A Neural Network is similar to the human brain with millions of interconnected neurons. It consists of three layers: input, hidden, and output units or variables. Every unit in the first layer is connected to every unit in the second and subsequent layers [11]. Inputs are derived from input units. Connection between each input unit and each hidden or output unit is based on relevance of the assigned value (weight) of that particular input unit. The higher the weight of an input unit, the more important it is. At the hidden (intermediate) units, all input units are processed and computed using a linear or radial function. The results are then passed to the output units using a transfer function. Neural Network algorithms use the Linear and Sigmoid transfer functions. At the output units, the values are processed and compared with the predicted values.

Neural Networks support both predictive and classification modeling objectives. They are suitable for training huge amounts of data with little input. However, for large number of input units, finding patterns can be more time consuming than Decision Trees and Naïve Bayes. Neural Networks are often used in commercial applications for analyzing complex input data, e.g., voice and handwriting recognition, fraud detection of credit, customer churn analysis. The downside is: it is more difficult to interpret results. Relationship between one input attribute and the other input attributes is also harder to understand. Neural Network is best used when other data mining techniques are unsatisfactory.

5.0 METHODOLOGY

HDDSS uses the CRISP-DM methodology. All six phases of this methodology are used to build the three data mining models [3]. Data Mining Extensions (DMX), a SQL-style query language for data mining, is used for building and accessing contents of the models. Tabular and graphical visualizations are incorporated to improve analysis and interpretation of results.
5.1 Data Source

A total of 909 records with 15 medical attributes (factors) were obtained from the Cleveland Heart Disease database [2]. The records were split equally into two datasets: training dataset (455 records) and testing dataset (454 records). Records for each set were selected randomly to avoid bias. In this research, we have chosen the classification-modeling objective as Naïve Bayes algorithm supports only categorical (discrete) attributes. Decision Trees and Neural network algorithms both support categorical and continuous attributes. To ensure consistency, we have used categorical attributes for all three models. All the medical attributes shown in Table 1 were transformed from numerical to categorical data. We have identified the medical attribute “Diagnosis” as the predictable attribute with value “1” for patients with heart disease and value “0” for patients with no heart disease. The attribute “PatientID” (patient identification number) was used as the key and the rest are input attributes. It is assumed that data quality such as noise and missing, inconsistent and duplicate data have been resolved in the datasets.

<table>
<thead>
<tr>
<th>Predictable Attribute</th>
<th>Key Attribute</th>
<th>Input Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis (value 0: &lt; 50% diameter narrowing (no heart disease); value 1: &gt; 50% diameter narrowing (has heart disease))</td>
<td>PatientID – Patient’s identification number</td>
<td></td>
</tr>
<tr>
<td>Sex (value 1: Male; value 0: Female)</td>
<td>Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)</td>
<td></td>
</tr>
<tr>
<td>Fasting Blood Sugar (value 1: &gt; 120 mg/dl; value 0: &lt; 120 mg/dl)</td>
<td>Restecg – resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)</td>
<td></td>
</tr>
<tr>
<td>Exang – exercise induced angina (value 1: yes; value 0: no)</td>
<td>CA – number of major vessels colored by floursopy (value 0 – 3)</td>
<td></td>
</tr>
<tr>
<td>Slope – the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: downsloping)</td>
<td>Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)</td>
<td></td>
</tr>
<tr>
<td>Trest Blood Pressure (mm Hg on admission to the hospital)</td>
<td>Serum Cholestoral (mg/dl)</td>
<td></td>
</tr>
<tr>
<td>Thalach – maximum heart rate achieved</td>
<td>Oldpeak – ST depression induced by exercise relative to rest</td>
<td></td>
</tr>
<tr>
<td>Age in Year</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.2 Mining Models

Data Mining Extension (DMX) query language was used for model creation, model training, model prediction and accessing the contents of the mining models. All parameters were set to default setting except parameters “Minimum Support = 1” for Decision Tree and “Minimum Dependency Probability = 0.005” for Naïve Bayes [16]. The trained model was evaluated against the testing dataset for their accuracy and effectiveness before they are deployed in HDDSS. Lift Chart and Classification Matrix were used in the models.

5.3 Model Validation

Lift Chart was used to validate the three models to determine if there was sufficient information to learn patterns in response to the predictable attribute. Columns in the mining structure of a trained model were mapped to columns in the testing dataset. The models, predictable column to chart against, and the state of the predictable column to predict patients with heart disease (predict value = 1) also had to be selected. Fig. 1 shows the Lift Chart output. The X-axis shows the percentage of the test dataset that is used to compare the predictions. The Y-axis shows the percentage of values predicted to the specified state. The blue and green lines show the random-guess and ideal models respectively. The purple, yellow and red lines show the Neural Network, Naïve Bayes and Decision Tree models respectively.

The top line (green) shows the ideal model; it captures 100% of the target population for patients with heart disease using 46% of the testing dataset. The bottom line (blue) shows the random line which is always a 45-degree line across
the chart. It indicates that if we are to randomly guess the result for each case, 50% of the target population would be captured using 50% of the testing dataset. All three model lines (purple, yellow and red) fall between the random and ideal lines. This shows that all three models have sufficient information to learn patterns in response to the predictable state.

![Data Mining Lift Chart for Mining Structure: Heart Disease Training](chart)

**Fig. 1: Result of Lift Chart with predictable value**

5.4 Testing Model Effectiveness

The effectiveness of the models was tested using two methods: Lift Chart and Classification Matrix. The purpose was to determine which model gave the highest percentage of correct predictions for diagnosing patients with heart disease.

5.4.1 Lift Chart with No Prediction Value

The steps for producing Lift Chart are similar to that given in Section 5.3 except that the state of the predictable column must be left blank. It does not include a line for the random-guess model. It tells how well each model fared at predicting the correct number of the predictable attribute. Fig. 2 shows the Lift Chart output. The X-axis represents the percentage of testing dataset used to compare predictions. The Y-axis represents the percentage of predictions that are correct. The blue, purple, green and red lines represent the ideal, Neural Network, Naïve Bayes and Decision Trees models respectively. It shows the performance of the models across all possible states. The ideal line (blue) is at a 45-degree angle, indicating that if 50% of the testing dataset is processed, then 50% of testing dataset is predicted correctly.

The chart shows that when 50% of the population is processed, Neural Network has the highest percentage of correct predictions (49.34%) followed by Naïve Bayes (47.58%) and Decision Trees (41.85%). When the whole population is processed, Naïve Bayes model appears to be superior to the other two as it has the highest number of correct predictions (86.12%) followed by Neural Network (85.68%) and Decision Trees (80.4%).

Processing less than 50% of the population causes the Lift lines for Neural Network and Naïve Bayes to be always higher than that for Decision Trees, indicating that Neural Network and Naïve Bayes are better at making high percentage of correct predictions than Decision Trees. Along the X-axis the Lift lines for Neural Network and Naïve Bayes overlap, indicating that both models are equally good at making correct predictions.

When more than 50% of population is processed, Neural Network and Naïve Bayes appear better as they yield high percentage of correct predictions than Decision Trees. This is because the Lift line for Decision Trees is always below that of Neural Network and Naïve Bayes. For some population range, Neural Network appears better than Naïves Bayes and vice-versa.
5.4.2 Classification Matrix

Classification Matrix displays the frequency of correct and incorrect predictions. It compares the actual values in the testing dataset with the values predicted by the trained model. In our example, the testing dataset contained 208 patients with heart disease and 246 patients without heart disease. Fig. 3 shows the results of the classification matrix for the three models. The rows represent the predicted values while the columns represent the actual values (1 for patients with heart disease; ‘0’ for patients with no heart disease). The left-most columns show the values predicted by the models. The diagonal values show correct predictions.

Table 2 summarizes the results for the three models. The most effective model appears to be Naïve Bayes as it has the highest percentage of correct predictions (86.53%) for patients with heart disease, followed by Neural Network (with a difference of less than 1%) and Decision Trees. However, Decision Trees appears to be best for predicting patients with no heart disease (89%) compared to the other two models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>No. of cases</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree</td>
<td>Patients with heart disease, predicted as having heart disease</td>
<td>146</td>
<td>Correct</td>
</tr>
<tr>
<td></td>
<td>Patients with no heart disease, predicted as having heart disease</td>
<td>27</td>
<td>Incorrect</td>
</tr>
<tr>
<td></td>
<td>Patients with no heart disease, predicted as having no heart disease</td>
<td>219</td>
<td>Correct</td>
</tr>
<tr>
<td></td>
<td>Patients with heart disease, predicted as having no heart disease</td>
<td>62</td>
<td>Incorrect</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>Patients with heart disease, predicted as having heart disease</td>
<td>180</td>
<td>Correct</td>
</tr>
<tr>
<td></td>
<td>Patients with no heart disease, predicted as having heart disease</td>
<td>35</td>
<td>Incorrect</td>
</tr>
<tr>
<td></td>
<td>Patients with no heart disease, predicted as having no heart disease</td>
<td>211</td>
<td>Correct</td>
</tr>
<tr>
<td></td>
<td>Patients with heart disease, predicted as having no</td>
<td>28</td>
<td>Incorrect</td>
</tr>
</tbody>
</table>
### 5.5 Evaluation of Mining Goals

Five mining goals were defined based on exploration of the heart disease dataset and objectives of this research. The goals were evaluated against the three-trained models. We found that all three models achieved the objectives of the mining goals suggesting that they could provide good decision support to healthcare practitioners in diagnosing patients and discovering the medical factors associated with heart disease. The defined mining goals are given below.

#### 5.5.1 Mining Goal 1

The first goal is “Given patients’ medical profiles, predict those who are likely to be diagnosed with heart disease.” All three models are able to answer this question using the singleton query and batch or prediction join query. Both queries are used to predict based on a single case and a table-full of multiple inputs cases respectively. HDDSS supports prediction queries based on “what if” scenarios. Users input values of medical attributes to diagnose patients with heart disease. For example, entering the attributes Age = 70, CA = 2, Chest Pain Type = 4, Sex = M, Slope = 2 and Thal = 3 into the models, would produce the results shown in Fig. 4. All three models concur that the person diagnosed will have heart disease. Naïve Bayes gives the highest probability (95%) with 432 supporting cases, followed closely by Decision Tree (94.93%) with 106 supporting cases and Neural Network (93.54%) with 298 supporting cases. As these high figures, doctors could recommend that the patient undergo further heart examination. Performing “what if” scenarios could thus help prevent a potential heart attack.

#### 5.5.2 Mining Goal 2

The second mining goal is “Identify the significant influences and relationships in the medical inputs associated with the predictable state – heart disease.” The Dependency viewer in Decision Trees and Naïve Bayes models shows the results from the most significant to the least (weakest) medical predictors. The viewer is especially useful when there are many predictable attributes. Fig. 5 and Fig. 6 show that in both models, the most significant factor influencing heart disease is “Chest Pain Type”. Other significant factors include Thal, CA and Exang. Decision Trees model shows ‘Trest Blood Pressure’ as the weakest factor while Naïve Bayes model shows ‘Fasting Blood Sugar’ as the weakest factor. Naïve Bayes appears to be better than Decision Trees because it gives the significance of all input attributes. Doctors can use this information to further analyze the strengths and weaknesses of all medical attributes associated with heart disease.
5.5.3 Mining Goal 3

The third mining goal is “Identify the impact and relationship between the medical attributes in relation to the predictable state – heart disease.” Identifying the impact and relationship between the medical attributes in relation to heart disease is only found in Decision Trees viewer (Fig. 7). It gives the highest probability (99.61%) that patients with heart disease are found in the relationship between these attributes (nodes): “Chest Pain Type = 4 and CA = 0 and Exang = 0 and Trest Blood Pressure >= 146.362 and < 158.036.” Using this information, doctors can perform medical screening on these four attributes instead on all attributes on prospective patients who are likely to be diagnosed with heart disease. This will reduce patients’ medical expenses, administrative costs and diagnosis time. Information on least impact is found in the relationship between the attributes: “Chest Pain Type not = 4 and Sex = F” with a probability of 5.88%. The relationship between attributes for patients with no heart disease is also given. Results show that the relationship between the attributes: “Chest Pain Type not = 4 and Sex = F” has the highest impact with a probability of 92.58%. The least impact with a probability 0.2% is found in the attributes: “Chest Pain Type = 4 and CA = 0 and Exang = 0 and Trest Blood Pressure >= 146.362 and < 158.036”. Additional information such as identifying patients and their medical profiles based selected nodes can be accessed using the drill through function. Doctors can use the Decision Tree viewer to perform further analysis.
5.5.4 Mining Goal 4

The fourth mining goal is "Identify characteristics of patients with heart disease." Only Naïve Bayes model identifies characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state. Fig. 8 shows that 80.8% of the heart disease patients are males (Sex = 1) and 43.5% are between ages 56 and 63. Other significant characteristics are high probability of fasting blood sugar with less than 120 mg/dl reading, chest pain type is asymptomatic, slope of peak exercise is flat, etc. Fig. 9 shows the characteristics of patients with no heart disease with high probability in fasting blood sugar with less than 120 mg/dl reading, no exercise induced, number of major vessels is zero, etc. These can be further analyzed.

![Fig. 7: Decision Tree Viewer](image)

![Fig. 8: Naïve Bayes attribute characteristics viewer in descending order for patients with heart disease](image)

![Fig. 9: Naïve Bayes attribute characteristic viewer in descending order for patients with no heart disease](image)

5.5.5 Mining Goal 5

The last mining goal is "Determine the attribute values that differentiate nodes favoring and disfavoring the predictable states: (1) patients with heart disease (2) patients with no heart disease." This query can be answered by analyzing the results of attribute discrimination viewer from both Naïve Bayes and Neural Network models. The viewer provides information the impact of all attribute values related to the predictable state. Naïve Bayes model (Fig. 10) shows the most important attribute favoring patients with heart disease: "Chest Pain Type = 4" with 158 cases and 56 patients with no heart disease. The input attributes "Thal = 7" with 123 (75.00%) patients, "Exang = 1" with 112 (73.68%) patients,"Slope = 2" with 138 (66.34%) patients, etc. also favor predictable state. In contrast, input attributes "Thal = 3" with 195 (73.86%) patients, "CA = 0" with 198 (73.06%) patients, "Exang = 0" with 206 (67.98%), etc. favor predictable state for patients with no heart disease.

Neural Network model (Fig. 11) shows that the most important attribute value that favors patients with heart disease is "Old peak = 3.05 – 3.81" with a probability of 98%. Other attributes that favor heart disease include "Old peak >= 3.05 – 3.81", etc. favor predictable state for patients with no heart disease.
3.81”, “CA=2”, “CA=3”, etc. We can see that medical attributes like “Serum Cholesterol >= 382.37”, “Chest Pain Type = 2”, “CA=0”, etc. also favor the predictable state for patients with no heart disease.

Fig. 10: A Tornado chart for attribute discrimination viewer in descending order for Naïve Bayes

Fig. 11: Attribute discrimination viewer in descending order for neural network

6.0 CONCLUSION AND FUTURE WORK

A prototype heart disease decision support system is developed using three data mining classification-modeling techniques. The system extracts hidden knowledge from a historical heart disease database. DMX query language and functions are used to build and access the models. The models are trained and validated against a testing dataset. Lift Chart and Classification Matrix methods are used to evaluate the effectiveness of the models. All three models are able to extract patterns in response to the predictable state. The most effective model to predict patients who are likely to have a heart disease appears to be Naïve Bayes followed by Neural Network and Decision Trees.

Five mining goals are defined based on business intelligence and data exploration. The goals are evaluated against the trained models. All three models could answer complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy. Naïve Bayes model could answer four out of the five goals; Decision Trees model, three, and Neural Network model, two. Though not the most effective, Decision Trees model results are easier to understand and interpret. The drill through feature to access detailed patients’ profiles is available only in the Decision Trees model. Naïve Bayes model fared better than Decision Trees model as it could identify all significant medical predictors. The relationship between attributes that provides by the Neural Network model is more difficult to understand.

Further work can be done to enhance and extend HDDSS. For example, it can incorporate other data mining techniques such as Time Series, Clustering and Association Rules. Text Mining can also be used to mine the vast amount of unstructured data available in many healthcare databases. Another challenge would be the integration of data mining and text mining known as duo-mining (Weiguo, Rich and Zhongju, 2006).

REFERENCES

Proceeding of the 2nd International Conference on Informatics, 2007


**BIOGRAPHY**

Sellappan Palaniappan obtained his PhD from University of Pittsburgh in 1978. He is currently Associate Professor at the Department of Information Technology, Malaysia University of Science and Technology. His research interests include collaborative software development, web services, ontology for healthcare information integration and healthcare decision support systems using OLAP and data mining.

Rafiah Awang obtained her Master of Science in Information Technology from Malaysia University of Science and Technology in 2007. She is currently working as a System Analyst. Her research interests include decision support systems using data mining.
INVESTIGATING MULTIOBJECTIVE OPTIMIZATION USING AN AUGMENTED COEVOLUTIONARY SPEA2 ALGORITHM

Tse Guan Tan¹, Jason Teo², Hui Keng Lau³
Centre for Artificial Intelligence,
Universiti Malaysia Sabah, 88999 Kota Kinabalu, Sabah, Malaysia.
Email: ¹tseguantan@gmail.com, ²{jtewe, hklau}@ums.edu.my

ABSTRACT

A new algorithm for solving multiobjective optimization problems with three to five objectives, namely SPEA2-CE-KR, is proposed. This introduced algorithm is an extension of a state of the art multiobjective evolutionary algorithm: Strength Pareto Evolutionary Algorithm 2 (SPEA2). SPEA2-CE-KR is a hybridization of SPEA2 and Competitive Coevolution (CE) concept with K-Random Opponents (KR) competitive fitness strategy. Comparison performance between SPEA2-CE-KR and SPEA2 are demonstrated with a set of DTLZ test problems, from DTLZ1 to DTLZ7. The experimental results reveal that, this presented algorithm outperforms SPEA2 in terms of convergence to the true Pareto front and the coverage of the obtained nondominated solutions.

Keywords: competitive coevolution, K-random opponents, multiobjective, Pareto optimization, SPEA2.

1.0 INTRODUCTION

A variety of optimization solution techniques have been introduced for solving multiobjective problems or tasks. Among these techniques, Evolutionary algorithms have been established as one of the most widely used method. Evolutionary Algorithms (EAs) [6] are capable to find a set of optimal solutions in a single run. Through this advantage feature, EAs are useful in the circumstance of multiobjective optimization, which makes the problems or tasks of approximating Pareto front of optimal solutions. Recently, Multiobjective Evolutionary Algorithms (MOEAs) have been shown very helpful to solve a lot of real life problems that have numerous objectives need to be optimized.

In this paper, we propose a novel multiobjective evolutionary algorithm called SPEA2-CE-KR, which consists of hybrid between SPEA2 and K-Random competitive coevolution to solve DTLZ test suite, from DTLZ1 to DTLZ7 with 3 to 5 objectives. The DTLZ test suite is a scalable set introduced by Deb et al [4]. SPEA2-CE-KR is compared with the original SPEA2 and the performance of these algorithms is evaluated by using generational distance and coverage.

The remainder of this paper is organized as follows. Section 2 describes the structure of SPEA2-CE-KR. The characteristics of performance metrics will be illustrated in Section 3. The experimental setup, results and discussion are given in Section 4. Section 5 defines the conclusion and future work.

2.0 THE HYBRID ALGORITHM: SPEA2-CE-KR

The algorithm which integrates SPEA2 and K-Random (KR) competitive coevolution (CE), referred to as the SPEA2-CE-KR will be presented. The SPEA2 is selected as the modified algorithm because it is one of the current state-of-the-art MOEAs. The main motivation for hybridizing SPEA2 with a competitive coevolution strategy is that through natural competition and selection among individuals of the population, the performance of the multiobjective evolutionary optimization process can be improved since there would now be no reliance on an external fitness measure to explicitly rank the goodness of evolved solutions. Rather this ranking will now be obtained through the implicit tournament selection strategies inherent within the coevolutionary framework. The general framework for SPEA2-CE-KR is shown in Fig. 1.

```plaintext
gen = 0
Pop, (gen) = randomly initialized population
Fitness_assignment Pop,(gen)
Opponents_selection Pop,(gen)
Reward_assignment Pop,(gen)
Environmental_selection Pop,(gen)
while termination = false do begin
  gen = gen + 1
end
```
Generally, the framework for the SPEA2-CE-KR is similar to the framework for SPEA2 [3] with the exceptions of two additional methods, Opponents_selection and Reward_assignment. In the evolution process, after finishing the calculation of the raw fitness value for each individual in the population, the Opponents_selection method will randomly select individuals as the opponents based on the $K$-Random Opponents strategy [5] from the same population without repeating the identical opponents and prohibits self-play. The $K$ is tested with the values of 20, 40, 60, and 80. After that, each individual will compete against the entire set of opponents. During the tournament, the reward value will be calculated for each competition based on the reward function. The reward value will be summed up as the fitness score for the individual up to the $K$ number of competitions, using the Reward_assignment method. Below is the description of the reward function. $I$ represent the participating individual, while $O$ represents the opponent. $R$ is the raw fitness value, max ($R$) is the maximum raw fitness value and the min ($R$) is the minimum raw fitness value. The range for values in this function is within [-1, 1]. If $Reward(I, O) = 0$, it corresponds to the competition as draw.

\[
Reward(I, O) = \frac{R(O) - R(I)}{max(R) - min(R)}
\]  

### 3.0 PERFORMANCES METRICS

**Generational Distance ($GD$)**: This metric was proposed by Van Veldhuizen and Lamont [1] which is used for estimating how far the elements in the Pareto front obtained are from the true Pareto front of the problem. This metric is defined as:

\[
GD = \frac{\sum_{i=1}^{n} d_i^2}{n}
\]

where $n$ is the number of non-dominated vectors found by the algorithm being analyzed and $d_i$ is the Euclidean distance (measured in objective space) between each of these and the nearest member of the true Pareto front. A value of $GD = 0$ indicates that all the elements generated are in the true Pareto front of the problem. Therefore, any other value will indicate how “far” the obtained solutions are from the global Pareto front of our problem.

**Coverage Metric ($C$)**: This metric was proposed by Zitzler et al. in [2]. By using this metric, two set of nondominated solutions can be compared to each other. Consider $X'$, $X^*$ as two sets of phenotype decision vectors. $C$ is defined as the mapping of the order pair ($X', X^*$) to the interval $[0, 1]$: 

\[
C(X', X^*) = \frac{|\{a' \in X'; \exists a' \in X^*: a' \preceq a\}|}{X^*}
\]

If the value $C(X', X^*) = 1$ means that all the decision vectors in $X^*$ are dominated by $X'$. Otherwise, if value $C(X', X^*) = 0$ represent the situation when none of the points in $X^*$ are dominated by $X'$.

