Journal of Theoretical and Applied Information Technology
15th October 2011. Vol. 32 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PSEUDO GENETIC AND PROBABILISTIC-BASED FEATURE SELECTION METHOD FOR EXTRACTIVE SINGLE DOCUMENT SUMMARIZATION

[1,2]ALBARAA ABUOBIEDA M. ALI, [1]NAOMIE SALIM, [3]RIHAB ELTAYEB AHMED, [1]MOHAMMED SALEM BINWAHLAN, [1,4]LADDA SUANMALI, [1,2]AHMED HAMZA.

[1] Faculty of Computer Science and Information Systems,

University Technology Malaysia, 81310, Johor, Malaysia

[2] Faculty of Computer Studies,

International University of Africa, 2469, Khartoum, Sudan

[3] College of Computer Science and Information Technology,

Sudan University of Sciences & Technology,407,Khartoum, Sudan.

[4]Faculty of Science and Technology,

Suan Dusit Rajabhat University, Bangkok, Thailand 10300

E-mail: albarraa@hotmail.com, naomie@utm.my, rahbon@hotmail.com, moham2007med@yahoo.com, nongnu_@hotmail.com, ahmedagraa@hotmail.com

ABSTRACT

Text features, as a scoring mechanism, are used to identify the key ideas in a given document to be represented in the text summary. Considering all features within same the level of importance may lead to generate a summary with low quality. In this paper, we present a feature selection method using (pseudo) Genetic probabilistic-based Summarization (PGPSum) model for extractive single document summarization. The proposed method, working as features selection mechanism, is used to extract the weights of features from texts. Then, the weights will be used to tune features' scores in order to optimize the summarization process. In this way, important sentences will be selected for representing the document summary. The results show that, our PGPSum model outperformed Ms-Word and Copernic summarizers benchmarks by obtaining a similarity ratio closest to human benchmark summary.

Keywords: *Summarization, Text Features, Genetic, Probabilistic, Similarity, Sentence Score, Features Weights, Binary Selection*

1. Introduction

Text summarization is defined as an operation of summarizing texts into a condensed form. A summary [1] is a new shorter text generated from one or more text sources. The main property of the new text is the inclusion of salient parts of the original text. [2] introduced first automatic text summarization (ATS) method which aims to let the machine have the ability to summarize texts.

ATS methods generate a summary from a large amount of text in different types, input sizes, and methodologies. The types of summaries are either indicative or informative. Indicative-based summary refers to a summary points to some important parts of the original text (article or document), while an informative-based summary aims to cover all relevant information that is provided in the text. ATS can employ two types of input sources: single document and multi document summarization. The developers of ATS can implement either extractive-base, abstractive-base methods summarizer or both of them. Extractive-base methods focus in selecting sentences that represent the text topic or concept, whereas abstractive-base methods generate a summary with phrases that may not be found in the original text.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In extractive summary the features of sentences are important to generate a good summary. A few features have been introduced in area of text summarization.

Recently, many researches handled the issue of features selection (FS) process. Due to its importance, FS affects the quality of the systems performance [3]. FS aims in identifying which features are important and can represent the data. [4] demonstrated that, the systems employing FS will improved its performance in several ways. FS reduces the dimensionality, removes irrelevant data, and skips the use of redundant features. In machine learning approach, FS can reduce the amount of data which are needed. Consequently, it improves the quality of system results. In automatic text summarization, the work of FS is not new. In section 2 we reviewed a number of works that discussed FS in the area of automatic text summarization and other areas.

In this paper, we introduce an extractive (pseudo) Genetic Probabilistic-base Single Document Summarization (PGPSum) model. The model presents a new FS method. Its mechanism is a mixture between the concept of genetic algorithm [5] and a simple probabilistic theory.

The rest of this paper is organized as follow. Section 2 introduces some literature review on text summarization. The features are described in section 3, while our proposed methodology is presented in section 4. Section 5 shows the experimental results and discussion. Section 6 concludes the paper findings.

2. RELATED WORKS

Luhn [2] first proposed a method that highlights important sentences to build abstract of scientific papers using IBM's data processing systems. In order to determine which sentences need to be included in the abstract, the "significant" sentences are identified. Two measurements have been proposed: word occurrences and sentence relative position. In addition, preprocessing steps are also applied which include: stop words removal and words stemming. The system then specifies sentences with high scores to be included in the abstract.

Later Baxandale [6] proposed a sentence selection measurement by its location in the text. Each sentence located at the beginning or at the end of the paragraph is considered to be important candidate and is included in the summary. For evaluation, Baxendale tested his methodology on 200 paragraphs: 85% of the paragraphs hold the sentence topic, while 7% ends with a topic sentence.

Ten years later, [7] presented two features in addition to two features presented by [2] and [6]. Edmondson used those two features to score sentences, and added two other features, which are pragmatic words: cue words, title and heading words.

Several approaches have been proposed for features selection in text processing, particularly in text summarization, using optimization techniques. Tu et al [8] used the PSO in order to select the optimal subset of features