### 6.0 EXPERIMENTAL RESULTS AND DISCUSSION

The performance between SPEA2-CE-KR is compared against SPEA2. In order to have a fair comparison, all runs considered are implemented with the same real-valued representation, simulated binary crossover (SBX), polynomial mutation and tournament selection. Also, the number of evaluations in each run is fixed at 60,000. Table 1 lists all the parameter settings for each evolutionary multiobjective optimization algorithm.
According to the literature review, comparing multiobjective optimization algorithms against each other can be hard. One would like an algorithm to minimize the distance of the Pareto front obtained with respect to true Pareto front and provide the solution set coverage. Therefore, comparisons become multiobjective optimization problems themselves. Can coevolution help to improve the performance of evolutionary multiobjective? Or can K-Random competitive coevolution assist to enhance the SPEA2 optimization performance? With this in mind we illustrate the results.

The results for test problems from DTLZ1 to DTLZ7 with respect to GD are summarized in Table 2 to Table 8. The graphical presentations in box plots of this metric are shown in Fig. 2. The leftmost box plot relates to SPEA2 while from the second left box plot to the rightmost box plot relate to SPEA2-CE-KR. The dark dash is the median, the top of the box is the upper quartile, and the bottom of the box is the lower quartile. The graphical presentations for the test problems in terms of the coverage metric are shown in Fig. 3 to Fig. 5. Each rectangle contains seven box plots representing the distribution of the C value; the leftmost box plot relates to DTLZ1 while the rightmost box plot relates to DTLZ7. The scale is 0 (no coverage) at the bottom and 1 (total coverage) at the top per rectangle.

Generational distance (GD): The nondominated solutions found by SPEA2-CE-KR have an excellent convergence that SPEA2. SPEA2-CE-KR enhances the GD values in all the DTLZ test problems with 3 to 5 objectives. The entire test problems depict noticeable improvements when the number of objectives is increased. It means that SPEA2-CE-KR has a far improved capability to escape from sub-optimal local solutions in its exploration process for a more optimal solution. Based on these encouraging results, it would appear that the proposed algorithm will outperform the standard algorithm even more as the multiobjective problem increases in its number of optimization objectives. However, this would only be verifiable in future work when the proposed algorithm is tested on problems with much higher number of objectives.

Coverage (C): SPEA2-CE-KR shows regular coverage of the nondominated solutions for the DTLZ test problems with 3 objectives. Based on the 4 and 5 objectives problems, coverage improved when compared against SPEA2 except for DTLZ5. Hence, SPEA2-CE-KR is superior to SPEA2 when the number of objectives increases.

Table 1: Parameter settings

<table>
<thead>
<tr>
<th>Parameter</th>
<th>SPEA2 / SPEA2-CE-KR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population size</td>
<td>100</td>
</tr>
<tr>
<td>Number of decision variables</td>
<td>12</td>
</tr>
<tr>
<td>Number of objectives</td>
<td>3 to 5</td>
</tr>
<tr>
<td>Number of generations</td>
<td>600</td>
</tr>
<tr>
<td>Mutation probability</td>
<td>0.08</td>
</tr>
<tr>
<td>Crossover probability</td>
<td>1</td>
</tr>
<tr>
<td>Polynomial mutation operator</td>
<td>20</td>
</tr>
<tr>
<td>SBX crossover operator</td>
<td>15</td>
</tr>
<tr>
<td>Number of repeated runs</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 2: Generational distance for DTLZ1 test suite with 3 to 5 objectives

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>3 Objectives</th>
<th>4 Objectives</th>
<th>5 Objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>St Dev</td>
<td>Mean</td>
</tr>
<tr>
<td>SPEA2</td>
<td>9.692</td>
<td>2.987</td>
<td>33.74</td>
</tr>
<tr>
<td>SPEA2-CE-KR-20</td>
<td>1.619</td>
<td>0.818</td>
<td>1.817</td>
</tr>
<tr>
<td>SPEA2-CE-KR-40</td>
<td>1.619</td>
<td>2.002</td>
<td>1.439</td>
</tr>
<tr>
<td>SPEA2-CE-KR-60</td>
<td>1.114</td>
<td>0.868</td>
<td>1.301</td>
</tr>
<tr>
<td>SPEA2-CE-KR-80</td>
<td>1.046</td>
<td>1.114</td>
<td>1.113</td>
</tr>
</tbody>
</table>

Table 3: Generational distance for DTLZ2 test suite with 3 to 5 objectives

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>3 Objectives</th>
<th>4 Objectives</th>
<th>5 Objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>St Dev</td>
<td>Mean</td>
</tr>
<tr>
<td>SPEA2</td>
<td>0.003689</td>
<td>0.001755</td>
<td>0.06290</td>
</tr>
<tr>
<td>SPEA2-CE-KR-20</td>
<td>0.002051</td>
<td>0.00193</td>
<td>0.006257</td>
</tr>
<tr>
<td>SPEA2-CE-KR-40</td>
<td>0.002036</td>
<td>0.00265</td>
<td>0.006388</td>
</tr>
<tr>
<td>SPEA2-CE-KR-60</td>
<td>0.001942</td>
<td>0.00224</td>
<td>0.006105</td>
</tr>
<tr>
<td>SPEA2-CE-KR-80</td>
<td>0.001964</td>
<td>0.00227</td>
<td>0.006164</td>
</tr>
</tbody>
</table>
Table 4: Generational distance for DTLZ3 test suite with 3 to 5 objectives

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>3 Objectives</th>
<th></th>
<th>4 Objectives</th>
<th></th>
<th>5 Objectives</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>St Dev</td>
<td>Mean</td>
<td>St Dev</td>
<td>Mean</td>
<td>St Dev</td>
</tr>
<tr>
<td>SPEA2</td>
<td>19.110</td>
<td>5.381</td>
<td>73.00</td>
<td>7.87</td>
<td>118.61</td>
<td>5.70</td>
</tr>
</tbody>
</table>

Table 5: Generational distance for DTLZ4 test suite with 3 to 5 objectives

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>3 Objectives</th>
<th></th>
<th>4 Objectives</th>
<th></th>
<th>5 Objectives</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>St Dev</td>
<td>Mean</td>
<td>St Dev</td>
<td>Mean</td>
<td>St Dev</td>
</tr>
<tr>
<td>SPEA2</td>
<td>0.003235</td>
<td>0.001623</td>
<td>0.04741</td>
<td>0.02913</td>
<td>0.12583</td>
<td>0.03894</td>
</tr>
<tr>
<td>SPEA2-CE-KR-20</td>
<td>0.001778</td>
<td>0.000435</td>
<td>0.006551</td>
<td>0.001080</td>
<td>0.016043</td>
<td>0.002673</td>
</tr>
<tr>
<td>SPEA2-CE-KR-40</td>
<td>0.001758</td>
<td>0.000395</td>
<td>0.006022</td>
<td>0.001062</td>
<td>0.015638</td>
<td>0.003222</td>
</tr>
<tr>
<td>SPEA2-CE-KR-60</td>
<td>0.001799</td>
<td>0.000389</td>
<td>0.006244</td>
<td>0.001189</td>
<td>0.015299</td>
<td>0.002971</td>
</tr>
<tr>
<td>SPEA2-CE-KR-80</td>
<td>0.001903</td>
<td>0.000336</td>
<td>0.006196</td>
<td>0.001285</td>
<td>0.015659</td>
<td>0.003024</td>
</tr>
</tbody>
</table>

Table 6: Generational distance for DTLZ5 test suite with 3 to 5 objectives

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>3 Objectives</th>
<th></th>
<th>4 Objectives</th>
<th></th>
<th>5 Objectives</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>St Dev</td>
<td>Mean</td>
<td>St Dev</td>
<td>Mean</td>
<td>St Dev</td>
</tr>
<tr>
<td>SPEA2</td>
<td>0.000378</td>
<td>0.000105</td>
<td>0.13761</td>
<td>0.00387</td>
<td>0.16845</td>
<td>0.00555</td>
</tr>
<tr>
<td>SPEA2-CE-KR-20</td>
<td>0.000407</td>
<td>0.000081</td>
<td>0.05504</td>
<td>0.00894</td>
<td>0.13182</td>
<td>0.01260</td>
</tr>
<tr>
<td>SPEA2-CE-KR-40</td>
<td>0.000356</td>
<td>0.000067</td>
<td>0.08016</td>
<td>0.00397</td>
<td>0.12874</td>
<td>0.01382</td>
</tr>
<tr>
<td>SPEA2-CE-KR-60</td>
<td>0.000349</td>
<td>0.000065</td>
<td>0.07803</td>
<td>0.03857</td>
<td>0.12383</td>
<td>0.01428</td>
</tr>
<tr>
<td>SPEA2-CE-KR-80</td>
<td>0.000385</td>
<td>0.000151</td>
<td>0.08389</td>
<td>0.03445</td>
<td>0.12423</td>
<td>0.01686</td>
</tr>
</tbody>
</table>

Table 7: Generational distance for DTLZ6 test suite with 3 to 5 objectives

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>3 Objectives</th>
<th></th>
<th>4 Objectives</th>
<th></th>
<th>5 Objectives</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>St Dev</td>
<td>Mean</td>
<td>St Dev</td>
<td>Mean</td>
<td>St Dev</td>
</tr>
<tr>
<td>SPEA2</td>
<td>0.10545</td>
<td>0.03423</td>
<td>0.51659</td>
<td>0.04963</td>
<td>0.80639</td>
<td>0.00566</td>
</tr>
<tr>
<td>SPEA2-CE-KR-20</td>
<td>0.014478</td>
<td>0.001036</td>
<td>0.07869</td>
<td>0.00946</td>
<td>0.4851</td>
<td>0.1509</td>
</tr>
<tr>
<td>SPEA2-CE-KR-40</td>
<td>0.014515</td>
<td>0.001231</td>
<td>0.07024</td>
<td>0.00901</td>
<td>0.3746</td>
<td>0.0959</td>
</tr>
<tr>
<td>SPEA2-CE-KR-60</td>
<td>0.013800</td>
<td>0.001118</td>
<td>0.06854</td>
<td>0.00793</td>
<td>0.3406</td>
<td>0.1393</td>
</tr>
<tr>
<td>SPEA2-CE-KR-80</td>
<td>0.014164</td>
<td>0.001340</td>
<td>0.06787</td>
<td>0.00931</td>
<td>0.3279</td>
<td>0.1243</td>
</tr>
</tbody>
</table>

Table 8: Generational distance for DTLZ7 test suite with 3 to 5 objectives

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>3 Objectives</th>
<th></th>
<th>4 Objectives</th>
<th></th>
<th>5 Objectives</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>St Dev</td>
<td>Mean</td>
<td>St Dev</td>
<td>Mean</td>
<td>St Dev</td>
</tr>
<tr>
<td>SPEA2</td>
<td>0.02163</td>
<td>0.01589</td>
<td>0.6708</td>
<td>0.1387</td>
<td>1.5718</td>
<td>0.1486</td>
</tr>
<tr>
<td>SPEA2-CE-KR-20</td>
<td>0.001646</td>
<td>0.000846</td>
<td>0.00505</td>
<td>0.000789</td>
<td>0.02187</td>
<td>0.00909</td>
</tr>
<tr>
<td>SPEA2-CE-KR-40</td>
<td>0.002002</td>
<td>0.001753</td>
<td>0.00538</td>
<td>0.001250</td>
<td>0.01889</td>
<td>0.00922</td>
</tr>
<tr>
<td>SPEA2-CE-KR-60</td>
<td>0.002886</td>
<td>0.002978</td>
<td>0.00534</td>
<td>0.000761</td>
<td>0.02153</td>
<td>0.00988</td>
</tr>
<tr>
<td>SPEA2-CE-KR-80</td>
<td>0.001952</td>
<td>0.001602</td>
<td>0.005124</td>
<td>0.000923</td>
<td>0.02152</td>
<td>0.00988</td>
</tr>
</tbody>
</table>

GD
Fig. 2: Box plots for generational distance ($GD$)

Fig. 3: Box plots of function $C(X, Y)$ indices for each pair of SPEA2 and SPEA2-CE-KR with 3 objectives

Fig. 4: Box plots of function $C(X, Y)$ indices for each pair of SPEA2 and SPEA2-CE-KR with 4 objectives
5 Objectives

\[ C(\text{SPEA2-CE-KR, SPEA2}) \]

Opponents Size \[ C(\text{SPEA2, SPEA2-CE-KR}) \]

20

40

60

80

Fig. 5: Box plots of function \( C(X, Y) \) indices for each pair of SPEA2 and SPEA2-CE-KR with 5 objectives

7.0 CONCLUSION AND FUTURE WORK

SPEA2-CE-KR is presented in this paper. This algorithm is integration between SPEA2 and K-Random competitive coevolution method. SPEA2-CE-KR is benchmarked against the original SPEA2 using seven scalable DTLZ test problems with three to five objectives. The proposed algorithm has been found to be a very successful optimization approach in multiobjective problems compared with the SPEA2 based on the results obtained from generational distance and coverage metrics. As future work, it would be interesting to examine whether SPEA2-CE using different competitive fitness strategies would be able to further improve the performance of MOEAs hybridized with coevolutionary techniques. Also, the proposed algorithm should be tested on problems with much higher number of objectives to further verify its vastly improved performance over the original algorithm.

ACKNOWLEDGMENTS

The first author is supported by a National Science Fellowship (NSF) from the Ministry of Science, Technology and Innovation (MOSTI), Malaysia.

REFERENCES


BIOGRAPHY

Tse Guan Tan was born in Kuala Lumpur, Malaysia, on February 10, 1982. He received the Bachelor of Computer Science (Software Engineering) from Universiti Malaysia Sabah, Malaysia, in 2006. He is working toward a Master’s degree in artificial intelligence at Universiti Malaysia Sabah, Malaysia.

Jason Teo is a senior lecturer in computer science at the School of Engineering and Information Technology and the deputy director of the Centre for Artificial Intelligence, Universiti Malaysia Sabah. He received his doctorate in information technology from the University of New South Wales, Australia, researching Pareto artificial evolution of virtual legged organisms. He has over 70 publications in the areas of artificial life, evolutionary robotics, evolutionary computing and swarm intelligence. His current research interests focus on the theory and applications of evolutionary multiobjective optimization algorithms in co-evolution, robotics and metaheuristics.

Hui Keng Lau is a lecturer in School of Engineering and Information Technology in Universiti Malaysia Sabah. He attended Universiti Malaysia Sarawak where he obtained his BIT (Hons.) degree in 2001. In 2004, he obtained his MSc in Computer Science from Universiti Teknologi Malaysia. His research interests cover topics in neural networks, artificial immune systems, pattern recognition and AI techniques.
AN INVESTIGATION OF AN INTEGRATION OF INDIVIDUAL AND SOCIAL LEARNING IN AN EVOLUTIONARY APPROACH TO THE GAME OF TIC-TAC-TOE

Razali Yaakob1, Graham Kendall2
1Department of Computer Science, Faculty of Computer Science and IT, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor. Email: razaliy@fsktm.upm.edu.my
2School of Computer Science and IT, University of Nottingham, Jubilee Campus, Wollaton Road, NG8 1BB, Nottingham, United Kingdom. Email: gxk@cs.nott.ac.uk

ABSTRACT

In this paper, we investigate an integration of individual and social learning, utilising evolutionary neural networks. Individual learning takes place by playing against the nearly perfect player. Social learning allows poor performing players to learn from those players, which are playing at a higher level. There are two types of experiments in this work, i.e. the experiment with social learning and the experiment without the social learning. The feed forward neural networks are evolved via evolutionary strategies and no knowledge incorporate in this work. Our experiments show that learning the experiment with social learning statistically significant giving better result than the experiment without social learning.

Keywords: Evolutionary neural networks, individual learning, and social learning.

1. INTRODUCTION

Game playing, as a testbed for investigating artificial intelligence techniques has a long history, and some notable results have been achieved, e.g. Deep Blue in chess [1], Chinook in checkers [2], Victoria (Go-moku) [3], Logistello (Othello) [4], TD-Gammon [5] and Neurogammon [6] (Backgammon), and Connect-Four [7, 8, 9]. Deep Blue achieved world champion status when, in 1997, it beat Garry Kasparov [1]. Deep Blue utilised search algorithms to analyse up to 200 million positions per second by utilising custom-built hardware [10].

Chinook, developed by Jonathan Schaeffer’s team at The University of Alberta, won the world checkers title in 1994 [2]. Chinook used an opening and endgame database together with extensive checkers knowledge. In contrast, Fogel and Chellapilla developed a checkers program, which did not rely on human expert knowledge [11,12]. In [11, 12], the program learned checkers strategy using a co-evolutionary approach without utilising any pre-programmed knowledge. The neural networks play against themselves for a number of generations. At each generation, the players only receive points based on whether they have won, lost, or drawn. Without any expert knowledge, Fogel and Chellapilla have demonstrated that a program can learn to play a game and reach the level of a human expert. The work in here is inspired from [13], where the evolved neural networks learnt to play by playing against a nearly perfect player. The evolved neural networks always played first.

A learning methodology that does not rely on human expertise is also the aim of the work we present here. Our objective is to investigate an integration of individual and social learning in evolutionary neural networks for the game of Tic-Tac-Toe. A feed forward neural network player is played against the nearly perfect player, and an evolutionary strategy is used to evolve these networks. We call this individual learning. After a period of individual learning we allow the players to “learn” from one another. We call this social learning.

In real-life, humans have a variety of techniques in order to develop strategies to defeat other humans. Humans can improve their strategy by themselves or through the experience of competing against other humans. Humans can also copy strategies from a better player and develop their own strategy based on this copy. Fogel and Chellapilla showed in [11, 12] that an automated player can learn to play checkers by competing against other computer players, and in [13], the evolved player played against the nearly perfect player. However, none of these players copied their strategy from other, superior, players. In this work, we give an opportunity for a player to learn to play the game through its own experience and via the experience of others.

According to Vriend [14] and Tesfatsion [15], in the context of agent-based computational economics, in individual learning, the agents learn exclusively from its own experience, and in social learning, the agents learn from the experience of other agents.

©FCSIT, UM 2007
The techniques we present in this paper are motivated from [16] with some minor modifications. In [16], a simulated stock market uses co-evolving neural networks, which are integrated with individual and social learning. The results show that the artificial stock traders perform better than a baseline buy-and-hold strategy.

2. BACKGROUND

Some early work on evolutionary Tic-Tac-Toe was carried out by Fogel [13], where the evolved neural networks learnt to play by playing against a nearly perfect player. The nearly perfect player uses the following rules to play the game,

1. Receive the current state of the game as an input
2. With 0.1 probability make a random move, else
3. If a win block is available, place a marker in the winning square, else
4. If a block is available, place a marker in the blocking square, else,
5. If two open squares are in line with an "O", randomly place a marker in either of the two squares, else
6. Randomly move in any open square.

In [13], a single feedforward neural network was used to evolve the strategy with nine input and output nodes, where they represent the nine squares of the Tic-Tac-Toe. The number of hidden nodes will be chosen randomly between 1 and 10.

Social learning research has largely based on in the context of agent-based computational economics [16, 14] and according to [14], in individual learning, the agents learn exclusively from its own experience and in social learning, the agents learn from the experience of other agents. In this work, the player will learn individually based on experience by playing against a nearly perfect player and after a period of time, we allow the player to "learn" from one another in a process called social learning.

In a context of automated game playing, individual learning is where a player learns and creates a strategy by himself through experience by playing against other players. In this type of learning, the player never copies another strategy from other players or replaces its strategy with a new random strategy. In contrast, the idea of social learning is to give the player a chance to copy or create a new strategy. The player still have a chance to create it own strategy through individual learning and also if the strategy is not good enough or the player is not happy with the current strategy, they player can choose to either copy a better strategy or create a new random strategy.

2.1 Rules of Tic-Tac-Toe

Tic-Tac-Toe is played between two players, with the normal size board being a three-by-three grid. The game starts with empty grid, and players (usually referred to as 'X' and 'O') take alternate turns. Player 'X' always starts first. The game ends when either one of the players places three of its marker in a time (horizontally, vertically or diagonally, e.g. Fig. 1) or both players cannot achieve this goal, in which case the game is drawn.

```
X | X | O
O | X | O
X | O | O
```

Fig. 1: Tic-Tic-Toe: Showing winning positions for 'X'

3. METHODOLOGY

In this work, the evolved players will receive points based on whether they win, lose or draw (i.e. +1, -1 and 0 respectively), when played against the nearly perfect player. This point is used as the fitness (or individual score), $F$, for each evolved player. The experiment starts by instantiating population of 50 players. The feed forward neural networks are used with nine input nodes, nine hidden nodes and an output node. The hyperbolic tangent is used as the evaluation function. Each player is represented as a neural network with the weights, initially, assigned random values. All these players play against the nearly perfect player and at each generation, the worst 25 players will be removed and replaced with a mutation (offspring) of the best 25 players. The players outside of the social pool are called individual player, where all these players try to develop their own strategy and the players in the social pool are drawn from the best players at the specific times. There are two types of experiments in this work, i.e. an experiment without social learning

---

1 A beatable player where 10% random moves.
and an experiment with social learning. All weights in the neural network are mutated at each generation using an evolutionary strategy. Each weight has an associated self-adaptive parameter, \( \sigma \), which serves to control the step size of the search for the mutated parameters of the neural network \[11\]. Initially, a Gaussian random generator is used to generate all the weights. The self-adaptive parameter and the weights are then adapted by

\[
\sigma'_j = \sigma_j \cdot \exp(\tau N_j(0,1)) \\
w'_j = w_j + \sigma'_j \cdot N_j(0,1)
\]  

(1)  

(2)

The \( w_j \) is a weight in the neural network, \( j = 0, 1, ..., N_w \), where \( N_w \) is a number of weights in the neural network. The \( N_j(0,1) \) is a standard Gaussian random variable resampled anew for every \( j \) and \( \tau \) is a learning rate, which is \( \tau = (2N_w^{0.5})^{-0.5} \). In fact, formula (1) and formula (2) are taken from \[11, 12\].

### 3.1 Social Learning

In this approach, we create a pool to keep the best strategy at a certain period (called the social pool). This concept is quite similar with the concept of hall of fame, introduced by \[17\], where all the best evolved players at every generation are kept and can be used for future testing to see the progress of the learning process. During social learning, the player can choose to replace their current strategy with a (hopefully) better strategy, drawn from the social pool. All strategies in the social pool have their own score, which is updated through time. The activities in social learning are as follows:

1. Rank the players from the highest to the lowest.
2. Keep the best player or players (if more than one) in the social pool.
3. For the rest of the players, there are two possibilities,
   i. If the players are satisfied with their current strategy,
      - keep the current strategy, else
   ii. If the players are not satisfied with their current strategy,
       - three alternatives are available,
       a. Copy a strategy from the pool, or
       b. Create a new random strategy, or
       c. Remain with the current strategy.

### Experiment without Social Learning

The structure of the experiment without social learning (also known as an experiment with individual learning) will follow the work presented in \[13\], with some modifications. This experiment will be used as a comparison for the experiment with social learning. The difference between the work in \[13\] and the initial experiment conducted here are as follows:

1. Neural network architecture, where the revised architecture consists of nine input nodes, nine hidden nodes and an output node. All hidden nodes are connected with bias. In \[13\], the neural networks have nine input and output nodes, and the number hidden nodes are depends on the mutation during the experiment.
2. The architecture in \[13\] is mutated during the experiment, whereas, there are no mutation activities on the architecture of neural networks.
3. In \[13\], the neural network chooses a best move based on the highest value from one of the nine output nodes from the neural network. However, in this work, the evolved neural networks will search for the best move using 1-ply search using the following algorithm:
   i. Read the current state of the game.
   ii. Put a marker ‘X’ on any available square.
   iii. Pass the new current state to the neural network.
   iv. Store the real value from the output.
   v. Remove the marker placed in step (ii) from the square.
   vi. Repeat steps (ii) to (v) until all available squares have been evaluated by the neural network and has an associated real value stored.
4. Select the highest output as the best move.
**Experiment with Social Learning**

In the experiment with social learning, social learning activities occurred at every 100\textsuperscript{th} generation. The algorithm in social learning as follows:

1. Normalise all the 50 evolved players, $i$, between 0.0 and 1.0 based on their individual fitness, $F_i$, using Eq. (3) and sort in descending order.

   \[
   V_i = \text{MIN} + (\text{MAX} - \text{MIN}) \times (F_i - d_{\text{min}}) / (d_{\text{max}} - d_{\text{min}})
   \]  
   \[\text{(3)}\]

   where,
   - $V_i$ is the normalised value for player $i$.
   - MIN and MAX is the lowest and highest value for range of the normalised value (0.0 and 1.0 respectively).
   - $F_i$ is the fitness of player $i$ before normalised.
   - $d_{\text{min}}$ and $d_{\text{max}}$ is the lowest and highest score among all players.

2. If ($V_i = 1.0$) and the strategy has never been published,
   - Publish the player $i$ into the social pool and assign pool score, $P_j$, using Eq. (4), where $j$ is an index of the best player in the social pool.

3. If ($V_i = 1.0$) and the strategy has previously been published to the social pool,
   - Do not publish the player $i$ but update the pool, score, $P_j$ using the Eq. (4).

4. If ($V_i \geq 0.9$) and ($V_i < 1.0$),
   - The player $i$ is satisfied with their current strategy and the player stays in the population for the next generation.

5. If ($V_i \leq 0.9$),
   - There are three probabilities ($p = 0, 1$ or $2$), which are chosen randomly:
    i. If ($p = 0$),
       - The player chooses to replace their current strategy with a strategy from the social pool.
       - Based on the scores assigned to each player in the pool, $P_j$, roulette wheel selection is used to select a strategy from the social pool.
    ii. If ($p = 1$),
       - The player chooses to remain with their current strategy for the next generation.
    iii. If ($p = 2$),
       - The player chooses to replace their current strategy with a new randomly created strategy.

The pool score in the social pool for each best player, $P_j$, is updated every time social learning occurred using Eq. (4), where the value of the pool score is based on how long the player has been in the pool and what was its individual score, $F_i$ when it was published into the pool. The objective of this function is to give better players, which have recently been introduced (or updated) to the pool, a higher probability of being selected to be copied as replacements for poorly performing individual players. The age of the player is controlled by $g_j / \Sigma g$, where the previous best players will receive less points than the current best players. However, the final calculation for calculating the pool score are also dependent on the individual score of the best player when publishing into the social pool and how many times they been called back\textsuperscript{2} since they were first published.

\[
P = \frac{g_j}{\sum_{i=1}^{g} \exp(I_j + C_j)}
\]  
\[\text{(4)}\]

Where,
- $j$ is an index of the best player in the social pool,
- $g$ is a generation when published into the social pool,
- $g_j$ is a generation for player $j$ when publish into the social pool.

---

\textsuperscript{2} the strategy that has been published in the best pool and receives $V_i = 1$ on the next social learning.
- $I$ is normalised value of individual score, $F$,
- $C$ is a sum how many times player $j$ been called back.

The number of strategies can be copied from the social pool are limited to three times of the best player in the social pool. For example, if the social pool has three best players, total number of best strategies can be copied in the social pool are nine strategies. The objective is to control from the same strategy been copied so many times and to make sure that arms race continues over the generation.

4. RESULTS

All experiments were run for 1000 generations and 50 times. In the experiment with social learning, we introduce minor and major social learning. Minor social learning occurs at every 50th generation, and major social learning occurs at every 100th generation. In minor social learning, the best player(s) at that time will keep in the social pool, but no other social activities taking place at this time. The actual social activities occur in major social activities.

Fig. 2 shows the mean of a highest score (in percent) for individual player at each generation for the experiment with and without social learning (marked with WSL and NSL respectively on the graph). The y-axis is an average of a highest score for 50 trials at each generation. Based on the Fig. 2, at the end of the experiment without social learning, the highest score is at about 74% and 80% for the experiment with social learning.

![Fig. 2: Experiment with and without social learning](image)

Based on Fig. 2, the experiment with social learning is constantly better than the experiment without social learning from the beginning until the end with the difference between them is statistically significance at about over 95% confidence level.

5. CONCLUSIONS AND FUTURE WORK

Based on the results, some learning has occurred in both experiments even though no knowledge or information was given in the program. The programs also do not have any look ahead. The experiment with social learning constantly gives a better result than the experiment without social learning from the beginning of the experiment. The players in the experiment with social learning learn quicker to play the game and this show that the integrated of individual and social learning did help the learning process.

Evolving neural network has the potential as a technique for evolving game playing. Even though the experiment with social learning better than without social learning, the performance of the player still the same, where both of them can play at competent level. We believe that to have a good player, it has to come from a good population and that will be our focus in the future work.

REFERENCES


BIOGRAPHY

Razali Yaakob obtained his Master of Science (Computer Science) from Universiti Putra Malaysia in 1999. Currently, he is a lecturer at the Faculty of Computer Science and IT, Universiti Putra Malaysia. His research areas include artificial neural network, pattern recognition, and evolutionary computation in game playing.

Professor Graham Kendall is a member of Automated, Optimisation, Scheduling and Planning Group in School of Computer Science and IT at the University of Nottingham. His research interests include Meta and Hyper Heuristics, Evolutionary and Adaptive Computation, Artificial Life, Case Based Reasoning, and Scheduling. His web page is http://www.cs.nott.ac.uk/~gxk/
LEARNING REACTION MECHANISMS THROUGH QUALITATIVE SIMULATION: TOWARDS THE QUALITATIVE REASONING APPROACH

Alicia Tang¹, Sharifuddin Mohd. Zain², Noorsaadah Abd. Rahman² and Rukaini Abdullah³
¹Department of Computer Science, College of Information Technology, Universiti Tenaga Nasional, 43009 Kajang, Selangor, Malaysia. Email: aliciat@uniten.edu.my
²Department of Chemistry, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia. Email: {smzain, noorsaadah}@um.edu.my
³Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia. Email: rukaini@um.edu.my

ABSTRACT

This work discusses the application of an Artificial Intelligence (AI) technique called Qualitative Reasoning (QR) coupled with the Qualitative Process Theory (QPT) ontology to model, simulate and explain chemical behaviours of reaction mechanisms. We have tested the new approach on two types of organic mechanisms under nucleophilic substitution reaction. This paper describes one specific type of the mechanisms called S_N1 to demonstrate how the qualitative models can be constructed for chemical processes such as “make-bond” and “break-bond”. The construction of these models is automated based on a set of QR algorithms. Each qualitative model represents some aspects of the chemical theories that are required to understand the subject. Several cases of causal explanation generation are also included to demonstrate how natural behavioural explanation can be derived from the automatically constructed models. A prototype called QRIOM based on the QR technology will serve as cognitive tool fostering the acquisition of conceptual understanding when learning about reaction mechanisms.

Keywords: Qualitative Reasoning, Reaction Mechanisms, Models, Simulation, Causal Explanation.

1.0 INTRODUCTION

The Qualitative Reasoning (QR) community pursues research at the intersection of Artificial Intelligence, Cognitive Science, Engineering, and Science. The main aim of QR is to develop representation and reasoning techniques that will enable a program to reason without having precise quantitative data needed by traditional analysis such as numerical simulators [1]. People can make interesting predictions using only qualitative information about situations they have never encountered. The desire to capture such commonsense knowledge initiated research in QR, which explains why this research is also known as ‘naïve physics’. Qualitative Process Theory (QPT) [2] is one of the prominent QR ontology that is widely used to represent the behaviour of dynamical systems. The work described in this paper was modelled using QPT. Among other well-known QR ontology are component-centred [3], and constraint-based [4]. Education is an important practical application areas of QR. CyclePad [5] is a system that teaches analysis and design of thermal cycles using QPT. Other representatives include a model-based reasoning framework for high school level mathematics Neuper and Wotawa [6], CPRODS [7], and QALSIC [8, 9]. Common features in these systems are the ability to predict and explain the behaviour of physical systems in qualitative terms. Even though there are many applications of Artificial Intelligence (AI) techniques in organic chemistry and reaction mechanisms but none has involved QR approach in their problem solving. On the other hand, a reaction mechanism describes a step-by-step sequence of reactions by which overall chemical change occurs, going from the starting material or substrate to the final product and these chemical changes occur by specific routes. Most of the time, the organic chemists could work out the mechanisms by only using commonsense developed from their chemical intuition and knowledge. In this scenario, the chemists are doing organic synthesis by following the mechanisms they proposed. This is a very suitable field for QR as commonsense reasoning and qualitative description are necessitated.

2.0 OBJECTIVES

Our past survey showed that a large number of chemistry students had difficulty in understanding reaction mechanisms such as the S_N1 (unimolecular nucleophilic substitution) and the S_N2 (bimolecular nucleophilic substitution). They learn the subject by memorizing the basic facts, steps and formulas of each reaction which are easily forgotten. Poor conceptual understanding about the cause effect interaction that exists in reactions is the main reason why most students are unable to solve new or more complex problems. The simulator provides a user-friendly learning environment that assists learners in understanding the ‘How’, ‘Why’, ‘Why-not’, and ‘What’ aspects of the general principles of organic
reactions. There have been many strives for innovation in teaching and learning chemistry using software. Most of the
techniques are unable to consider intention and goal in their delivery. We believe that by studying the cause effect chain
of a chemical phenomenon via a process-based ontology such as QPT can serve as an alternative technology in learning
of chemistry subject such as reaction mechanisms. The research paper will describe the design and methodology, and
present the results of the application of qualitative reasoning and simulation to the facilitation of learning with the new
approach.

3.0 PREVIOUS WORKS

We have reported in [10] the modelling decisions and problems faced when trying to cast the expert knowledge into
qualitative models using QPT. In [11], we justified the problem as a suitable domain by comparing inorganic chemical
reactions and organic reaction mechanisms. The paper also explained chemical behaviours through cause effect
inspection via QPT modelling constructs. We have also identified “make-bond” and “break-bond” as reusable components
that can be extended to modelling other reaction mechanisms such as the electrophilic addition and elimination mechanisms [12]. In the work, we have also grouped all reacting species as either a nucleophile (charged/neutral) or an electrophile (charged/neutral); upon which chemical processes are selected.

8.0 QUALITATIVE PROCESS THEORY (QPT)

QPT provides the means to embody notions of causality which is important to explain behaviour of chemical systems.
QPT provides a number of modelling constructs (design primitives) for representing qualitative knowledge, and the
notion of processes needed in expressing and modelling chemical reaction steps. In QPT, a structural description of the
model is given by a set of Individual Views and Processes. The individual views describe objects and their general
characteristics while the processes support changes in system behaviour. A process is described by five slots:
Individuals, Preconditions, Quantity-conditions, Relations and Influences. Quantity-conditions contain inequalities
involving quantities (object’s characteristics, e.g. charge, no. of covalent bond), which is crucial in determining the
status of a process (active/inactive). Relations are statements about functional dependencies among quantities. Two
important design primitives for describing the relationships between quantities are the correspondences and qualitative
proportionalities. Qualitative proportionalities (denoted by P+/P-) propagate the effects of processes that express
unknown monotonic functions (increasing/decreasing/unchanged) between two quantities. The main ideas of the
reaction of nucleophiles with electrophiles are modelled using this construct. In this formalism, dynamic aspects are
expressed by the notion of direct influence, that can only appear in QPT processes, represented in the slot called Influences as either I+ or I-. Note: words typed in italics are QPT modelling constructs. Further discussion on QPT is not the focus of this paper. Readers may refer to [2] for a complete reading of the ontology.

9.0 METHODS

In developing the simulator called Qualitative Reasoning in Organic Mechanism (QRIOM), the system methodology
used can be divided into a number of main activities, as follows:
• Chemical properties identification for chemical processes in order to cast them into QPT process model
• Establishment of functional dependencies among the chemical parameters (covalent bonds, charges, etc.) of objects
  (substrates, reagents, etc.) using the ontological modelling constructs
• Automate QPT processes construction (E.g. “make-bond” and “break-bond” based on the substrate)
• QR algorithms development
• Qualitative simulation based on the algorithms
• Causal reasoning to generate explanation on demand

9.1 Chemical Behaviour Abstraction

We will use Equation 1 to simulate the behaviour of SN1. When SN1 is used to explain the production of alkyl halide, in
the first stage, the alcohol oxygen (the \( O \) from the \( OH \) group) is protonated. This is to make the \( O^+H^2 \) a better leaving
group. Once broken, a tertiary carbocation is produced. In the second stage, the incoming nucleophile (\( Cl^- \)) can bond to
the carbocation to form a neutral and stable final product [13].

\[(\text{CH}_3)_2\text{C} = \text{OH} + \text{HCl} \rightarrow (\text{CH}_3)_2\text{C} = \text{Cl} + \text{H}_2\text{O}\]

.. Equation 1
To benefit readers of non-chemistry background, the thought processes for Equation 1 is explicitly stated as a series of steps as follows.

Step 1: **Protonation** of tert-butyl alcohol to produce an oxonium ion. This is a “make-bond” process.

\[
\begin{align*}
(CH_3)_3C - O: & \quad + \quad H^+ - Cl^- \\
\text{tert-butyl alcohol} & \quad \leftrightarrow \\
\text{hydrogen chloride} & \quad \text{tert-butyloxonium ion} \\
\end{align*}
\]

Step 2: **Dissociation** of tert-butyloxonium ion to give a carbocation intermediate. This is a “break-bond” process.

\[
\begin{align*}
(CH_3)_3C^+ & \quad + \quad \text{H}^+ \\
\text{tert-butyloxonium} & \quad \leftrightarrow \\
\text{tert-butyl cation} & \quad \text{water} \\
\end{align*}
\]

Step 3: **Capturing of tert-butyl cation** by chloride ion to produce tert-butyl chloride. This is a “make-bond” process.

\[
\begin{align*}
(CH_3)_3C^+ & \quad + \quad Cl^- \\
\text{tert-butyl cation} & \quad \rightarrow \\
\text{chloride ion} & \quad \text{tert-butyl chloride} \\
\end{align*}
\]

We have coded the general principles sought in the properties identification stage as chemical theories via the qualitative proportionality modelling construct of QPT.

### 9.2 Qualitative Model Development

Model development involves the mapping of general chemical properties into ontological primitives. The casting of the expert knowledge into a “make-bond” process called “protonation” is described in next section. Note: Chemical processes are automatically generated by the Qualitative Model Constructor (module 2 in Fig. 1) of the QRIOM.
9.3 Instantiation of the “Protonation” Chemical Process

Protonation is a “make-bond” process that necessitates a proton (an electrophile) and a nucleophile which has lone pair of electrons to be donated. This information is cast into the Quantity-Conditions (Lines 6–7 in Fig. 2). “Protonation” is normally done on poor leaving group such as the hydroxyl functional group (OH). QPT provides Precondition slot to hold information that remains valid throughout a reaction. In this case, it is defined as leaving_group (OH, poor) (Line 5). Through chemical intuition, the process will form a covalent bond between the O and the $H^+$; this relationship is modelled as direct changes, placed into the Influence slot (Line 17). Changes will then propagate to other dependent quantities via qualitative proportionality construct (Lines 13–16) that defined functional dependencies among chemical parameters of reacting species. When the above line numbers are put together, the qualitative model in Fig. 2 is composed. Chemical processes for the other reaction steps can also be constructed using the same approach. Qualitative proportionality also forms the basis for causal explanation needed at a later stage. Obviously, a large number of chemical facts are required to support, and these are stored as chemical facts and chemical theories. Discussion about chemical knowledge base component is beyond the scope of this paper. We have developed algorithms for constructing QPT models based on the properties of the reacting species (nucleophile/electrophile). As such, all complexities described above are hidden from the users of QRIOM. Qualitative models can be inspected by learners at any stage of a learning process. This helps to sharpen a learner’s logical and critical thinking in the way that the learner has to think hard for why the statements in each slot are relevant or negligible. Learning using qualitative models and causal explanation manifestation will be given in the Section 6 and Section 7.

10.0 LEARNING WITH QUALITATIVE MODELS

Good reasoning skill is required when performing model inspection. In this section we will show how a qualitative model can help to articulate ideas about a learning task. We divided the model inspection activity as a series of learning assignments. A scenario of learning the QPT process in Fig. 2 is as follows. Note: Line numbers are based on the enumeration used in the figure.

- **Learning task 1**: The $H^+$ from $\text{HCl}$ is needed by the reaction formula in Equation 1. Learners would be able to find this by inspecting the Individuals slot. Briefly, the slot says that, in order to begin the first chemical process, there needs a proton ($H^+$, in our example) which serves as an electrophile together with a species which is nucleophilic. In this case, the nucleophile is the $O$ from the $\text{OH}$ group (termed as alcohol oxygen) which has lone-pair electron to be donated. Line 1 and Line 2 show exactly the existence of hydrogen ion together with the $\text{OH}$ functional group from the alcohol which helped to explain why the a “make-bond” process is activated.

- **Learning task 2**: Lines 3, 4 and 5 collectively say that the number of covalent bond at $O$ is two; $(\text{CH}_3)_3\text{C-OH}$ is reactive and $\text{OH}$ is a poor leaving group. These are basic information about properties that remain true throughout a reaction for the involved substances.

- **Learning task 3**: The inequality lone-pair-electron $\geq$ min-electron-pair for the $O$ from oxygen alcohol speaks for “there is at least one pair of lone-pair electron to donate to $H^+$”. This indirectly ascertains that a covalent bond could be made between the two species.

- **Learning task 4**: By using chemical intuition, when the chemical process begins, the $O$ will have an extra covalent bond. That is to say, there will be a bond activity transpired (in this case, it is a “make-bond” process). This is defined in Line 17 via the direct influence slot of QPT.

11.0 CAUSE EFFECT INSPECTION

A causal diagram is a representation of the cause-effect relationships of a model. For example, when given ‘X causes Y’, we believe that if we want to obtain Y we would bring about X. As such, when we observe Y we will think that X might be the reason for it. We will demonstrate how the ontological design primitives of QPT can provide this nature of explanation which is casual about chemical behaviour.

When “protonation” process (Fig. 2) is described in qualitative terms, the notion of causality is visible. A causal diagram (Fig. 3) for the said process is included to explicitly represent the cause effect notion in the process.
We included the following two qualitative proportionalities (abstracted from the left branch of Fig. 3) in order to manifest the explanation generation ability of the QR approach:

\[
\begin{align*}
\text{lone-pair-electron(O)} & \quad \text{P}^+ \quad \text{no-of-bond(O)} \quad \text{.. qp1} \\
\text{charges(O)} & \quad \text{P}^- \quad \text{lone-pair-electron(O)} \quad \text{.. qp2}
\end{align*}
\]

Based on the above, a set of hypothetical Q & A can be devised as follows:

Q1: How would the above qualitative proportionalities explain the ‘O’ becoming positive charge?
A1: The number of lone-pair electron will decrease when more covalent bonds are made at O atom (via the inverse qualitative proportionality defined in qp1). In qp2, when the lone-pair electron of O decreases the charge of O will increase.

Q2: How would you explain a decrease in the lone-pair electron on the ‘O’?
A2: A possible conclusion would be “We know that the immediate cause of the process is the number of covalent bond of O increases. This quantity influenced the lone-pair electron of O, and the influence is strictly decreasing through the inverse proportionality relationship”.

The inspection of cause effect chain can help a learner to pick up the underlying concept better than merely memorizing the reaction steps or basic facts. QRIOM is able to provide this type of explanation on demand.

12.0 SIMULATION THROUGH QUALITATIVE REASONING

The top level of QR algorithm for reaction mechanism simulation is depicted in Fig. 4 (a) – Fig. 4 (c). By applying these algorithms, the simulator would be able to reproduce the behaviours of nucleophilic substitution reaction, and to explain why a particular final product is produced. The prediction of output is based on qualitative model reasoning, as opposed to the forward search method in traditional approach which needs large knowledge base and pre stored results for search used.
13.0 DISCUSSION AND CONCLUSION

Qualitative representation captures the intuitive, causal aspects of many human mental models. This new representation and computational approach, when applied to model construction and simulation through qualitative reasoning can help to enhance the critical thinking of learners. Causal models inspection can outperform traditional learning (such as reading from a textbook), in that their cognitive capabilities can be further developed. We have demonstrated the use of QR approach coupled with QPT ontology to represent the chemical intuition and knowledge in order to understand reaction mechanisms. The capability of the models in providing explanation has also been demonstrated in representative cases. The modules in QRIOM altogether served as embedded intelligence to the simulator. Qualitative reasoning using QPT as knowledge capture tool can lead to deeper and systematic understanding of chemical processes and phenomenon. We believe that organic reaction mechanisms (and other sub fields of organic chemistry) can serve as a test domain for further development of the QPT since the understanding of these subjects requires the application of knowledge at intuitive level.

REFERENCES


BIOGRAPHY

Alicia Tang is a lecturer at Computer Science Department, UNITEN. She obtained her Master of Science from The University of Manchester, UK. She is currently pursuing her PhD at University of Malaya in the field of Qualitative Reasoning applied to Reaction Mechanisms Simulation. She is a member of IEEE and IEEE Computer Society.

Sharifuddin Mohd. Zain is currently the Head of Chemistry Department at University of Malaya. His research interests include Computational Chemistry, Application of Computers in Chemical Research and Education, Artificial Intelligence in Chemistry, and Environmental Modelling.

Noorsaadah Abd. Rahman is a Professor at Chemistry Department, University of Malaya. She obtained her PhD from Cambridge University, UK. Her areas of specialization are Organic Synthesis and Synthetic Methods, Drug Design and Modelling.

Rukaini Abdullah obtained her PhD from University of Leeds, UK. She is attached to the Faculty of Computer Science and Information Technology, University of Malaya. Her research interests are Natural Language Processing, AI in Education and Computer Based Learning.
Precise Fingerprint Enrolment Through Projection Incorporated Subspace Based on Principal Component Analysis (PCA)

Md. Rajibul Islam, Md. Shohel Sayeed, Andrews Samraj
Faculty of Information Science and Technology (FIST)
Multimedia University, Jalan Ayer Keroh lama, 75450 Melaka, Malaysia
E-mail: {md.rajibul.islam05, shohel.sayeed, andrews.samraj}@mmu.edu.my

Abstract

Despite recent advances in the area of fingerprint identification, fingerprint enrolment continues to be a challenging pattern recognition problem. The first step to this problem is the enhancement of landmarks as well as precise minutiae points (ridge bifurcation and ridge ending), core, plain ridges from a print. Once enhanced, these fingerprint images are then ready to extract features and store into a database. Later these are compared to all sets on file in search of a match. The accurate fingerprint image is the basis for the entire identification and matching process. Various enhancement approaches have been proposed in the literature, each with its own merits and degree of success. The most common approach is to enhance and store the precise fingerprint image through normalization, orientation, frequencies calculation, contextual filtering and then binarisation and masking. Our emphasis in this paper is to enhance and store the fingerprint image accurately using Projection Incorporated Subspace based on Principal Component Analysis (PCA). In particular, we have implemented the methods based on eigenspace representations and neural network classifiers. Moreover, we present preliminary results of an attempt to mingle the outputs of these methods using a clustering algorithm unique to this type of problem.

Keywords: Fingerprint enrolment, Minutiae, Projection Incorporated Subspace, PCA, Region merging.

1.0 Introduction

Fingerprint identification is one of the most admired biometric technologies and is used in biometric personal identification, criminal investigations, commercial applications, and so on. The performance of a fingerprint image-matching algorithm depends heavily on the quality of the input fingerprint images [1]. It is very significant to acquire good quality images but in practice a significant percentage of acquired images are of poor quality due to some environmental factors or user’s body condition [2]. The poor quality images cause two problems: (1) many spurious minutiae may be created and (2) many genuine minutiae may be ignored. Therefore, a novel approach is necessary to increase the performance of the minutiae extraction algorithm.

In this paper, we have presented a subspace method based on Principle Component Analysis which incorporates the projection information of the fingerprint images. This method uses a subspace of lower dimension than that used by PCA. Also, its correct recognition rates are superior to PCA. The rest of this paper is structured as, section 2 briefly introduces the Projection Incorporated Subspace and PCA, section 3 the region merging technique (assume that section 2 and section3 are the proposed approach), section 4 presents some experimental results. Finally, section 5 concludes this article.

2.0 Projection Incorporated Subspace

Projection Incorporated Subspace has been described by Wu Jianxin, Chen Zhaoqian and Zhon Zhihua in [4]. The reason we use the projection incorporated subspace method that it requires less eigenvectors. Using fewer eigenvectors means that fewer computational power and processing time is needed. In real-world applications the fingerprint database may have thousands of individuals or even more, therefore the saving of computational cost may be quite significant.

Suppose $P(x, y)$ be an intensity image of size $N_1 \times N_2$, satisfies $x \in [1, N_1]$ , $y \in [1, N_2]$ , $P(x, y) \in [0,1]$. The vertical and horizontal integral projections are defined respectively as:

$$V_P(x) = \sum_{y=1}^{N_2} P(x, y) \quad (1)$$
$$H_P(y) = \sum_{x=1}^{N_1} P(x, y) \quad (2)$$

©FCSIT, UM 2007
Define the projection map \( M_p(x, y) \) for \( P(x, y) \) as
\[
M_p(x, y) = \frac{v_p(x) R_p(y)}{N_1 N_2 P} \tag{3}
\]
in which \( \overline{P} \) is the image's mean intensity, defined as
\[
\overline{P} = \frac{\sum_{x=1}^{N_1} \sum_{y=1}^{N_2} P(x, y)}{N_1 N_2} \tag{4}
\]
Then, a projection incorporated version of \( P(x, y) \) is defined as
\[
P_\alpha(x, y) = \frac{p(x, y) + \alpha M_p(x, y)}{1 + \alpha} \tag{5}
\]
in which \( \alpha \) is called combine parameter. Since \( P_\alpha(x, y) \) may go out of \([0, 1]\), exhibit of the fingerprint image may be imprecise although the recognition results will not be affected using projection map. For better display, the fingerprint image in Fig. 1 has been adjusted according to:
\[
P'_\alpha(x, y) = \frac{P_\alpha(x, y) - \min P_\alpha(x, y)}{\max P_\alpha(x, y) - \min P_\alpha(x, y)} \tag{6}
\]
In the above process, PCA are performed on the projection incorporated version of the fingerprint image instead of on the original image. We call the subspace find by this technique projection incorporated subspace method. And all the subspaces are shown by some boundary boxes (see Fig. 1).

![Fig. 1: Fingerprint image using projection incorporated subspace method based on PCA. Image (n) is the enhance part of image (m).](image)

### 2.1 Principal Component Analysis (PCA)

PCA is a useful statistical technique that has found application in fields such as fingerprint recognition and image compression, and is a common technique for finding patterns in data of high dimension [8]. The main effect of PCA is dimensionality reduction, that is, mapping the \( n \)-dimensional vector \( \mathbf{x} \) into an \( m \)-dimensional space, where \( m \ll n \). The vector \( \mathbf{x} \) can be approximated as a linear combination of a set of orthonormal vectors \( \mathbf{u}_i \),
\[
\hat{X} = \sum_{i=1}^{m} z_i \mathbf{u}_i \tag{7}
\]

### 3.0 REGION MERGING

In this section, we briefly present region merging technique researched by Hyun geun Yu [7]. Here a region merging post-processing phase is elucidated to reimburse for the over-segmentation problem. The most natural method to overcome the over-segmentation of watersheds transformation is to merge the small regions in a homogeneous region since they may possess certain homogeneous characteristics in intensity, texture or statistical properties. There is a traditional method for image segmentation, called Split/Merging. The Split/Merging method takes an intensity image as an input and splits it into small grids usually using quadtree structure (Fig. 2.0). Finally, the procedure merges small grids according to their statistical properties.
The region merging as post-processing for watersheds transformation takes a labeled image as input instead. This labeled image coincides with a quadtree of Split/Merging method (Fig. 2.1). The watersheds transformation algorithm processes the original image into a labeled image with boundary pixels; each label represents a different region. Two important keys for merging different regions together are:

1. If the regions are adjacent or not
2. How dissimilar/similar the regions are to each other.

3.1 Region adjacency graph

A simple label image is shown in Fig. 3.0. It can be transformed into a Region Adjacency Graph (RAG), to indicate whether two regions are adjacent or not; only adjacent regions are connected with bars.

Two regions that have minimum cost dissimilarity (a concept that will be explained in the section 3.2) are merged together if the dissimilarity satisfies a given criteria. Fig. 3.1 shows the merging step between regions a and b where it is assumed that those regions have minimum cost dissimilarity.

This merging step is repeated until there is no pair of regions satisfying the dissimilarity condition or the number of regions in repetition is the same as a given number of regions. Each merging step reduces the number of regions by 1. Based on conceptual analysis, the pseudo-code for region merging can be summarized. Given the RAG of the initial K-partition (K-RAG), the RAG of the suboptimal (K-n)-partition ((K-n)-RAG) is constructed by:

**Input:** RAG of the K-Partition (K-RAG)

**Iteration:** For $i = 0$ to $n - 1$

- Find the minimum cost edge in the (K-i)-RAG
- Merge the corresponding pair of regions to get the (K-i-1)-RAG
- Update RAG and dissimilarity

**Output:** RAG of the (K-n)-partition ((K-n)-RAG)

When region merging is applied to a labeled image being implemented in a program language, the image can be processed by first, scanning the image with a 3x3 window and the comparing labels in the window. Then, the table containing the RAG information is built in matrix form. This table (see Fig. 3.2) represents the status of adjacency with a flag (0=off, 1=on).
The dissimilarity cost is calculated only for the region pairs that have their RAG flags on.

### 3.2 Region dissimilarity function

Similarity between two regions can be simply described by the difference of statistical properties like average, variation or both of intensity values for each region. In this paper, the dissimilarity function defined as the equation below is used. This objective cost function is the square error of the piecewise constant approximation of the observed image, which yields a measure of the approximation accuracy and is defined over the space of partitions. If $R^*_M$ is the optimal $M$-partition with respect to the squared error, then the optimal $(M-1)$-partition is generated by merging the pair of regions of $R^*_M$, which minimizes the dissimilarity function.

$$
\delta(R^*_M, R^*_M) = \frac{||R^*_M||^2}{||R^*_M|| + ||R^*_M||} \left[\mu(R^*_M) - \mu(R^*_M)\right]^2 I(i, j)
$$

Where

$$
I(i, j) = \begin{cases} 
1, & \text{if regions } R^*_M, R^*_M \text{ are adjacent} \\
+\infty, & \text{otherwise}
\end{cases}
$$

According to the above formulation, the most similar pair of regions is the one minimizing square error.

The determination of the optimal number of segments $K^*$ is performed by checking the value of $\delta(\cdot, \cdot)$. If $\delta$ is greater than a certain threshold, then the merging process is terminated. This threshold value can be obtained through hypothesis testing on noise distribution; however, the desired number of regions can be simply given to stop the merging process if the threshold value is not certain.

In our experimental result shows the effect of region merging. The original image is processed with the watersheds transformation and region merging is applied to reduce the number of regions. In this case, the program stops the merging process if the average intensity difference between the optimal pair of regions being merged is greater than 12.

### 4.0 EXPERIMENTAL RESULTS

The principle of this research is to get the whole fingerprint image accurately for enrolment or store in the database and make them more suitable for the minutiae extraction algorithm. Especially during the enhancement from a poor fingerprint image if we fail to enhance some parts of a fingerprint image, we’ll definitely fail to get some minutiae information. By using our proposed approach we can overcome this problem. The ultimate criterion for evaluating such a proposed approach is the total amount of “quality” improvement when the experiment is applied to the poor input fingerprint images. Such an improvement can be assessed subjectively by a visual inspection of a number of typical experiment results. However, a precise consistent characterization of the quality improvement is beyond the capability of subjective evaluation. Examples of the experiment results are shown in Fig. 4.1 and Fig. 4.2.

To test the proposed approach, we use the FVC2002 database [5]. From this dataset, we randomly selected a total of 80 fingerprints, eight impressions per finger. Three impressions for one of the same finger in the database are shown in Fig. 4.0(a)(b)(c). In our first session we have used the code loosely follows the approach presented by Hong, L., Wan, Y., and Jain, A. K.[3] are shown in Fig. 4.0(d)(e)(f). We used Peter Kovesi’s implementation [6] for this purpose. In our
first experiment, only those images taken during the first session were used. The results obtained using the proposed Projection Incorporated subspace method based on PCA approaches are shown in Fig. 4.1 (g)(h)(i).

Fig. 4.0: (a), (b), (c) are the original fingerprints taken from the same person and (d), (e), (f) are obtained from (a), (b), (c) respectively by using the code loosely follows the approach presented by Hong, L., Wan, Y., and Jain, A. K.[3].

Fig. 4.1: Fingerprints which are obtained through proposed approach (PCA Subspace method).

After first experiment we combine all the subspaces by sorting and reconstruction but when we remove the boundary boxes from the fingerprint image we have seen some imperfect reconstruction which are shown by circles in Fig. 4.2(k) below. Then we use region merging technique to sort and reconstruct perfectly and got good results (see Fig. 4.2(l)(m)).

Fig. 4.2: (j) – sort and reconstructed fingerprint image with all boundary boxes, (k)-sort and reconstructed fingerprint image without boundary boxes and imperfect reconstructions are shown by the circles, (l)- merged fingerprint image using Region Merging on fingerprint image. (m)- enhance part after implementing region merging to solve imperfect reconstruction problem.

Basically we evaluate our experimental results by the rate of detecting minutiae specially the ridge-ending and bifurcation and matching in the fingerprint image. We access and analyze here three fingerprints impressions from a same finger after enhanced them only and the fingerprint which has obtained using the proposed approach. Experimental results are recorded in terms of two classes, namely True and False. The True class is subdivided into True Acceptance (TA), i.e. correctly classified minutiae, Minutiae Detection rate (MDR), i.e. how many ridge ending and bifurcation detected and True Rejection (TR), and i.e. missed minutiae. The False class is subdivided into False
Acceptance (FA), i.e. wrong minutiae and False Rejection (FR), i.e. correctly rejected pixels. Only TA, TR and FA are tabulated in the results (see Table 1).

Table 1: Averaged result with and without using proposed approach

<table>
<thead>
<tr>
<th>Fingerprint Image</th>
<th>TA (%)</th>
<th>MDR (%)</th>
<th>TR (%)</th>
<th>FA No.</th>
<th>FA (%)</th>
<th>Matched TA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Ridge-ending</td>
<td>90.1</td>
<td>93.6</td>
<td>9.9</td>
<td>91</td>
<td>0.39</td>
<td>94.2</td>
</tr>
<tr>
<td>Bifurcation</td>
<td>63.2</td>
<td>72.1</td>
<td>36.8</td>
<td>54</td>
<td>0.24</td>
<td>91.3</td>
</tr>
<tr>
<td>2 Ridge-ending</td>
<td>84.0</td>
<td>81.3</td>
<td>16.0</td>
<td>90</td>
<td>0.37</td>
<td>90.9</td>
</tr>
<tr>
<td>Bifurcation</td>
<td>59.7</td>
<td>62.9</td>
<td>40.3</td>
<td>66</td>
<td>0.27</td>
<td>90.7</td>
</tr>
<tr>
<td>3 Ridge-ending</td>
<td>88.9</td>
<td>89.1</td>
<td>11.1</td>
<td>91</td>
<td>0.39</td>
<td>93.9</td>
</tr>
<tr>
<td>Bifurcation</td>
<td>62.3</td>
<td>70.5</td>
<td>37.7</td>
<td>57</td>
<td>0.25</td>
<td>90.3</td>
</tr>
<tr>
<td>4 Ridge-ending</td>
<td>92.8</td>
<td>97.4</td>
<td>7.2</td>
<td>90</td>
<td>0.38</td>
<td>97.2</td>
</tr>
<tr>
<td>(obtained by proposed approach)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The averaged matched true acceptance for all the fingerprint impressions is above 90%. But we obtained quite good TA and MDR from the fingerprint image using our proposed approach. Because all fingerprints are taken from one finger but different impressions and we have combine three impressions of the same fingerprint image to get all the ridge ending and bifurcation which might be impossible in one impression. And using our proposed approach we obtained better result which is however still considered quite high.

5.0 CONCLUSION

In this paper we have presented a novel approach for the precise fingerprint enrolment and to improve the performance of the minutiae extraction algorithm. Our proposed approach is mainly Projection Incorporated Subspace based on PCA to combine some different impressions of a fingerprint image which has taken from a same finger. Actually we have taken those fingerprint impressions which has enhanced already and strictly from a same finger. We have presented this approach because most of the time, for the poor quality fingerprint images when we enhance and enroll, we fail to get all minutiae information. Therefore to get accurate minutiae information from an individual finger we have presented the proposed approach.

In this experiment we obtained a better result and got a better TA, MDR and Matched TA from the fingerprint image using our proposed approach which has described in Table 1.

REFERENCES


**BIOGRAPHY**

Md. Rajibul Islam received his BCA (Bachelor in Computer Application) degree from the Indira Gandhi National Open University, New Delhi, India, in 2004. He is currently pursuing his M.Sc degree in Information Technology at Multimedia University, Melaka, Malaysia. His research interests include Pattern Recognition, Image Processing, Fingerprint Matching, Artificial Intelligence and Network security system.

Md. Shohel Sayeed obtained his M.Sc degree in Information Technology from National University of Malaysia in 2000. Currently, he is a lecturer at the Faculty of Information Science and Technology (FIST), Multimedia University, Melaka, Malaysia. His research interests include Image Processing and Signal Processing, Biometrics and Web Technology.

Andrews Samraj received his M.Sc in Computer Science from Bharathidasan University M.Phil degree in Computer science from Bharathiyar University, India in 2002. His research is focused in the fields of Bio Cybernetic Systems using EEG, Signal Processing, Statistical Data Analysis, Biometrics and Bioinformatics. He is presently working in Multimedia University as a Lecturer in the Faculty of Information Science and Technology.
PROBABILISTIC MODEL FOR RESPONSE GENERATION WITH APPLICATION OF BAYESIAN NETWORKS

Aida Mustapha, Md. Nasir Sulaiman, Ramlan Mahmod, Hasan Selamat
Faculty of Computer Science and Information Technology,
University Putra Malaysia, 43400 Serdang, Malaysia. Email: {aida, nasir, ramlan, hasan}@fsktm.upm.edu.my

ABSTRACT

This paper presents Bayesian Networks for probabilistic response generation in a mixed-initiative, transaction dialogue system. The probabilistic approach is to compute the probability of each server response based on the training corpus, and returns to user the response with the highest probability of answering the semantic input from user utterance. Apart from the dialogue acts as provided by the DAMSL-annotated corpus, we proposed two additional types of random variables for the Bayesian Networks, which are the Surface Linguistic Features and the Dialogue Context Features. The BN representation of the dialogue corpus is hoped to allow the system to reason about the dialogue utterances available so far and predict the next system response that will satisfy the user utterance both linguistically and contextually.

Keywords: Dialogue system, Response generation, Natural language generation, Bayesian Networks

1.0 INTRODUCTION

As opposed to research on text generation systems that generate paragraph-length sentences based on rhetorical ground, research on natural language generation in dialogue systems avoid detailed linguistic realization for two main reasons. Firstly, dialogue utterances are typically short, single-sentenced and each carry individual dialogue act (intention). Because dialogues are speech-act-driven, the major concern in a dialogue system is, therefore, the coherency of the utterances that should reflect the overall context in the dialogue conversations. This means, the preceding utterance often gives stronger impacts on the appearance of the next utterance to be generated rather than its syntactic attributes. Secondly, the dialogue utterances are often incomplete and may even be grammatically incorrect but yet are acceptable, as long as it is coherent with the previous utterance. The merit to measure a coherent dialogue is by the relevancy of the response utterances to its preceding utterance. A sequence of utterances is relevant with respect to the dialogue context by means of coherence relation, which corresponds to exchange patterns that indicate the well-formed sequence of initiatives and responses [1].

In 1986, Sperber and Wilson presented the Relevance Theory that connects relevance to the cognitive processing load. When people understand an utterance, they often try to maximize relevance by choosing the context in which the utterance is the most relevant [2]. Our approach is in line with this theory, whereby to rank the best response to user utterance is by finding the response that is highly relevant in context of the input utterance. Relevant in context means the response is timely and semantically correct for the given input. In other words, server response must satisfy one of the intentions that have been raised by the previous utterance. In computational linguistics, the approach to relevance is based on plans and goals paradigm. Carberry proposed a plan recognition strategy to determine if a particular utterance is relevant [3]. According to her, an utterance is considered relevant when it provides a contribution to a plan that is used to achieve the current goal. The relevance issues are often difficult to access because it is related to the topic of the conversation. However, we believe that given the domain-specific corpus, topic can be recognized by using named entity recognition methods.

To date, most natural language generation systems mostly use variations of n-gram models. N-gram is not sufficient for learning dialogues because the final utterance should not only make sense grammatically (which can be learnt using n-gram) but it should answers the previous utterance satisfyingly. We argue that the problem to faithfully adapt the Overgeneration-and-ranking [4] approach in dialogue system goes back to the fundamental characteristics of the target output. An ideal response utterance may not necessarily be complete and grammatically correct, but must be coherent to the input utterance and relevant to the context. Along the line with the previous research in statistical natural language generation both in text-based generation [5, 6] and dialogue-based generation [7, 8], we would like to propose a probabilistic approach to response generation in dialogue systems with Bayesian Networks.
2.0 PROBABILISTIC INTENTION-BASED MODEL

The study of probabilistic and decision-theoretic inference in the area of dialogue systems has been explored in a preliminary fashion by [9] for conversational games, [10] for dialogue act recognition, and [11] for study in anaphora resolution. Probabilistic approach is promising because the presence of uncertainties in dialogue utterances, and also features representation that is very subtle and context-dependent. The response generator can handle uncertainty by using methods of probability and decision theory, but first they must learn the probabilistic theories of the dialogue patterns from experience. The central problem of an automatic response generator has therefore shift from how to respond (generating surface structure of a response) to what to respond (generating the most accurate response) at a particular, specific turn. In this type of problem, there are many hidden variables that are not observable from the surface of the dialogue, but can be tapped from the Dialogue Act Markup in Several Layers DAMSL [12] annotation scheme and available for learning, mainly from the forward-looking functions (FLF) and the backward-looking functions (BLF). Dialogue acts can also be used to identify context of utterance in a particular turn.

In a mixed-initiative, transaction dialogue, an automatic response generator must be able to understand what kind of response is appropriate for a particular user utterance. The main challenge for the generator now lies beyond returning the most suitable response to the user, which is to return the most accurate response that satisfies the intentions of the user at a correct time of conversation. To learn this from the dialogue corpus, the generator has to first determine the intentions, the context of the dialogue at that point (i.e. what is its expected role?; what is the state of conversation now?) and then formulate the relationships between a particular server utterance with the previous utterance based on the corpus utterance data.

To formulate the relationships is really a form of uncertainty. The server must decide whether the current response should answer the user's utterance immediately or the system should interrupt, where it holds the answer in order to ask a question instead. To add to the intricacy of the decisions, the system must also recognize which utterance that it is currently addressing to, in case of multiple utterances at each turn. Given the complexities of the dialogue structures in a mixed-initiative transaction dialogues, a probabilistic Bayesian learning model is proposed to model the knowledge acquired from the dialogue utterance analysis. The fundamental insight of our Bayesian model is to build multiple relationships for the user utterance input, in parallel to the server responses, compute the probability of each relationship, and choose the relationship between a user utterance and the corresponding server response with maximum probability.

3.0 DIALOGUE CORPUS

SCHISMA is a Theater Information and Ticket Reservation system with the main objective to enable users to reserve a particular show from a wide range of available options. The dialogue corpus is a collection of 64 text-based dialogues, obtained through a series of Wizard of Oz experiments. It contains 920 user utterances and 1127 server utterances in total. SCHISMA is a type of mixed-initiative, transaction dialogue.

The mixed-initiative model consists of two types of interactions, which are inquiry and transaction [13]. During inquiry, the system is user-initiated whereby users inquire about details of the shows like the dates, artists, reviews or authors, while the system answer to all the questions. When the conversation arrives at the point where the users indicate that they would like to make reservation, the system will shift into transaction mode where the system takes over the initiative. Starting from this point onwards, the system will ask users series of questions like the number of tickets to reserve, the discount cards and others. User will answer the questions to complete the reservation details.

In transaction dialogue, before reaching the point of reservation, both user and system must collaborate to achieve an agreement to several issues like the ticket price, the seating arrangement or the discount availability. This model is far more complex than the usual question-answering systems because at any point, both parties may request information from each other and the user particularly, may retract any previous decisions and proceed conversation is a total opposite direction. Fig. 1 illustrates the complexities in mixed-initiative, transaction dialogues.
S: Valid reductions are CJP, JTK, Normal and Senior Pass.
S: Do you have a reduction card?
U: I have CJP
S: The price of show "Eugen Onegin" with CJP is 75.00.
U: What expensive.
S: Sorry, I do not understand you, can you reformulate that?
U: Forget about that reservation.

Fig. 1: SCHISMA dialogue extract

3.1 Dialogue acts

SCHISMA corpus is tagged using dialogue act annotation scheme based on DAMSL framework by [14]. SCHISMA-DAMSL consists of five layers, each of which covers different aspect of communicative functions. This research concerns on two levels only, the forward-looking and backward-looking functions. Both levels indicate the communicative functions of an utterance. FLF tags indicate the type of speech act that the utterance is conveying, for example, assert, info-request and commit. BLF tags indicate how the particular utterance relates to the previous utterance and include answers (positive, negative or no-feedback) to questions, degree of understanding or disagreement. Table 1 list out the FLFs and BLFs in SCHISMA corpus.

<table>
<thead>
<tr>
<th>FLF</th>
<th>BLF</th>
</tr>
</thead>
<tbody>
<tr>
<td>conventional</td>
<td>signal understanding</td>
</tr>
<tr>
<td>commit</td>
<td>signal non understanding</td>
</tr>
<tr>
<td>offer</td>
<td>positive answer</td>
</tr>
<tr>
<td>action directive</td>
<td>negative answer</td>
</tr>
<tr>
<td>open option</td>
<td>no answer feedback</td>
</tr>
<tr>
<td>query if</td>
<td>correction feedback</td>
</tr>
<tr>
<td>query ref</td>
<td>accept</td>
</tr>
<tr>
<td>assert</td>
<td>Reject, reject part</td>
</tr>
<tr>
<td>exclamation</td>
<td>hold</td>
</tr>
<tr>
<td>explicit_performative</td>
<td>maybe</td>
</tr>
<tr>
<td>other ff</td>
<td>no blf</td>
</tr>
</tbody>
</table>

Table 1: FLF and BLF for SCHISMA

Dialogue act or intention recognition is not being carried out by the proposed probabilistic response generator due to the availability of the dialogue act information from the corpus. However, research in dialogue act recognitions (DAR) are also based on surface linguistic features of the dialogue utterance. We argue that in the absence of DA-annotated corpus, our response generator can be extended to include the DAR component, which can be run in tandem with analysis of our utterance type (UT) as described in the following section. Even though dialogue acts are important in classifying the specific intention of utterance, it is insufficient to perform interpretation of the utterance because the other component of equal importance is the content of the utterance.

Apart from the set of dialogue acts (intentions) provided by the DAMSL-annotated corpus like SCHISMA, we proposed two sets of features for automatic relationship detection between the user and system utterances. The first feature set is the surface level features like cue words, word order and lexical entries. The second feature set is the context features that are designed specifically for mixed-initiative, transaction dialogue like SCHISMA.

3.2 Surface linguistic features

Clustering and tagging utterances on the basis of surface features can be done using machine learning techniques. The result will be classes of utterances described by a combination of surface properties and can be used to identify change of initiatives and determine the correct surface structure of the expected response. We use two kinds of surface features, which are utterance type and content type based on surface information from the utterance.

3.2.1 Utterance Types (UT)

Utterance Type (UT) denotes the general structure of the server utterances, whether it is a declarative or imperative sentences or wh-question or yes/no questions. However, SCHISMA corpus, being a human/machine dialogue, also
contains incomplete sentences like short answers, commands or remarks, consisting of single verb phrase (VP), noun phrase (NP) or prepositional phrases (PP). A separate class for text is used to categorize long utterance that contains multiple utterances, usually of server explaining the synopsis or the review of a performance.

Table 2 shows the UT classification for each user and server utterances, and is purely determined by the surface linguistic features. The feature classes are modified from the utterance types developed for classification of utterances in the SCHISMA corpus, developed by [15]. The utterance type is very important cue to signal the server utterance of the structure of its expected response. For example, in choosing the best response for a question, the server may exclude considering the probabilities of *wh*- and *yn*-questions. The convention is determined by the grammatical structure of the sentence and the lexical type of the constituents.

<table>
<thead>
<tr>
<th>Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>dec</td>
<td>no <em>wh</em>-word, finite verb on 2nd position</td>
</tr>
<tr>
<td>whq</td>
<td><em>wh</em>-word 1st position, finite verb on 2nd position</td>
</tr>
<tr>
<td>ymq</td>
<td>finite verb on 1st position, subject on 2nd position</td>
</tr>
<tr>
<td>imp</td>
<td>finite verb on 1st position, no subject</td>
</tr>
<tr>
<td>short</td>
<td>short answer i.e. prepositional phrases (PP), noun phrases (NP), proper names, adjectives, adverbs and numbers.</td>
</tr>
<tr>
<td>meta</td>
<td>Greetings i.e. thanks and bye, confirmation i.e. yes and negation i.e. no</td>
</tr>
<tr>
<td>text</td>
<td>long text i.e. reviews, list of genres, titles, authors, artists</td>
</tr>
</tbody>
</table>

Table 2: Utterance types

### 3.2.2 Content Types (CT)

Content Type (CT) indicates the content intention of each utterance in the dialogue. Content intentions can also be seen as semantic tags that are assigned to every utterance based on the cue words detection to indicate the intended information to be conveyed from the utterance. To understand the content of a query utterance, we analyze the choice of *wh*-word i.e. *what* for title, *which genre* for genre or *what time* as time. Combination of the utterance dialogue act and the content will provide the system with semantic of the user input, and lead to the expected answer in the correct response structure. For example, the *when* question indicates that the expected response from server should be a date. We proposed 5 classes of CT, which are application-dependant to SCHISMA Ticket Reservation System as shown in Table 3. Content types is important to determine the semantic input for a given user utterance, which will provide the word class of what the user wants within additional parameters that are set by the user.

<table>
<thead>
<tr>
<th>Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>performance</td>
<td>title, genre, artist, time, date, review</td>
</tr>
<tr>
<td>reservation</td>
<td>reserve, ticket, cost, avail, reduc</td>
</tr>
<tr>
<td>theater</td>
<td>seat, theater</td>
</tr>
<tr>
<td>person</td>
<td>author, director, composer</td>
</tr>
<tr>
<td>other</td>
<td>utterance contains none of the above cue words</td>
</tr>
</tbody>
</table>

Table 3: Content types

### 3.3 Dialogue Context Features

Dialogue analysis for context features are made possible by the presence of two most important dialogue acts according to DAMSL layers; the forward-looking and backward-looking functions. Unlike surface linguistic features that are determined by the surface structure of utterances, context features classify each utterance according to its positioning and time of presence during the conversation. The determination is based on theory of Dialogue Games and Negotiation Games with the advantage of DA-annotated corpus that provides the dialogue acts.

In a dialogue game, (exchanging utterance), each user and server plays a specific role and expects the opponent would do the same too [16]. In case of mixed-initiative corpus like SCHIMA, both parties must respect the party that is currently holding the initiative, and must know when it is time for initiative to change sides. We argue that if the system can learn its expected role from a given user utterance, it can maximize its contextual knowledge to find the highest probability response that satisfy the input utterance.
The idea for negotiation game [17] is an important base for SCHISMA analysis because in ticket reservation system, both participants must collaborate together to achieve an agreement before a reservation for a particular ticket show is reserved. The different phases during the negotiation process can be captured based on the dialogue acts [1]. His work covered a rough characterization of the dialogue acts that occur in each phase and provided a strong basis to understand the contribution of different types of dialogue acts during the negotiation.

Within our scope of research, we argue that the relationships between input and response utterance can be better postulated if we have the information of which phase the system is currently in, so it can limit its response choices. Clearly, knowledge about the initiatives alone is insufficient because sequence of speaker turns does not necessarily consecutive. This implies that initiatives may change at any time and does not always alternate regularly. We propose two dialogue context features that are the initiative-type and the negotiation-type. The characteristics of initiative and negotiation phase are summarized in Table 4.

<table>
<thead>
<tr>
<th>Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>initiative-type</td>
<td>inquiry, transaction, neutral</td>
</tr>
<tr>
<td>negotiation-type</td>
<td>open, inform, propose, confirm, close</td>
</tr>
</tbody>
</table>

Table 4: Dialogue context features

3.3.1 Initiative Types (IT)

Any dialogue utterance in general can either be an initiative, a response or neutral. An initiative requires a particular response from the opposing party, where the response can either be a direct answer or a signal of acceptance. In the case of SCHISMA, initiative is an inquiry for information and the response consists of positive or negative reply; or else change of initiative into transaction. Before going into transaction, both parties must agree on a specific issue and commit to proceed to the next level. During transaction, the exchange pattern is still the same as in inquiry state where queries are responded with assertives and assertions are followed by acknowledgements. Only the lead is passed to system as the new control party of the conversation. We believe that it is therefore possible to formulate relationship between the dialogue acts and the exchange patterns within the context of the conversation by identifying the party who is currently holding the right to initiate.

3.3.2 Negotiation Types (NT)

To relate a particular server utterance with the input user utterance, the analysis of dialogue context should be extended to the identification of type of activity that takes place at that particular utterance. A ticket reservation system like SCHISMA is a transaction dialogue because the user and the system must collaborate together to reach an agreement point of the details of particular reservation, before the reservation is confirmed and bound. The main activity during a transaction is negotiation, as described by [18] and [17] in the formal model of negotiation within the scope of dialogue systems.

A negotiation can be characterized as an exchange of information that will eventually lead to agreement to that satisfy both parties’ information needs. The final agreement depends on the choices made during the negotiation process. In our proposed dialogue context features, we use the negotiation types as the indicator for negotiation phases during the conversation. The negotiation types are illustrated in Fig. 2.

![Fig. 2: Phases in negotiation](image)

4.0 Experiments and Results

The general problem for a probabilistic response generator is to choose the highest probability response that answers the user input accurately. Our task is therefore to compute the probability of each server utterance and return to user the
response that has the highest probability. It is important that the generator also learn other features of the utterance: is it an initiative or response? How does it relate its content to the context? What kind of dialogue acts it has (is it assertive, interrogative, directives)? Is the content positive or negative?

In relation to the SCHISMA corpus, the analysis of highest probability response will be in terms of goals and planning of ticket reservation. The goal is to reserve ticket for user. The plan is to confirm the title, date and the total ticket to reserve, with agreement of the user. The different state of information-seeking process (first the title, then the tickets followed by the cost), leads to the importance of negotiation phase during the conversation.

Constructing a probabilistic model for automatic response generation requires all dialogue acts, surface linguistic and context features as random variables in setting up the network. Given information about random variables in the evidence set \( X_E \), what is the configuration of the server utterance \( Y \), that maximize its probability

\[
y^* = \arg \max_{Y} P(Y \mid X_E)
\]

However, the SCHISMA corpus contains a total of 1127 server utterances, hence, too many attributes to be used as the observed node. In order to reduce the number of attributes within the server utterance classes, we have manually prepared the server utterances into templates and classify them as shown in Table 5.

<table>
<thead>
<tr>
<th>Classes</th>
<th>Utterance Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>confirmRT</td>
<td>Utterances in the form “You have reserved/made reservations/tickets…”</td>
</tr>
<tr>
<td>requestRV</td>
<td>Utterances in the form of “Shall I reserve…”</td>
</tr>
<tr>
<td>shortUT</td>
<td>Short utterances like thanks, greetings, answers &amp; fragments</td>
</tr>
<tr>
<td>whQS</td>
<td>Wh-questions with 2\textsuperscript{nd} person pronoun as subject</td>
</tr>
<tr>
<td>ynQS</td>
<td>Yes/no questions 2\textsuperscript{nd} person pronoun as subject. Utterances in the form “Can/Do You…”</td>
</tr>
<tr>
<td>informMV</td>
<td>Inform utterances containing mental verb like the verb “to have/look/find/request…”</td>
</tr>
<tr>
<td>informALT</td>
<td>Inform utterances containing alternative performances. 2\textsuperscript{nd} person pronoun as subject</td>
</tr>
<tr>
<td>informLST</td>
<td>Inform utterances about list of genre, type of reductions, type of rooms and address of the theater.</td>
</tr>
<tr>
<td>informAT</td>
<td>Inform utterances about when and where artists play/perform, review of performance</td>
</tr>
<tr>
<td>informCT</td>
<td>Inform utterance about cost and availability of tickets</td>
</tr>
<tr>
<td>informPT</td>
<td>Inform utterances about performances</td>
</tr>
</tbody>
</table>

Table 5: Server utterance classes

5.0 Conclusions

To build automatic, robust, practical, and easy-to-use dialogue systems, it is essential to study actual dialogues from within the target application's domain in order. A corpus-based approach to dialogue generation will sidestep the need to prepare pattern-response rules that are both time consuming and labor intensive. This type of generation also promise a portability across other limited, task-oriented domain such as the theater reservation system, train ticket booking system or help desk system; provided the conversation transcripts from the real dialogues.

In our point of view, for limited task-oriented dialogue systems, it is sufficient for the generation engine to learn to acquire its response automatically from the corpus based on the dialogue acts, surface linguistic features and the dialogue context features. The basis of learning is that the output response must be coherent to the previous utterance and relevant to the context in the dialogue. Since, the study of surface linguistic features is established; we hope to venture deeper into pragmatics to find the context features that can adequately characterize dialogue utterances, hence, automating the learning process for generating response utterances in dialogue systems.

REFERENCES


A SET OF SCALAR FEATURES REPRESENTATION FOR 3D FACE RECOGNITION

Fatimah Khalid¹, Tengku Mohd. Tengku Sembok², Khairuddin Omar³

¹Department of Multimedia, Faculty of Computer Science and Information Technology
University Putra Malaysia, 43400 UPM Serdang,Malaysia. E-mail: patiaziz@yahoo.com
²Department of Information Science, Faculty of Technology and Information Science
³Department of Science and System Management, Faculty of Technology and Information Science
University Kebangsaan Malaysia, 43600 Bangi,Selangor,Malaysia.
E-mail: tmts@ftsm.ukm.my; ko@ftsm.ukm.my

ABSTRACT

The purpose of the feature selection is to remove the irrelevant or redundant features which may degrade the performance of face classification. Before doing the feature selection, we investigate automatic methods for detecting face anchor points with 199 3D-facial points of 29 individuals. There are 7 images per subject including views presenting light rotations and facial expressions. Each images have twelve anchor points which are Right Outer Eye, Right Inner Eye, Left Outer Eye, Left Inner Eye, Upper nose point, Nose Tip, Right Nose Base, Left Nose Base, Right Outer Face, Left Outer Face, Chin, and Upper Face. All the control points are based on the measurement on an absolute scale (mm). After all the control points have been determined, we will extract a relevant set of features. These features are classified in 3 : (1) distance of mass points, (2) angle measurements, and (3) angle measurements. There are fifty-three local geometrical features extracted from 3D points human faces to model the face for face recognition and the discriminating power calculation is to show the valuable feature among all the features. Experiment performed on the GavabDB dataset (412 faces) show that our algorithm achieved 83%, 86%, 90%, 93% of success when respectively the first one, two, three and seven matches were selected.

Keywords : local geometrical features, discriminating power

1.0 INTRODUCTION

The second step in face recognition after face detection is comparing between faces based solely on a set of scalar features calculated the distances and angles. A vector in feature space is used associated to each face in a set of features. The feature calculation begins with the 3 dimensional data which in text file and simple mathematical formula is used. This method is made on the basis of a small number of feature measurements. As a result, less memory is required for storage and the computationally is very simple. Although the feature calculation is obviously different from other researcher where they based on the curvature, we also used the curvature information to detect points. These features are calculated primarily from the set of face specific features and can be illustrated in figure 1.

Section 2 describes the related work of previous researchers. Section 3 describes the proposed feature representation for verification. Next, section 4 explained how useful each feature in face recognition. Then, the experiment is discussed in section 5 to show how well the set of scalar representation work. Lastly, section 6 concluded this paper.

2.0 RELATED WORK

Various algorithms have been proposed for feature representations. Gaile [1] used 12 feature vectors to make a comparison within 24 faces. The features are head width, nose height, nose depth, nose width, eye separation, maximum Gaussian curvature on the ridge line, average minimum curvature on the ridge above the tip of the nose, Gaussian curvature at the bridge, and Gaussian curvature at the base. The performance with a small feature set and very basic statistical methods for classification shows the recognition results were better than 70% in all cases. But they did not mention how accurate the recognition result whether in 1 image, 2 images or how many images they can be matched. Meanwhile, Ana et al.[2] employed a set of eighty six features using a database of 420 3D range images, 7 images per each one of a set of 60 individuals. After the feature discriminating power analysis, the first 35 features of the ordered list of features according to the Fisher coefficients were used to represent faces in their face recognition experiments. The features offering better recognition results were angles and distances measurements. Hallinan et al. [3] obtained a set of twelve 3D feature extracted from segmented regions using curvature properties of the surface. They do an experiment for face recognition using a database of 8 individual and 3 images per individual. As a result, they obtained 95.5% of recognition rate providing a previous 100% of correct feature extraction.

3.0 CALCULATION OF EXTRACTED FEATURE
After the anchor point determination task, a feature extraction stage is performed. Fifty-three non-independent features extracted from the anchor points were computed. Extracted features are categorized into 3 categories: distance measurements, angle measurements, and scalar space. Scalar space is a distance measurements in specific distance where details in the next section. The distance measurement is the length of these nine anchor points (as in Figure 1) while the angle measurement is the angle of three points. This calculation will be explained in the next topic.

3.1 Distance measurements

Distance is a numerical description of how far apart objects are at any given moment in time. In physics or everyday discussion, distance may refer to a physical length, a period of time, or an estimation based on other criteria (e.g. "two counties over"). In mathematics, distance must meet more rigorous criteria[4]. In neutral geometry, the minimum distance between two points is the length of the line segment between them. In algebraic geometry, one can find the distance between two points of the xy-plane using the distance formula. The distance between \((x_1, y_1)\) and \((x_2, y_2)\) is given by

\[
d = \sqrt{(\Delta x)^2 + (\Delta y)^2} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.
\]

Similarly, given points \((x_1, y_1, z_1)\) and \((x_2, y_2, z_2)\) in three-space, the distance between them is

\[
d = \sqrt{(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}.
\]

Which is easily proven by constructing a right triangle with a leg on the hypotenuse of another (with the other leg orthogonal to the plane that contains the 1st triangle) and applying the Pythagorean theorem. In the study of complicated geometries, we call this (most common) type of distance Euclidean distance, as it is derived from the Pythagorean theorem, which does not hold in Non-Euclidean geometries. This distance formula can also be expanded into the arc-length formula.

The distance calculation can be seen in Table 1 and referred to the nine anchor points as in Figure 1. There are thirty-five features produced from this distance measurement category.

<table>
<thead>
<tr>
<th>Table 1 : Distance, Relation and Mean of points</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Distance between mass points</strong></td>
</tr>
<tr>
<td><strong>Id. num</strong></td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
</tbody>
</table>

©FCSIT, UM 2007
3.2 Calculate the angles

The calculation of the angles is based on any three points [5]. These three points performed a triangle which has internal angles ‘A’, ‘B’ and ‘C’, and sides of length ‘a’, ‘b’ and ‘c’. This triangle can be illustrated in figure 2.

The distance formula as mentioned previously can be used to find the distance between two points (the length of a, b, and c). Once the three side measurements are known, then the internal angles ‘A’, ‘B’ and ‘C’ can be found as well.

When no angles are known, the cosine rule is the only option. The first step is by using the cosine rule to find the largest angle. This is because, there can only be one angle in a triangle that is obtuse (greater than 90°). If a triangle has an obtuse angle, then this will be it. The reason for finding it first is that in the next step we will use the sine rule to find the second angle. The inverse sin operation that we will use can only give us acute angles (less than 90°), so we avoid a possible wrong answer by first eliminating the only possibility of obtuse angle. The largest angle is always opposite to the largest side. The cosine rule is defined as:

\[ b^2 = a^2 + c^2 - 2ac \cos B \]
\[ \cos B = \frac{a^2 + c^2 - b^2}{2ac} \]

Then, find the inverse cos B to get the angle. The remaining angles can be found by using the sine rule as defined below:
\[
\frac{c}{\sin C} = \frac{b}{\sin B} \\
\sin C = \frac{c \sin B}{b}
\]

Lastly find the inverse sin C to get the angle. Table 2 shows the angle calculation based on the anchor points in figure 1 and nine features from this angle category.

### Table 2: Angle and Mean angle of points

<table>
<thead>
<tr>
<th>Angle between three points</th>
<th>Id. num</th>
<th>Feature description</th>
<th>Id. num</th>
<th>Feature description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ang(P1,P3,P2)</td>
<td>5</td>
<td>ang(P5,P3,P6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 ang(P1,P3,P6)</td>
<td>6</td>
<td>ang(P1,P2,P6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 ang(P4,P3,P6)</td>
<td>7</td>
<td>ang(P4,P6,P5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 ang(P2,P3,P6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mean of angles

| 8 ½[ang(P1,P3,P6)/ang(P2,P3,P6)] | 9 ½[ang(P4,P3,P6)/ang(P5,P3,P6)] |

### 3.3 Calculate the Scalar Features

The measurements of the face can be described from the most basic set of scalar features. Its calculation is based from the twelve anchor points detected. The scalar features are Head height, Head width, Left eye width, Right eye width, Eye separation, Span of eyes, Nose depth, Nose height, and Nose width.

The measurements will be on an absolute scale (mm) which obtained from the 3D scanner and based on the 3D points (x,y,z). The head height is made based on the point features marking upper head point and chin point while head width based on the left head tip and right head tip. It is the distances between four 3D coordinates.

The eyes measurements are from the point features marking inside and outside corner cavities of the eyes. The width of each eye is the difference between outer and inner eye points. Similarly, the separation of the eyes is the distance between two inner points and the total width of the eyes is the difference between two outside points.

The nose height is the distance between upper nose point and the nose tip. Nose width is defined by the distance measured between left and right nose base point. The nose depth is measured based on the nose tip as the maximal depth in the nose region. Then we measure the distance of the base nose points and take the average of them.

All these scalar features can be illustrated in table 3.

### Table 3: Distance of specific points (scalar space)

<table>
<thead>
<tr>
<th>Id. num</th>
<th>Scalar Space</th>
<th>Feature description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Face height</td>
<td>dist(P9,P10)</td>
</tr>
<tr>
<td>2</td>
<td>Face width</td>
<td>dist(P11,P12)</td>
</tr>
<tr>
<td>3</td>
<td>Left eye width</td>
<td>dist(P2,P5)</td>
</tr>
<tr>
<td>4</td>
<td>Right eye width</td>
<td>dist(P1,P4)</td>
</tr>
<tr>
<td>5</td>
<td>Eye separation</td>
<td>dist(P1,P2)</td>
</tr>
<tr>
<td>6</td>
<td>Span of eyes</td>
<td>dist(P4,P5)</td>
</tr>
<tr>
<td>7</td>
<td>Nose depth</td>
<td>½[dist(P6,P7)+ dist(P6,P8)]</td>
</tr>
<tr>
<td>8</td>
<td>Nose height</td>
<td>dist(P3,P6)</td>
</tr>
<tr>
<td>9</td>
<td>Nose width</td>
<td>dist(P7,P8)</td>
</tr>
</tbody>
</table>

### 4.0 DISCRIMINATING POWER
For features to be useful in automatic face recognition they must ideally satisfy two criteria. First they must be automatically be robustly detectable; their measurement must be consistent for the same face over reasonable variation in view position, expression, age, weight, etc [1]. Second their values must vary distinctly over the range of different individuals [1].

Table 1 describes and classifies these features into three categories which areas, distances, and angles. All the 53 features have their own associated identification number and the ordered discriminating power of the feature where the high discriminating power have a small rank position while the low discriminating power have a high rank position. Discriminating power estimation of each feature $\Phi$ has been computed using Fisher coefficient, which represents the ratio of between-class variance to within-class variance, according to the formula [6]

$$\Phi = \frac{\sum_{i=1}^{c} n_i (m_i - m)^2}{\sum_{i=1}^{c} \sum_{x \in \Phi_i} (x - m)^2}$$

where $c$ is the number of classes or subjects, $\Phi$ is the set of feature values for class $i$, $n_i$ is the size of $\Phi_i$, $m_i$ is the mean of $\Phi_i$, and $m$ is the global mean of the feature over all classes [2].

There are 29 classes corresponding to the number of distinct individuals in the database. Same as Moreno et al.[5], although there are seven images per person, when computing the Fisher coefficients, only the 3D facial images having the whole set of features correctly extracted have been employed (a total of 199 images). Table 4 includes the mathematical definition of each one of the 53 more discriminating features. Each feature has an associated feature description and a ranking position.

### Table 4: Extracted features from the points determination, and their position in an ordered list according to their discriminating power.

<table>
<thead>
<tr>
<th>Rank pos</th>
<th>Feature description</th>
<th>Rank pos</th>
<th>Feature description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Eye separation</td>
<td>28</td>
<td>dist(P2,P6)</td>
</tr>
<tr>
<td>2</td>
<td>$\frac{1}{2}$[dist(P1,P7)/dist(P2,P8)]</td>
<td>29</td>
<td>ang(P5,P3,P6)</td>
</tr>
<tr>
<td>3</td>
<td>$\frac{1}{2}$[dist(P1,P9)/dist(P2,P9)]/dist(P3,P9)</td>
<td>30</td>
<td>dist(P5,P8)</td>
</tr>
<tr>
<td>4</td>
<td>$\frac{1}{2}$[dist(P1,P6)/dist(P2,P6)]</td>
<td>31</td>
<td>Nose height</td>
</tr>
<tr>
<td>5</td>
<td>dist(P1,P6)</td>
<td>32</td>
<td>$\frac{1}{2}$[dist(P1,P3)/dist(P2,P3)]</td>
</tr>
<tr>
<td>6</td>
<td>$\frac{1}{2}$[dist(P4,P9)/dist(P5,P9)]/dist(P3,P9)</td>
<td>33</td>
<td>dist(P1,P2)/dist(P3,P6)</td>
</tr>
<tr>
<td>7</td>
<td>ang(P1,P3,P2)</td>
<td>34</td>
<td>Nose width</td>
</tr>
<tr>
<td>8</td>
<td>$\frac{1}{2}$[dist(P4,P3)/dist(P5,P3)]</td>
<td>35</td>
<td>Left eye width</td>
</tr>
<tr>
<td>9</td>
<td>dist(P2,P8)</td>
<td>36</td>
<td>dist(P4,P5)/dist(P3,P9)</td>
</tr>
<tr>
<td>10</td>
<td>$\frac{1}{2}$[dist(P4,P5)/dist(P5,P6)]</td>
<td>37</td>
<td>Head width</td>
</tr>
<tr>
<td>11</td>
<td>ang(P2,P3,P6)</td>
<td>38</td>
<td>Right eye width</td>
</tr>
<tr>
<td>12</td>
<td>dist(P4,P3)</td>
<td>39</td>
<td>dist(P2,P3)</td>
</tr>
<tr>
<td>13</td>
<td>$\frac{1}{2}$[dist(P4,P7)/dist(P5,P8)]</td>
<td>40</td>
<td>dist(P1,P9)</td>
</tr>
<tr>
<td>14</td>
<td>dist(P5,P3)</td>
<td>41</td>
<td>dist(P2,P9)</td>
</tr>
<tr>
<td>15</td>
<td>ang(P1,P2,P6)</td>
<td>42</td>
<td>$\frac{1}{2}$[ang(P1,P3,P6)/ang(P2,P3,P6)]</td>
</tr>
<tr>
<td>16</td>
<td>$\frac{1}{2}$[dist(P1,P9)/dist(P2,P9)]</td>
<td>43</td>
<td>ang(P4,P6,P5)</td>
</tr>
<tr>
<td>17</td>
<td>ang(P1,P3,P6)</td>
<td>44</td>
<td>dist(P4,P5)/dist(P3,P6)</td>
</tr>
<tr>
<td>18</td>
<td>Head height</td>
<td>45</td>
<td>dist(P8,P9)</td>
</tr>
<tr>
<td>19</td>
<td>dist(P4,P9)</td>
<td>46</td>
<td>ang(P4,P3,P6)</td>
</tr>
<tr>
<td>20</td>
<td>dist(P1,P7)</td>
<td>47</td>
<td>dist(P7,P9)</td>
</tr>
<tr>
<td>21</td>
<td>dist(P5,P9)</td>
<td>48</td>
<td>$\frac{1}{2}$[dist(P4,P9)/dist(P5,P9)]</td>
</tr>
<tr>
<td>22</td>
<td>dist(P9,P3)</td>
<td>49</td>
<td>Span of eyes</td>
</tr>
</tbody>
</table>
5.0 EXPERIMENT AND RESULT

Experiment has been done with 412 3D-facial points (without texture) of 60 individuals. There are 7 images per subject including pose variation and facial expressions. Best recognition results using the 24 more discriminating features based on eigenvalues that are more than one were obtained when frontal views were tested. To obtain this lower dimensional representation, the principal component analysis (PCA) is used. 83%, 86%, 90%, 93% of success when respectively the first one, two, three and seven matches were selected. This matching procedure was based on the minimum Euclidean distance classifier.

6.0 CONCLUSION

This paper presented a 53 features extraction from 12 anchor points. The local geometric features are calculated basically using Euclidean distance and angle measurement. Although there are 53 features, 24 of them have been used for recognition and shows an improvement compared to Ana[5] works with less features.

REFERENCES


BIOGRAPHY

Fatimah Khalid obtained her Master of Information Technology from University Kebangsaan Malaysia in 1998. Currently, she is a lecturer at the Faculty of Computer Science and Information Technology, University of Putra Malaysia. Her research areas include computer vision and image processing. She has published a number of papers related to these areas.

Tengku Mohd. Tengku Sembok (Prof. Dr.) is currently a Vice Canselor at Universiti Pertahanan Malaysia. His research areas include Information Retrieval, Multimedia System, NLP, and Intelligence System.

Khairuddin Omar (Assoc. Prof. Dr.) is a lecturer at Department of System Science and Management, Universiti Kebangsaan Malaysia. His research areas include Information Retrieval, Knowledge Based Systems, and Intelligent Information Mining.
A STUDY ON NORMALIZATION AND FEATURE EXTRACTION TECHNIQUES FOR THE RECOGNITION OF ISOLATED KANNADA HANDWRITTEN NUMERALS

Surekha Patil¹, Meenaksi Patil²

¹Lecturer Department of MCA, BMS Institute Of Technology, Bangalore, India.
   Email: patil.surekha@rediffmail.com
²Senior Lecturer, Department of Electronics and Communication, PDA Engineering College, Gulbarga, India.
   Email: meena_a_patil@yahoo.co.in.

ABSTRACT

The evaluation of various techniques is important to select the correct options in developing character recognition system. Normalization is an important preprocessing technique for character recognition and feature extraction is the core of OCR. In this paper we conduct experiments for different normalization functions and different feature extraction strategies. It is hoped that this study improves the performance of character recognition system and will find wider use to the researchers.

Keywords: Handwritten numeral recognition, normalization, aspect ratio, direction features, gradient features.

1.0 INTRODUCTION

Numeral recognition is the most challenging and tantalizing field. Handwritten numeral recognition is an important step in many document processing application like office or library automation, bank and postal services, tax forms and other handwritten forms. The correct interpretation of digit in numeral recognition is very important. Over past three decades, many different methods have been explored by large number of scientists. The complexity of the problem is increased by noise in the data and infinite variability of handwriting as a result of mood of the writer and nature or style of writing.

Generally a character recognition system includes three main tasks: pre-processing, feature extraction, and classification. Pre-processing, includes the size and aspect ratio of normalized image, the interpolation of techniques of pixel values, etc. In feature extraction various types of features and extraction techniques are available such as geometric features, directional features etc. For classification, many new methods are available including statistical classifiers, artificial neural networks (ANN), support vector mechanism (SVM), hybrid classifiers, etc. In this paper we analyze the effect of size normalization and feature extraction on the recognition of handwritten numerals. Improved normalization and feature extraction strategies will evaluate the performance of the recognition techniques.

Normalization is considered to be one of the important preprocessing factors for character recognition [1]. Normally, the character image is linearly mapped onto the standard plane by the interpolation or extrapolation. The size and position is controlled by the x-y dimension of the normalized plane. The basic normalization strategies are linear, nonlinear and moment normalization. Feature extraction is at the core of character recognition and a variety of feature types and extraction strategies have been proposed [2]. The chain code feature extraction is widely adopted, where as the gradient feature is applicable to gray-scale images and binary images. Selecting the best method improves the recognition rate.

The paper is organized as follows. Section 2 describes character acquisition; Section 3 describes the preprocessing steps like normalization and digitization; Section 4 describes the feature extraction techniques; Section 5 presents the experimental results; section 6 provides concluding remarks.
2. CHARACTER ACQUISITION.

The first step in recognition process is to acquire handwritten numeral characters. We used flatbed scanner for digitization. The images are in gray tone and digitized at 300 dip and stored as Tagged Information File (TIF) format. The Kannada handwritten numerals were collected from different people, with different age group and different occupation for our study. Fig 1 shows few samples of Kannada handwritten numeral (0 – 9).

3. PREPROCESSING.

3.1 Digitization.

In digitization, the image is converted into binary form by binarization method. We have used a histogram based global binarizing algorithm to convert them into two-tone (0 and 1) images (Here ‘1’ represents object point and ‘0’ represents background point). The digitized image may contain spurious noise points and irregularities on the boundary of the numerals, leading to undesired effects on the system. This digitized image is put through preprocessing routines that smooth the image and eliminate noise, artificial holes (by thinning process) and other artifacts produced by the digitization process.

3.2. Normalization.

Normalization is one of the important pre-processing factors for character recognition. Normally, the character image is linearly mapped on to a standard plane by interpolation/extrapolation [3, 4]. The size and position of the character is controlled such that the X/Y dimension of the normalized planar field. The implementation of interpolation/extrapolation is influential to character recognition. By linear mapping, the character shape is not deformed except the aspect ratio changes. Hence the use of Aspect Ratio Adaptive Normalization (ARAN) technique is used. In ARAN technique, the dimensions of the standard plane are not necessarily filled [5]. Depending on the aspect ratio the normalized image is centered in the plane with one dimension filled. Assuming the standard plane is squared and the side length is denoted by L. Denote the width and height of the normalized character image as W2 and H2, the aspect ratio is defined by

\[ R_2 = \begin{cases} \frac{W_2}{H_2}, & \text{if } W_2 < H_2 \\ \frac{H_2}{W_2}, & \text{otherwise} \end{cases} \]  

(1)

If the normalized image fills one dimension, then max (W2, H2) = L.

To implement ARAN method, the normalized character is filled into another plane of flexible size W2 X H2 and then the flexible plane is shifted to overlap the standard plane by aligning the boundaries or centroid. The transformation can be accomplished by forward mapping or backward mapping. Denote the original image and the normalized image as f(x, y) and g(x', y') respectively. The normalized image is generated by \( g(x', y') = f(x, y) \) based on coordinate mapping. The forward mapping and backward mapping are given by

\[ x' = x'(x, y), \quad x = x(x', y'), \]
\[ y' = y'(x, y), \quad y = y(x', y') \] respectively.

The forward mapping and backward mapping of linear normalization and moment normalization are given in the table 1. Here \( \alpha \) and \( \beta \) denote the ratios of transformation given by
\[ \alpha = \frac{W_2}{W_1} \quad \text{and} \quad \beta = \frac{H_2}{H_1} \]

Table 1. Coordinate mapping of normalization methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Forward mapping</th>
<th>Backward mapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>( x' = \alpha x )</td>
<td>( x = x' / \alpha )</td>
</tr>
<tr>
<td></td>
<td>( y' = \beta y )</td>
<td>( y = y' / \beta )</td>
</tr>
<tr>
<td>Moment</td>
<td>( x' = \alpha (x - x_c) + x'_c )</td>
<td>( x = (x' - x'_c) / \beta + x_c )</td>
</tr>
<tr>
<td></td>
<td>( y' = \beta (y - y_c) + y'_c )</td>
<td>( y = (y' - y'_c) / \beta + y_c )</td>
</tr>
</tbody>
</table>

Here the moment normalization refers to linear transformation without rotation, with the center and size of the normalized image determined by moments \((x_c, y_c)\) denotes the centroid of the original image given by

\[ x_c = \frac{m_{10}}{m_{00}} \quad \text{and} \quad y_c = \frac{m_{01}}{m_{00}}, \]

where \( m_{pq} \) denotes the geometric moments:

\[ m_{pq} = \sum_x \sum_y x^p y^q f(x, y) \]

and \((x'_c, y'_c)\) denotes the geometric center of the normalized plane, given by

\[ x'_c = \frac{W_2}{2} \quad \text{and} \quad y'_c = \frac{H_2}{2} \]

In the following, we will give 7 normalization functions with varying aspect ratio mapping. For dimension based normalization, the actual image width and height are taken as the dimensions and for moment based normalization, the original image is viewed to be centered at the centroid and the boundaries are reset to \( [x_c - 2\sqrt{\mu_{20}} x_c + 2\sqrt{\mu_{20}}] \) and \( [y_c - 2\sqrt{\mu_{02}} y_c + 2\sqrt{\mu_{02}}] \). Accordingly, the dimensions are reset to \( W_1 = 4\sqrt{\mu_{20}} \) and \( H_1 = 4\sqrt{\mu_{20}} \). The image plane is expanded or trimmed so as to fit the image in this range. The aspect ratio of the original image is then calculated by

\[ R_1 = \begin{cases} \frac{W_1}{H_1} & \text{if } W_1 < H_1 \\ \frac{H_2}{W_2} & \text{otherwise} \end{cases} \]

The normalization strategies and aspect ratio mapping functions are listed below:

a. Linear normalization with fixed aspect ratio \( R_2 = 1 \).

b. Linear normalization with aspect ratio preserved \( R_2 = R_1 \).

c. Linear normalization with square root of aspect ratio \( R_2 = \sqrt{R_1} \).

d. Linear normalization with cubic root of aspect ratio \( R_2 = 3^{\sqrt{R_1}} \).

e. Linear normalization with piecewise linear aspect ratio

\[ R_2 = \begin{cases} 0.25 + 1.5R_1 & \text{if } R_1 < 0 \\ 1 & \text{otherwise} \end{cases} \]

f. Moment normalization with aspect ratio preserved

g. Moment normalization with square root of aspect ratio.

h. Moment normalization with cube root of aspect ratio.

4. Feature extraction techniques.

Feature extraction plays an important role in character recognition. Feature extraction is at the core of character recognition & a large variety of feature types & extraction strategies have been proposed [6]. The distribution of local stroke direction (direction feature) is most popularly used due to high performance and easy of implementation, these can be measured from skeleton [7], Chain code [8] or gradient[9]. The chain code feature is widely adopted, whereas the gradient feature is applicable to gray-scale images as well as binary images. Feature extraction extracts the well defined characteristics, which classifies the character in classification stage. Here we extract two types of direction features: chain code feature and gradient feature.

For chain code feature and gradient feature extraction, the normalized image is decomposed into four orientation or eight direction planes. The convolution with blurring mask is equivalent to low-pass filtering and sampling. Usually, a Gaussian mask is used:

\[ h(x, y) = \left(1/2\pi\sigma^2\right) \exp\left(-x^2 + y^2\right) / 2 \sigma^2 \]
In chain code feature extraction, the contour pixels of the normalized image are assigned 8-direction codes as in Fig 2 and the contour pixels of each direction are assigned to the corresponding direction plane. In this procedure, a pixel of multiple connectivity can be assigned multiple chain codes. If 4-orientation feature is to be extracted, the direction planes of each pair of opposite directions are merged into one feature plane & blurring is performed on four planes.

For gradient feature extraction, we use the Sobel operator to compute the x/y components of gradient and the gradient image is decomposed into four orientation planes or eight direction planes.

The Sobel operator has been used by many researchers [10], here we have two masks to compute the gradient components in two axes. The masks are shown in Fig 3 and the gradient \( g(x, y) = [g_x, g_y] \) at location \((x, y)\) is computed by:

\[
\begin{align*}
g_x &= f(x+1, y+1) + 2f(x+1, y) + f(x+1, y-1) \\
    &\quad - (f(x-1, y+1) + 2f(x-1, y) + f(x-1, y-1)), \\

\end{align*}
\]

\[
\begin{align*}
g_y &= f(x-1, y+1) + 2f(x,y+) + f(x+1, y+1) \\
    &\quad - (f(x-1, y-1) - 2f(x,y-1) - f(x+1,y-1)).
\end{align*}
\]

The gradient strength and direction can be computed from the vector \([g_x, g_y]^T\). For character feature extraction, the gradient of every pixel on the normalized image is computed. The gradient feature can be extracted from either binary or gray-scale normalized image. From a binary character image, the gray-scale normalized image is generated by forward binary-to-gray pixel mapping as in section 3.2.

5. Results.

Regarding normalization functions, the moment normalization function (f) to (h) perform best. The other linear normalization functions (a)- (e) perform fairly well. With the feature vectors, the chain code feature and gray-scale gradient feature perform best in the two cases. The superiority of gray-scale gradient feature is due to the reliable computation of the gradient. Comparing the 4-Orientation and 8-direction features, it is clear that the performance of 8-direction feature is superior, at the cost of higher dimensionality.
6. Conclusion.

In this paper, we present different normalization and feature extraction methods to increase the handwritten numeral recognition accuracy. These methods were implemented on a dataset representing (Kannada) digits 0–9 collected from 500 people of different age group. Thus according to our study, the proper selection of normalization and feature extraction technique will definitely affect the speed of automatic recognition system.

REFERENCES


BIOGRAPHY

Surekha Patil obtained her Master of Computer Application from Gulbarga University in 1995. Currently, she is a lecturer at BMS Institute of Technology, Bangalore, India. Her research areas include Image Processing and Pattern Recognition. She is also a member of ISTE.

Meenakshi Patil obtained her B.E and M.Tech in Electronics Engineering from Gulbarga University in 1998. Currently, she is a senior lecturer at Department of Electronics and Communication Engineering, PDA Engineering College, Gulbarga, India. Her research areas include Digital Signal Processing and Networking.
TOOL SUPPORT FOR COMPETENCY EVALUATION OF WEB-ONTOLOGIES

Muthukkaruppan Annamalai¹, Noor Hasimah Ibrahim Teo²
Faculty of Information Technology and Quantitative Sciences, MARA University of Technology, 40500 Shah Alam, Selangor, Malaysia.
Email: ¹mk@tmsk.uitm.edu.my, ²shimateo@tmsk.uitm.edu.my

ABSTRACT

Ontology is a key component of the semantic web. The web ontologies facilitate users to espouse a common understanding of the structure of knowledge in an application domain, enabling sharing of knowledge meaningfully on the web. The increasing use of the semantic web has escalated the demand for ‘competent’ ontologies, i.e., ontologies that can effectively support their purpose of use. Hence, we regard competency as an important quality that must be assessed when evaluating ontologies being built or adapted. The competency evaluation is a semantic analysis performed by a user to ascertain if the existing ontological commitments are adequate to support the ontology’s needs and requirements. The detail analysis inevitably needs to inspect the concepts defined in the ontology and trace the relationships that exist between them. Evaluating ontologies in this manner is a challenging task because it is cumbersome to trace the intricate conceptual relationships from within the ontology editors. Although some ontology editors provide visualisation support; it is still difficult to trace the conceptual relationships on the miniaturised representation of rich and large ontologies. In an attempt to circumvent this problem, we have developed a competency evaluation tool for the Protégé ontology editor that complements an ontology developer. The tool is currently equipped with only basic features, yet has demonstrated its potential as a useful tool support for competency evaluation of web ontologies.

Keywords: Semantic web, Knowledge modelling, Ontology evaluation, Software tool

1.0 INTRODUCTION

Ontology or knowledge model is a formal specification of a conceptualisation [1] that makes the underlying assumptions about the concepts in a particular domain and the relationships among them explicit; and hence, helps to elucidate the meaning of expressions in a shared knowledge. As a result, ontology is a key component of the semantic web, which is founded on meaning, and is used to support applications that perform indexing, searching, retrieving and using of the shared knowledge distributed on the web.

Web ontologies are generally regarded as simple models of knowledge represented using Frames and descriptive logic [2]. The concepts in the model are defined as classes with attributes, and binary relations that capture the logical association among the concepts. The extensions of these concepts are the instances describing entities in a particular knowledge area. Web ontologies are encoded in web ontology languages like OWL [3], DAML [4] and RDF [5], and are usually developed using ontology editors such as Protégé [6], OntoEdit [7] and WebODE [8]. OWL has become the defacto web ontology language and Protégé is a popular open-source web ontology editor.

Ontology evaluation is an important step in the ontology development process. In principle, there are three aspects of an ontology that are evaluated upon: the syntax, the structure and the semantics of its conceptual definitions. This paper is concerned about the latter. Traditionally, the evaluation of the underlying structure and semantics of an ontology is based on a list of desirable qualities such as the consistency, completeness, conciseness and expandability, often emphasising on generality to enable the reuse of the ontology. However, when developing a new ontology or adapting an existing ontology, the purpose of use ought to be considered [9], and so, competency is a consequential quality that must be present [10].

Because ontology is used to specify the relationships among concepts that give rise to meaning of expression, the competency evaluation is performed over the concepts and relationships defined in an ontology to ascertain if the existing ontological commitments are adequate to support the ontology’s purposive mechanisms. It requires inspecting the defined concepts and their associated relationships on their appropriateness and use – a challenging task when dealing with large ontologies enriched with relationships because it is simply not feasible to trace the conceptual relationships on the editor windows. Often, a developer has to undertake additional preliminary work such as re-engineering the ontology and producing its graphical representation to facilitate the tracing of the conceptual
relationships in the ontology being evaluated. Although some editors provide visualisation support, it is still cumbersome to trace the intricate links on the miniaturised graphical representation of those ontologies.

At present, the support for competency evaluation is noticeably weak. Therefore, we have taken the initiative to develop a tool support for the Protégé ontology editor at the conceptualisation level that can assist a developer to perform the competency evaluation more effectively, in less time, and with reduced effort. In particular, the tool circumvents the additional preliminary work that is currently performed as part of the competency evaluation. The tool can also help to overcome mistakes due to human oversights and reduce the subjectivity that could arise during the tracing of the conceptual relations in the ontology.

2.0 COMPETENCY EVALUATION OF A WEB ONTOLOGY

A practicable approach for evaluating the competency of an ontology is prescribed by Gruninger and Fox [11]. The ontology developer first prepares a set of ‘competency questions’ that express the users’ requirements of this ontology. The ensuing analysis needs to validate the suitability of the ontology with respect to its intended use by checking if the ontology has the requisite concepts and relationships to formulate the competency questions, as well as their corresponding answers, which warrants for the inspection of the concepts and relationships in the ontology.

We exemplify the competency evaluation using a graphically reproduced segment of the Family ontology in Fig. 1, and some example competency questions below. The concepts in the figure are denoted within the boxes, while the relationships between a pair of concepts are indicated on the edges of the arrows. In the ontology, Gender and Person are defined as high level concepts, and Male and Female are defined as subconcepts of Gender. Likewise, Man and Woman are defined as subconcepts of Person. The subtype association between these concepts is indicated by an is-a relationship. These concepts are also related to each other and with themselves through relationships such as has-gender, has-wife and mother-of.

![Fig. 1: A graphically reproduced segment of the example Family ontology.](image)

In general, there are five basic types of relationships in a web ontology. We identify them as the normal relation, inverse relation, parallel relations, self-referencing relation and indirect relations.

A normal relation is a primary relation defined for a concept that directly relates it to another concept. For example has-gender, is-a and has-wife are examples of normal relations defined for Man in the ontology. An inverse relation (whose name is bracketed) is the opposite of a normal relation. For example, son-of is defined as the inverse relation of has-son, and vice versa in this ontology. Parallel relations exist when there are two or more relations connecting a pair of concepts. For example has-wife and has-sister are examples of parallel relations from Man to Woman. Self referencing relation refers to a relation that relates a concept to itself. For example, has-son and has-father are examples of self-referencing relations defined for Man. Indirect relations refers to the sequence of relationships that connect two concepts through one or more intermediate concepts. For example, has-mother and has-gender are indirect relations that
relate the concepts *Man* to *Female* through *Woman*. For the sake of reasoning efficiency, web ontologies are generally free of cyclic relationships with exception of the inverse and the self-referencing relations.

The *Family* ontology is created using the Protégé ontology editor. The Protégé ontology source file is encoded *Clips* format, which can be translated into a chosen web ontology language prior to use. Fig. 2 shows an excerpt of the encoded *Family* ontology saved in the *Family.pont* source file.

The following four example competency questions for the *Family* ontology. The underlined terms in the questions have been identified as the key terms that need to be mapped to certain concepts or relations in the ontology.

1. What is the *gender* of a *person*?
2. How is the *feminine gender* associated with a *man*?
3. Who is the *son* of whom?
4. Who are the *siblings* of a particular *child*?

In Fig. 3 we describe the progression of the competency analysis based on the second question. Here, the two terms in the question, *feminine gender* and *man* can be mapped to the ontology concepts *Female* and *Man*, respectively. In this case, we have established that the requisite concepts associated with this question are found in the ontology, otherwise the missing concepts or relations ought to be defined.

![Family ontology encoded in Clips format.](image)

Consequently, we need to trace all the relationships between these concepts in order to ascertain if the requisite relationships between these concepts are defined in the ontology. The latter is a semantic analysis task that is subject to understandability, which only humans can perform. However, we have developed the tool to support the initial task of tracing the conceptual relationships in the ontology; a devious task especially when evaluating the competency of large, rich ontologies.

![Preliminary steps for checking the admissibility of a competency question.](image)
3.0 RELATED WORK

Evaluation is a crucial step in the ontology development process [11, 12, 13]. Generally, there are three aspects of an ontology that are evaluated upon, the language syntax, the conceptual structure and the underlying semantics. The competency evaluation is encompassed by the semantics aspect of ontology evaluation.

Automated support for evaluating the syntactic aspects of an ontology such as syntax checkers, parsers and validators are often incorporated within an ontology editor. On the other hand, the evaluation of the structural and semantics aspects are analytical task that only human can perform effectively. Traditionally, this evaluation is based on a list of desirable qualities of an ontology such as consistency, coherence, completeness, conciseness, extensibility, expressiveness and minimal encoding bias [1, 12, 13, 14] evolved from best practices in conceptual modelling and design principles in knowledge and software engineering. More recently, tool supports for evaluating certain structural and semantics aspects of an ontology, especially along the functional dimension have surfaced. Examples are the graph topology metrics for measuring the complexity of an ontology’s conceptual structure [15], means of organising the conceptual taxonomic hierarchy [16], graph theory based algorithms for detecting problems in concept taxonomies [17], and clustering techniques for measuring ‘structural fit’ between an ontology and the corpus of knowledge used to develop it [18], are to name a few.

We are concerned about the competency evaluation of the web ontologies. Following the approach of Gruninger and Fox [11], we ascertain whether the competency questions and their corresponding answers can be represented using the concepts and relations defined in this ontology. Interestingly, ontology editors like Protege, OntoEdit and WebODE provide query support over the instances. Instances are extensions of the defined concepts used in the description of content in the knowledge base, which unfortunately cannot be plausibly created until the ontology has been developed. On the other hand, we really require the competency evaluation support during the iterative development of an ontology that undergoes successions of additions, deletions and modifications at the conceptualisation level. Clearly any attempt to extend the evolving concepts during the development phase is time consuming and useless effort. So, the existing query support over the instances probably can help in the competency evaluation of an ontology that has been implemented, but definitely not during its development.

Alternatively, the graphical representation of an ontology being developed could aid its visualisation. Such tool supports are available but are often limited to the visualisation of the concept hierarchies (i.e., is-a relationships). While the visualisation tools can help to graphically capture the ontology as a whole, it is rather cumbersome to trace the conceptual relationships on the miniaturised representation. Besides, the more established ontology visualisation tool such as the Jambalaya [19] still lacks the expressiveness we want (as mentioned in following section).

4.0 USING THE COMPETENCY EVALUATION SUPPORTING TOOL

Since web ontologies only allow binary relationships between a pair of concepts, we have identified four options by which a user can request the tool to trace the conceptual relationships in the ontology, i.e., by providing the name(s) of:

I. a particular concept,
II. a particular relation,
III. a concept and a relation associated with it, or
IV. a pair of concepts.

In case I, the tool lists all the relations that originate from the named concept. In case II, the tool lists all pairs of concepts that are directly related through the named relation. In case III, all the concepts directly associated with the named concept through the named relation are listed. Lastly, in case IV, the application lists all the direct and indirect relationships between the named source and the target concepts.

For example, recall that the concepts in the ontology that are associated with the second competency question in Section 2.0 are Female and Man. We want to find the existing relationships between these concepts and check if these relationships can sufficiently provide for that question. In this situation, we select option IV, and identify Female as the source concept and Man as the target concept. As a result, the application generates all possible relationships that exist between these two concepts as partly depicted in Fig. 4. Subsequently, the user can make use of the structured result to verify if all the requisite relationships between those concepts are defined in the ontology.
5.0 TOOL DESIGN AND IMPLEMENTATION

The tool is developed using the Java programming language and is casted as a backend application for the Protégé ontology editor. Therefore, the application reads the input directly from the Protégé ontology source file. The necessary information in the input file are extracted, internally organised and appropriately manipulated before the tool is availed for use. Fig. 5 shows the steps involved in the design and implementation of this tool.

5.1 Creation of the Intermediate Files

Once the ontology to be evaluated is identified, the Protégé ontology source file encoded in *Clips* format (see Fig. 2) will be processed, and the information about the defined concepts and the normal relationships will be extracted. The extracted terms will be reorganised in the form “Predecessor; Normal-relation; Successor” and stored in an intermediate file called “normalRelations.txt”.

The inverse relationships between the concepts, which are represented differently from the way the normal relationships are represented in the ontology file will be extracted separately and reorganised in the form “Normal-relation; Inverse-relation”, and stored in another intermediate file called “inverseRelations.txt”.

---

Fig. 4: An excerpt of the structured output elucidating the relationships between Female and Man.

Fig. 5: The tool design and implementation steps.
5.2 Construction of the Conceptual Adjacency Matrix (CAM)

The conceptual relationships in a web ontology can be captured as a directed graph and internally organised in an adjacency matrix, which allows easy access and manipulation of the stored information. Fig. 6 delineates the organisation of the information in the conceptual adjacency matrix (CAM).

First, the information about the direct relationships between the concepts in the ontology will be described. We begin by representing the information about the normal relationships between concepts in the “normalRelations.txt” file. As a result, a particular cell in the CAM can hold one or more assertions about the normal relations between a pair of concepts, represented with the aid of the concept-relationship structure pointed by the arrow in Fig. 6. For example, the information “Man; has-wife; Woman” will structure Man as the Predecessor, has-wife as the Normal-relation and Woman as the Successor.

Next, the information about the inverse relationships in the “inverseRelations.txt” file will be updated to the CAM by copying the inverse relationship names corresponding to the normal relationship names, and vice versa, in the concept-relationship structures. For example, the information has-husband will be represented as the Inverse-relation in the concept-relationship structures whose Normal-relation is has-wife.

Finally, the indirect relationships in the ontology are determined and updated to the CAM. For this, we adapt the Warshall’s algorithm to compute the transitive closure of the binary relationships in the ontology. The original Warshall’s algorithm simply manipulates logical values that are used to indicate the presence or the absence of path connecting two vertices on a graph. We have modified the algorithm to utilise the CAM instead. We also employed a marking scheme described in the next paragraph to represent the tracking information that facilitates the generation of the output.

![Fig. 6: Outline of the information organised in the CAM.](image)

A piece of tracking information is added to the CAM for each indirect relationship that is identified between a source concept and a target concept during the computation of the transitive closure. The information describes the immediate predecessors related to the target concept and is stored in the cell holding the information about relationships from the source to the target concepts (i.e., the source-to-target cell indexed to the source concept on the left and the target concept at the top in Fig. 6). For example, an indirect relationship from Man to Female through Woman is represented by a tracking information “Woman; has-gender; Female; gender-of” in the Man-to-Female cell, indicating that the source concept Man is related to the target concept Female through the concept Woman, the immediate predecessor of the target concept. Note that Woman is related to the target concept Female through the normal relation has-gender whose inverse relation is gender-of. Subsequently, we draw on this tracking information when tracing the conceptual relationships in the ontology.

6.0 TESTING THE TOOL

The last step is the testing of the tool’s ability to correctly process the four types of user requests mentioned in Section 4.0 and to accurately trace the conceptual relationships involving the five types of relations identified in Section 2.0. For this we used the prototype ontologies we have developed, one of which is the Family ontology described in Section 2.0. Besides our ontologies, we also tested the tool on large ontologies, and ontologies with richly defined relationships that are developed by others, which can be found in the Protégé ontology library [20]. An example is the Wine ontology, which has 113 concepts and 111 relations. In all our test cases, the tool has demonstrated its capability to respond successfully to the user requests and produce the desired output.
We also conducted a simple usability analysis by involving three users who are familiar with ontology development. The users were given a set of competency questions and a test ontology, and were asked to verify whether the concepts and relations defined in the ontology are adequate to represent the key terms in the questions and their corresponding answers. They performed the analysis, first tracing the conceptual relationships on the ontology editor windows with the aid of paper and pencil, and then using the tool we have developed. All the users acknowledged that the tool has helped to reduce the time and effort required to perform the competency evaluation on the ontology.

7.0 CONCLUSION

This work is built upon the reasonable assumption that web ontologies must have adequate ontological commitments to support their purpose of use. Therefore, a semantic aspect of an ontology that must be evaluated is whether the definitions in the ontology competently satisfy its use and requirements. For this, we adopt an established approach to evaluate the competency of an ontology with the help of a predetermined set of ‘competency questions’. Currently the competency evaluation on an ontology being developed is by and large performed manually. Consequently, we have taken the initiative to develop a tool to support the competency evaluation of web ontologies. It is implemented as a backend application to Protégé, a popular web ontology editor by following the steps delineated in Section 5.0. We summarise below the significant benefits of using this tool.

- The preliminary steps involved in the competency evaluation under the conventional approach, such as skimming through the ontology editor windows to extract the relevant information about the ontology and/or reengineering the encoded ontology to produce its graphical representation have been circumvented.
- The issues related to the visualisation of rich and large ontologies, and the difficulty encountered when tracing the conceptual relationships in such ontologies has been overcome by focussed non-visual presentation of the ontology.
- The automated support dispense with much of the human effort by precisely tracing out all the conceptual relationships in the ontology based on the user requests, without succumbing to mistakes due to human oversights.
- The structured result serves as a useful and reliable source of reference for the developer who uses it to verify the adequacy of the concepts and relationships in the ontology, as part of the competency evaluation.

In future work, we intend to strengthen the functionality of the tool by accommodating admissible cyclic relations (other than inverse and self-referencing relations), extending the user request options, accepting less structured user input and providing a focussed visual presentation and summary to accompany the resulting output.

REFERENCES


BIOGRAPHY

Muthukkaruppan Annamalai is a computer science lecturer at the Faculty of Information Technology and Quantitative Sciences, UiTM, Malaysia. His current research areas include Semantic Web, Knowledge Management, Knowledge Modelling and Representation, with particular interest in the development of decision-support applications and tools.

Noor Hasimah Ibrahim Teo has just completed her Master of Computer Science at UiTM. At present, she is lecturing at the Faculty of Information Technology and Quantitative Sciences, UiTM, Malaysia. She is continuing her interest in ontology engineering and tool supports.
THE USE OF A DECISION SUPPORT FRAMEWORK IN A MULTI-CRITERIA PURCHASE DECISIONS

P.D.D. Dominic¹, Shakirah Mohd Taib², Emy Elyanee Mustapha³
Department of Computer and Information Sciences,
Universiti Teknologi PETRONAS, 31750 Tronoh, Perak, Malaysia.
E.mail: dhanapal_d@petronas.com.my, shakita@petronas.com.my, emyelyanee@petronas.com.my

ABSTRACT

Today’s global business environment characterized by unprecedented competitive pressures and sophisticated customers that demand speedy solutions creates a biggest set of potential suppliers to evaluate and select. Increasing global competition has made many companies recognize that competitive manufacturing in terms of low cost and high quality is crucial for success. Also they are facing rapid changes stimulated by technological innovations and changing customer demands. These companies are realized that the effort to obtain products at the right time, in the right quantity, with the right quality at the right source is crucial for their survival. This type of problem will require effective decisions of supplier selection for purchase of any products. To deal with the complexities of the supplier selection process, the authors proposed a Decision support system (DSS) framework for purchase of computers, which will helps the top and middle level managerial decision. The framework integrates the various elements of the system such as data warehouse, mathematical model, interface engine etc. The objective of the paper is to construct a model based decision support system framework for supplier selection for purchase of computers with numeric illustration in practice. Which maximize the benefit to the organization and minimizes the total cost of purchase.

Keywords: Decision Support framework, Supplier selection, what if analysis, Total cost of purchase, and attribute types.

1.0 INTRODUCTION

The selection of supplier for long period it has been considered as an important decision of purchasing department in any organization. Attributing to the significance of the problem a wide range of research has been done starting from 1960’s. The authors [1], [2] explained about selecting the right suppliers that will furnish them with the necessary products, components, and materials in a timely and effective manner significantly reduces the purchasing cost, quality problems, and long lead times to improve corporate competitiveness. Different kind of qualitative and quantitative approaches are proposed for supplier selection by various researchers. A common feature of all this research is that the problem of vendor selection is considered as a multi objective decision making problem. Also many researchers tried to quantify the qualitative aspect pertaining to this problem. Decision Support Systems are a specific class of computerized information system that supports business and organizational decision-making activities. A properly designed DSS is an interactive software-based system intended to help decision makers compile useful information from raw data, documents, personal knowledge, and/or business models to identify and solve problems and make decisions. Many authors [3], [4], [5], and [6] defines the concept of a decision support system is extremely broad and its definitions vary depending on the author’s point of view. It can take many different forms and can be used in many different ways, according to [7] it is an interactive, flexible, and adaptable computer-based information system, especially developed for supporting the solution of a non-structured management problem for improved decision making. The objective of the paper is to construct a model based decision support system framework for supplier selection for purchase of computers. Like any other DSS contains a user interface, a database, and a mathematical model management system. In the working of model base, weightage mathematical model has been proposed. The paper is organized as follows: we present the review of relevant literature in the next section. We then present the research methodology, followed by detailed discussion and conclusion.

2.0 LITERATURE REVIEW

The area of DSS has seen in many research papers. Some of the papers are reviewed during the research paper shed light on this critical decision making aid. The authors [8], [9] in his paper developed a model to deal with the complexities of altering schedules and acquiring additional resources when MIS plans are too ambitious. A mixed integer-programming model was developed for the information services division of a state governmental agency.

Decision support systems have been used in various areas said by [10]. Multi-dimensional taxonomy of systems is proposed as an organizing framework for research in the area. Three environmental contingencies are identified as
critical to GDSS design: group size, member proximity, and the task confronting the group. Potential impacts of GDSS on group processes and outcomes are discussed, and important constructs in need of study are identified. An agent-based framework for building decision support system proposes a framework for building decision support systems using software agent technology to support organizations characterized by physically distributed, enterprise-wide, heterogeneous information systems. Intelligent agents have offered tremendous potential in supporting well-defined tasks such as information filtering, data mining and data conversion.

The authors [11] says to deal with the complexities of the supplier selection process, an integration of Quality Function Deployment, Analytical Hierarchy Process and Preemptive Goal Programming techniques is proposed to form a matrix is used to display the degree of relationship between each pair of requirement for suppliers and supplier evaluating criterion.

The authors [12] developed a DSS for the scheduling of the operations of contractors who travel from farm to farm to harvest maize. The system provides a useful addition to the contracting scheduler’s toolkit and an interesting application of the DSS approach to a critical area of agriculture.

3.0 RESEARCH METHODOLOGY

Our paper critically evaluates all the criteria in a purchase process. The primary research objective of this paper is selection of a suitable supplier for purchase of computers by using a model based decision support framework. This Research paper is mainly intent to bring out a solution in the form of decision to select a supplier on the basis of some established criteria. The information about the factors influencing the purchase decision was observed and collected from the supplier through quotations and related websites. Also the information collected on DSS and their modeling were of secondary data type. With this, a suitable DSS framework was proposed for the study and the data of criteria values, applied to the model for analysis with the help of an illustration.

4.0 DATABASE MANAGEMENT SYSTEM

The database management system defined by [13] allows user to store data in an organized form and retrieve it on the basis of specified selection criteria. To a database expert, a defining characteristic of a true database is its ability to integrate several types of data, typically stored in several files. Data managers can be quite useful as personal filing systems, allowing a decision maker to recall data quickly when it is needed to make a decision. Data managers can also serve as “back end” to a system in a higher category. For example, a data manager could supply input into a model that calculates the result to be expected from a new marketing program aimed at selected customers from the list. When the data manipulation requirements of such DSS are modest, data managers provide an easy way to meet them. Any database reflects a conceptual data model. The data model specifies the entities about which the database contains data and the ways in which these entities are related.

5.0 MODEL BASE MANAGEMENT SYSTEM

A supplier selection problem typically consists of four phases, namely problem definition, formulation of criteria, qualification of suitable supplier and final selection of the ultimate supplier. In each phase a different set of models is appropriate. Many researchers [14], [15], [16], [17], [18], [19], and [20] suggested several methods for supporting the supplier selection process have been proposed in the literature. The most important of them are presented as follows: categorical methods, data envelopment analysis, cluster analysis, cased based reasoning, linear weighting models, total cost of ownership models, statistical models, discrete choice analysis, mathematical programming, Brown-Gibson and Artificial Intelligence. In this research paper the authors proposed Weightage mathematical model is given below for supplier selection. Where Wi is the weight of each attributes and Xi is the normalized values of corresponding attributes.

\[
\text{Total Score} = \frac{\sum W_i X_i}{\sum W_i}
\]

6.0 DATA NORMALIZATION
Data normalization has also to be performed so that all the data have a range of 0 to 1. In real life purchase decision for each of the product attributes, have different magnitudes and units. For example, one variable may be in Ringgits Malaysia with a value between 3000 and 3800. Another variable may be in warranty periods with a value between 2 and 3 years. When we use these variables directly in a weightage model, the price variables in a particular supplier having a much larger value will influence the model much more than the supplier having a lower value. But this is not the correct way of doing it, because any organization interested to buy the product from supplier has quoted a lower price value. To avoid this biasing of the weights by larger variables, all variables are normalized in the range of 0 to 1. Moreover our threshold attributes can be classified into two category, called Maximum type and Minimum type of attributes accordingly the following equations are formed. When all the variables are normalized, they will have unbiased influence on the attributes weights.

Maximum type of attribute = \frac{\text{Maximum Value} - \text{Vendor Value}}{\text{Maximum Value} - \text{Minimum Value}}

OR

\[ X' = \frac{X_{\text{max}} - X}{X_{\text{max}} - X_{\text{min}}} \]  ------ (1)

Minimum type of attribute = \frac{\text{Vendor Value} - \text{Minimum Value}}{\text{Maximum Value} - \text{Minimum Value}}

OR

\[ Y' = \frac{Y - Y_{\text{min}}}{Y_{\text{max}} - Y_{\text{min}}} \]  ------ (2)

7.0 DIALOG GENERATION MANAGEMENT SYSTEMS

In What - If analysis, an end user makes changes to variables, or changes in the weight of the attributes, or relationships among variables, and observes the resulting changes in the values of other variables using spreadsheet. The user can command this spreadsheet programme to instantly recalculate all affected variables in this spreadsheet. A managerial use would be interested in observing and evaluating any changes that occurred to values in the spreadsheet. This type of analysis would be repeated until the manager was satisfied with what the result revealed about the effects of various possible decisions. So, this research has taken the option of What – If analysis, so that any variations in the factors and the impact of the same on decisions can be analyzed.

For example, if the end user that is the decision maker interested to change the Price of a particular supplier L1 from table 2 values after negotiation process to RM 2838 which is currently at RM 2888, and asks the system for rank the supplier based on the new set of values. Then the model recalculates the affected variables and informs the end user about new ranking list.

8.0 LIMITATION

The weightages of each attributes are assigned by the decision makers based on the nature of the product of purchase or based on the end users experiences which may lead to variations in the final decision. However the decision makers are interested to reduce the gap between the ultimate correct decision and poor decision.

9.0 NUMERICAL EVALUATION
A numerical illustration is demonstrated in this section to understand the behavior of mathematical model used to make complex decision for purchase of computers.

Table 1: Multiple attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Unit</th>
<th>Threshold</th>
<th>Weightage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor (PRO)</td>
<td>MHz</td>
<td>Min</td>
<td>0.20</td>
</tr>
<tr>
<td>Hard-disk (HD)</td>
<td>GB</td>
<td>Min</td>
<td>0.10</td>
</tr>
<tr>
<td>Monitor (MO)</td>
<td>Inch</td>
<td>Min</td>
<td>0.05</td>
</tr>
<tr>
<td>L1/L2 Cache (CAC)</td>
<td>MB</td>
<td>Min</td>
<td>0.05</td>
</tr>
<tr>
<td>RAM (RAM)</td>
<td>MB</td>
<td>Min</td>
<td>0.15</td>
</tr>
<tr>
<td>Power Consumption (POW)</td>
<td>Watts</td>
<td>Max</td>
<td>0.05</td>
</tr>
<tr>
<td>Warranty (WR)</td>
<td>In years</td>
<td>Min</td>
<td>0.10</td>
</tr>
<tr>
<td>Delivery Time (DT)</td>
<td>In weeks</td>
<td>Max</td>
<td>0.05</td>
</tr>
<tr>
<td>Price (PRI)</td>
<td>In RM</td>
<td>Max</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Table 1 describes all selected attributes for evaluation with unit, threshold and their weightages. Table 2 describes a sample of 13 models with standard operating systems and processor; and their product attributes, corresponding data values are presented. These data collected through quotations, advertisements and from their official websites.

Table 3 figures are obtained from table 2 after data normalization. This is very important because, by nature different attributes are in different units of measurements for example warranty period is in years and price is in Malaysian Ringgits, the researchers interested to convert the different units into a uniform unit, this is possible through data normalization. Also an organization interested to buy a quality product with lower price in the same time they are interested to buy which vendor giving more years as a warranty period. These two attributes are in contradicting in nature, through data normalization the researchers converting the conflicting direction into a uniform direction.

In Table 4 the final decision making information are available. Higher the value leads to high advantages or benefits to the purchaser. From table 4 one who identified that vendor A2 got more score stood first in the vendor selection list, followed by L1 and D2 vendors or brands. So the decision maker can take the decision based on these figures. Still the organization or the user interested to make further action they can invite those vendors for negotiations, if there is any changes in their data values use sensitivity analysis through what if analysis and rework the final table and make final decision for placement of purchase order.

Table 2: Comparative statement

<table>
<thead>
<tr>
<th>S.No</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Att./Mod.</td>
<td>D1</td>
<td>D2</td>
<td>D3</td>
<td>A1</td>
<td>A2</td>
<td>HP1</td>
<td>HP2</td>
<td>L1</td>
<td>L2</td>
<td>L3</td>
<td>L4</td>
<td>L5</td>
<td>L6</td>
</tr>
<tr>
<td>PRO</td>
<td>2.8</td>
<td>2.8</td>
<td>1.8</td>
<td>3.06</td>
<td>3.4</td>
<td>3.4</td>
<td>3.2</td>
<td>3</td>
<td>3.4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2.8</td>
</tr>
<tr>
<td>HD</td>
<td>80</td>
<td>80</td>
<td>160</td>
<td>80</td>
<td>160</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>MO</td>
<td>17</td>
<td>17</td>
<td>19</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>CAC</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>RAM</td>
<td>512</td>
<td>512</td>
<td>512</td>
<td>256</td>
<td>512</td>
<td>512</td>
<td>512</td>
<td>512</td>
<td>512</td>
<td>256</td>
<td>256</td>
<td>256</td>
<td>256</td>
</tr>
<tr>
<td>POW</td>
<td>250</td>
<td>250</td>
<td>250</td>
<td>135</td>
<td>250</td>
<td>250</td>
<td>250</td>
<td>250</td>
<td>250</td>
<td>250</td>
<td>250</td>
<td>250</td>
<td>250</td>
</tr>
<tr>
<td>WR</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>DT</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>PRI</td>
<td>2149</td>
<td>2449</td>
<td>2949</td>
<td>2388</td>
<td>2838</td>
<td>2874</td>
<td>2414</td>
<td>2888</td>
<td>3798</td>
<td>2158</td>
<td>2468</td>
<td>2188</td>
<td>2588</td>
</tr>
</tbody>
</table>

Table 3: Normalized values of the attributes
Table 4: Final decision table

<table>
<thead>
<tr>
<th>S.No</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Att./Mod.</td>
<td>D1</td>
<td>D2</td>
<td>D3</td>
<td>A1</td>
<td>A2</td>
<td>HP1</td>
<td>HP2</td>
<td>L1</td>
<td>L2</td>
<td>L3</td>
<td>L4</td>
<td>L5</td>
<td>L6</td>
</tr>
<tr>
<td>PRO</td>
<td>0.63</td>
<td>0.63</td>
<td>0.79</td>
<td>1</td>
<td>1</td>
<td>0.88</td>
<td>0.75</td>
<td>1</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.63</td>
<td></td>
</tr>
<tr>
<td>HD</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MO</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CAC</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RAM</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>POW</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>WR</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>DT</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>PRI</td>
<td>1</td>
<td>0.82</td>
<td>0.51</td>
<td>0.86</td>
<td>0.58</td>
<td>0.56</td>
<td>0.84</td>
<td>0.55</td>
<td>0</td>
<td>0.99</td>
<td>0.81</td>
<td>0.98</td>
<td>0.73</td>
</tr>
</tbody>
</table>

10 CONCLUSION

Purchase decision is a central element of cost reduction, which may be used as a competitive advantage of any company under today’s competitive world. So organizations are interested to find out the means to minimize the total cost of purchase. Our research paper intended to examine all the influencing factors for selection of a supplier and place an order with that particular supplier not only this time for subsequent purchases over a period of time. In this paper, we present a systematic decision procedure to be used in supplier selection process, which has been traditionally based on expert opinions. This model based decision support system will enable the organization or user to produce a better result in terms of to find the best supplier and to place the optimum order quantities among suppliers such that the total cost of purchasing is minimum. The decision approach presented in this paper can be easily extended for any purchase of equipments. For further research include qualitative variables like performance of the supplier, supplier location nearness to the organization etc along with quantitative variables. In this case selection of mathematical model may be different in nature.

REFERENCES


©FCSIT, UM 2007


BIOGRAPHY

P.D.D.Dominic obtained his M.Sc degree in operations research in 1985, MBA from Regional Engineering College, Tiruchirappalli in 1991, Post Graduate Diploma in Operations Research in 2000 and Obtain his Ph.D during 2004 at Alagappa University, Karaikudi, India. Since 1992 he has held the post of Lecturer in the Department of Management Studies, Regional Engineering College, Tiruchirappalli- 620 015. Currently he is working as a Senior Lecturer in the department of computer and information sciences, Universiti Teknologi PETRONAS, Malaysia. His fields of interest are Operations Research, Scheduling, and Decisions Support Systems. He has published technical papers in International, National journals and conferences.

Shakirah Mohd Taib obtained her Master degree in Computing in 2004 from University of Tasmania, Australia. Since 2004 she is a lecturer in the Computer and Information Sciences Department, Universiti Teknologi PETRONAS. Her fields of interest are Knowledge Management, Artificial Intelligence, Web Mining and Decisions Support Systems. She has published technical papers in International, National journals and conferences.
Emy Elyanee Mustapha obtained her Master of Information Technology in 2005, from Australian National University, Australia. Since 2006 she is working as a Lecturer in the department of computer and information sciences, Universiti Teknologi PETRONAS, Malaysia. Her fields of interest are E-Commerce, Decision Support Systems and Human Computer Interaction. She has published technical papers in International, National journals and conferences.

**APPENDIX**

List of Abbreviation

<table>
<thead>
<tr>
<th>Att.</th>
<th>Attributes</th>
<th>L4</th>
<th>Lenovo IBM 4 (A52 Tower)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Dell 1 (E520)</td>
<td></td>
<td>L5</td>
</tr>
<tr>
<td>D2</td>
<td>Dell 2 (E520)</td>
<td></td>
<td>L6</td>
</tr>
<tr>
<td>D3</td>
<td>Dell 3 (E520)</td>
<td></td>
<td>PRO</td>
</tr>
<tr>
<td>A1</td>
<td>Acer 1(FH)</td>
<td></td>
<td>HD</td>
</tr>
<tr>
<td>A2</td>
<td>Acer 2 (2000-P34P)</td>
<td></td>
<td>MO</td>
</tr>
<tr>
<td>HP1</td>
<td>Hewlett-Packet 1 (dx2700)</td>
<td></td>
<td>CAC</td>
</tr>
<tr>
<td>HP2</td>
<td>Hewlett-Packet 2 (dx2100)</td>
<td></td>
<td>RAM</td>
</tr>
<tr>
<td>L1</td>
<td>Lenovo IBM 1 (M52 USFF)</td>
<td></td>
<td>POW</td>
</tr>
<tr>
<td>L2</td>
<td>Lenovo IBM 2 (M52 Tower)</td>
<td></td>
<td>WR</td>
</tr>
<tr>
<td>L3</td>
<td>Lenovo IBM 3 (E50 Tower)</td>
<td></td>
<td>DT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PRI</td>
</tr>
</tbody>
</table>
ABSTRACT

Currently, word-processing software for Malay language only exists in the form of words translator or spellchecker. This paper aims to complement the existing word-processing software by presenting a prototype of Malay sentence parser using top-down parsing technique. The prototype is able to illustrate the structure of a grammatically correct sentence, determine if a sentence is grammatically correct, and semantically parse a sentence. In evaluating the prototype, sentences which were inputs to the prototype, were randomly provided by experts in the structure of Malay language grammar. Experimental results showed that the prototype was able to correctly parse the provided sentences by approximately 81%.

Keywords: Artificial intelligence, natural language processing, parser, top-down parsing, grammar

1.0 INTRODUCTION

Grammar is the rules and patterns, which we combine words, clauses and phrases to provide meaning to our daily communication. The study of grammar relates to language parsing. According to Mohanty and Balabantaray [1], in the context of natural language processing, parsing may be defined as the process of assigning structural description to sequence of words in a natural language. The structural description assignment of words depends upon the grammar according to which the parser attempts to analyze the sequence of symbols presented to it.

The Malay language has been the national language of Malaysia (formally known as Malaya) since 1955. The language belongs to the western group of Austronesian family and is spoken by people throughout Malaysia, Brunei, Indonesia and Singapore [2]. The grammar of Malay language has three main morphological processes: reduplication, compounding, affixation [2]. Reduplication is the process in which a base or some part of the base is repeated. Two types of reduplication are reduplication proper (partial or full) and rhyming & chiming (reduplication with phonetic change). Meanwhile, the compound form is a construction having two or three free forms, as its constituent’s forms may either be a root or derived. Affixation is a process in which a base may be extended by one or more affixes. It is the most common and widely used of the three morphological processes. Affixes may be classified as prefixes, suffixes, infixes and circum-fixes. It plays an important role in the setting of the syntactical and the semantic features of a word form. Affixes could change the grammatical class of the word form or even change the meaning of the word form.

Currently, the existence of word-processing software of Malay language is in the form of word translator or spellchecker but not grammar checker. An example of available word-processing software of Malay language is Dewan Eja [3], which is a complete reference consisting of spellchecker, dictionary and encyclopedia. Dewan Eja’s main function is the spellchecker that can be integrated with such Microsoft Office applications as Microsoft Word and Microsoft Excel to show spelling mistakes and to suggest correct spellings. Its unique feature enables users to find the meaning of any word by only highlighting the word in any Windows application.

This paper presents a sentence parser for Malay language, which illustrates how a grammatically correct sentence is built and what phrases compose a complete sentence. The prototype is able to draw the structure of a correct sentence and identify whether a sentence is grammatically correct.

2.0 TECHNIQUES IN SENTENCE PARSING

There are systems developed to explain grammatical structure for English. However, no work on the development of Malay sentence parser that explains Malay grammatical structures has been reported in the literatures.

Fox and Bowden [4] developed a system called GRADES (Grammar Diagnostic Expert System), primarily for non-native English speakers. GRADES perform its diagnostic task through the application of classification and pattern
matching rules instead of through parsing. GRADES is a rule-based diagnostic expert system in which using a set of relevant rules for a certain type of error, it narrows down a specific category for identifying the type of error in the form of hierarchical classification.

Another intelligent parsing algorithm created by Mohanty and Balabantaray considers both syntactical and semantic analyses of a sentence during parsing [1]. Their algorithm is sensitive to a wide range of language details; hence, capable of providing better outputs.

3.0 MALAY LANGUAGE SENTENCE PARSER

In this project, the top-down approach [5] is applied to the Malay language grammatical rules. In addition, Malay words applicable to humans and animals are identified to achieve a basic level of semantic parsing on top of syntactical parsing. The developed prototype expects a user to input a sentence before parsing it for grammatical errors check. Next, the prototype should be able to display the tree diagram for a grammatical structure of the sentence so that the user would know the structure arrangement of a correct sentence.

3.1 System Architecture

Fig. 1 shows the system architecture with the main elements of the Malay sentence parser.

- The Text Parser

For the technical structure of the Text Parser, it consists of several parts. Each part will be linked to the parser, the central file loading or consulting all other parts and starts the application. Prolog is used to develop the checking engine.

- Lexicon

Lexicon contains all words arranged according to their types. The words are collected from Kamus Dwibahasa Oxford Fajar 2nd Edition [6]. In total, Lexicon consists of more than 3000 words.

- Grammar Rules
Grammar Rules contain syntax of all phrases of the text. The rules consist of part of speech arranged in the top-down parser form.

3.2 Rules of Malay Language Grammar

Although Malay language is spoken throughout Malaysia, Indonesia, Brunei and Singapore [2], for the purpose of this project, the authors include rules of Malay language grammar used in Malaysia, collected from Tatabahasa Dewan Edisi Baharu [7] and Malay Grammar for Academics and Professionals [8]. The rules are then constructed using the top-down parsing technique. Top-down parsing is a basic predictive parser [5] in which the next production to be applied is predicted. In the case of incorrect prediction, alternative predictions are stacked and will be revisited if necessary. The leading item of the current sentential form is crossed off if it matches against the next word in the sentence. The syntaxes formula is used to arrange and construct words until they become a complete sentence. This grammar formula is divided into two types: Rumus Struktur Frasa (Phrase Structure Rules) and Rumus Transformasi (Transformation Rules).

Fig. 2 shows the position and relationship between these two rules. The Phrase Structure Rules is situated on the base component to generate a structure description to produce basic sentence or head sentence. These rules are meant to generate a tree diagram that illustrates the grammatical relationship between many types of phrase in a sentence. A syntax representation or sentence is the inner structure of the overall structure. This inner structure may become an input to the component containing the Transformation Rules. The Transformation Rules is powerful enough to drop, add, change or rearrange the word elements.

The most basic grammar rules lined out in [7] and [8] are shown in Fig. 3.
FA $\rightarrow$ (KB) + (KPeng) + Adj + [Adj] + (Ket) + (AKomp)
FS $\rightarrow$ (KB) + SN + (KNArah) + FN + (AKomp)
FS $\rightarrow$ (KB) + SN + (KNArah) + FN + (Ket)

Fig 3: Malay grammar phrase structure general elements

The descriptions of the rules in Fig. 3 are as follows:

- **A** consists of **S** and **P**.
- **S** must consist of **FN**.
- **P** may consist of single **FN**, **FK**, **FA** and **FS** or combination of **FN**, **FK**, **FA** and **FS**.
- Optional elements are parenthesized. An element in bracket implies that the element may be part of the phrase to support the required element. For example, **FN** may consist of **Bil**, **PenjBil**, **Gel**, **KNInt**, **Pen** and **Pent** or combination of the stated subsection. Bracketed **KNInt** may be part of **FN** to support non-bracketed **KNInt**. Examples of the bracketed **KNInt** usage are *ibu bapa* and *kerusi meja*.
- Similar to **FN**, phrases in **FK**, **FA** and **FS** may consist of elements on the right-hand side of their respective rules or combination of the stated subsection in their respective rules. For **FA**, bracketed **Adj** is used to support non-bracketed **Adj**. Examples of the bracketed **Adj** usage are *hitam manis*, *tinggi lampai*, and *kaya raya*. Both *hitam* and *manis* are adjectives, and so are *tinggi*, *lampai*, *kaya*, and *raya*.
- The entire phrases are then identified to be applicable to humans or animals for the purpose of arranging the structure for the semantic part of the sentence.

Each element in the grammar rules used to develop the sentence parser prototype is described in Table 1.

Table 1: Description of elements used in Malay grammar phrase structure

<table>
<thead>
<tr>
<th>Element</th>
<th>Description in Malay (English)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Ayat (Sentence)</td>
</tr>
<tr>
<td>Adj</td>
<td>Adjektif (Adjective)</td>
</tr>
<tr>
<td>AKomp</td>
<td>Ayat komplemen (Complementary sentence)</td>
</tr>
<tr>
<td>Bil</td>
<td>Bilangan (Numeric)</td>
</tr>
<tr>
<td>FA</td>
<td>Frasa Adjektif (Adjective Phrase)</td>
</tr>
<tr>
<td>FK</td>
<td>Frasa Kerja (Verb Phrase)</td>
</tr>
<tr>
<td>FN</td>
<td>Frasa Nama (Noun Phrase)</td>
</tr>
<tr>
<td>FS</td>
<td>Frasa Sendi (Prepositional Phrase)</td>
</tr>
<tr>
<td>Gel</td>
<td>Gelaran (Title)</td>
</tr>
<tr>
<td>KB</td>
<td>Kata Bantu (Auxiliary)</td>
</tr>
<tr>
<td>Ket</td>
<td>Keterangan (Explanation)</td>
</tr>
<tr>
<td>KKtr</td>
<td>Kata Kerja Transitif (transitive verb)</td>
</tr>
<tr>
<td>KKttr</td>
<td>Kata Kerja Tak Transitif (intransitive verb)</td>
</tr>
<tr>
<td>KNArah</td>
<td>Kata Nama Arah (Direction)</td>
</tr>
<tr>
<td>KNInt</td>
<td>Kata Nama Inti (Head noun)</td>
</tr>
<tr>
<td>KPeng</td>
<td>Kata Penguat (Intensifier)</td>
</tr>
<tr>
<td>Obj</td>
<td>Objek (Object)</td>
</tr>
<tr>
<td>P</td>
<td>Predikat (Predicate)</td>
</tr>
<tr>
<td>Pel</td>
<td>Pelengkap (Complement)</td>
</tr>
<tr>
<td>Pen</td>
<td>Penerang (Description)</td>
</tr>
<tr>
<td>PenjBil</td>
<td>Penjodoh Bilangan (Numerical Coefficient or Classifier)</td>
</tr>
<tr>
<td>Pent</td>
<td>Penentu (Determiner)</td>
</tr>
<tr>
<td>S</td>
<td>Subjek (Subject)</td>
</tr>
<tr>
<td>SN</td>
<td>Sendi Nama (Preposition)</td>
</tr>
</tbody>
</table>

3.3 Sample Outputs

Fig. 4 shows the tree structures generated by the prototype for the input sentences “sekawan lembu sedang melintasi jalanraya tersebut” and “fatimah mengusahaakan perniagaan di Ipoh”. Each generated tree structure is based on the sentence applicability to either *haiwan* (animal) or *orang* (human).
The tree structure in Fig. 4(a) shows a correct sentence consisting of Animal FN as $S$, and FK as $P$. Inside the FN it has numeric value for Animal and noun for animal, while inside the FK, it has the auxiliary KB, transitive verb $KKtr$, object $Obj$ and explanation $Ket$. The second example in Fig 4(b) consists of Human FN as $S$ and Human FK as $P$. The FN consists of Human FN, while the FK consists of transitive verb $KKtr$, object $Obj$ and explanation $ket$.

Fig. 5 shows an example in which a user entered a sentence containing semantic error.

For the sentence ‘Bapa sedang meragut rumput’, the word ‘meragut’ is reserved for animal’s verb and the word ‘bapa’ falls under human noun. So, in the sentence, $S$ does not match with $P$.

4.0 RESULTS AND DISCUSSIONS

An evaluation for the prototype was conducted involving Malay language experts. The experts were Malay language teachers from primary schools (standard 1-6), lower secondary (form 1-5) and upper secondary (form 6) schools in two towns of Perak state: Kampar and Seri Iskandar. They were asked to provide sentences randomly. Fifty teachers participated and a total $N$ of 520 grammatically correct sentences was provided by those teachers to be inputs to the prototype. The average recall $Avg$ (Eq. 1) and the weighted average recall $WAvg$ (Eq. 2) of correctly parsed sentences were reported using the following equations:

$$\text{Avg} = \frac{\sum_{i=1}^{k} B_i}{A_i} \times 100\%$$  \hspace{1cm} (Eq. 1)

where $k = 3$ represents three school levels, $A$ the number of test cases, and $B$ the total number of sentences correctly parsed.
\[
WAvg = \frac{\sum_{i=1}^{k} B_i}{N} \times 100\%
\]

where \( k = 3 \) represents three school levels, \( N \) the total number of test cases, and \( B \) the total number of sentences correctly parsed.

The weighted average recall was calculated mainly because the number of test cases was different for the three school levels. The results are shown in Table 2.

<table>
<thead>
<tr>
<th>School Level</th>
<th>Test Cases Provided by Experts (A)</th>
<th>Total Sentences CorrectlyParsed (B)</th>
<th>Recall: ((B/A) \times 100%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary</td>
<td>150</td>
<td>135</td>
<td>90.0 %</td>
</tr>
<tr>
<td>Lower Secondary</td>
<td>250</td>
<td>208</td>
<td>83.2 %</td>
</tr>
<tr>
<td>Upper Secondary</td>
<td>120</td>
<td>85</td>
<td>70.8 %</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td>81.3 %</td>
</tr>
<tr>
<td>Weighted Average</td>
<td></td>
<td></td>
<td>82.3 %</td>
</tr>
</tbody>
</table>

Based on the experiment, the results show that the prototype was able to achieve the average recall (weighted average recall) rate of 81.3% (82.3%) from the sentences provided by the experts. The prototype scored higher for sentences provided by the Malay language teachers from primary and lower secondary schools while lower score was obtained from sentences given by the teachers from upper secondary school. The prototype was more suitable for students who were in the lower secondary schools and primary schools. The prototype was less suitable for the form six students because the Malay language taught for form six students was more complex than what the prototype offered.

5.0 CONCLUSION

Grammar checkers for other languages such as English or Swedish [9] already exist but currently, the existence of word-processing software of Malay language is only available in the form of words/phrase translator or spellchecker. The development of Malay sentence parser could complement existing features in word processors and potentially help students understand Malay grammar better. By using a simple top-down parsing approach, the prototype is capable of checking whether Malay sentences are structurally correct. It is also sensitive to the words used for animals or humans. Hence, the prototype will check for syntactic part and semantic sensitivity of the sentence. The prototype is able to illustrate users the structure of a correct sentence by displaying a tree diagram with detailed section of the sentence structure. Chin in [10] suggested that a good understanding of grammar helps improve writing skills. A Malay sentence parser potentially helps students understand Malay grammar and construct a sentence better; hence, improving their writing skills. However, the prototype could not identify any single error of grammar’s component. Generally, it only detects whether the sentence entered by user contains errors. This will lead to the future work of the project.

REFERENCES


BIOGRAPHY

Ahmad I. Z. Abidin obtained his MS Computer Science from State University of New York at Binghamton, USA. Currently, he is a lecturer at the Department of Computer and Information Sciences (CIS), Universiti Teknologi Petronas (UTP). His research areas include artificial intelligence, data mining and e-commerce.

S. P. Yong obtained her Master of Information Technology from UKM. She is a lecturer at the Department of CIS, UTP. She is currently pursuing her PhD at University of Otago, New Zealand. Her research areas are artificial intelligence and data mining.

Rozana Kasbon received her Master’s degree in Information Technology from UiTM. She is currently a lecturer at the Department of CIS, UTP. Her research areas are intelligent systems and databases.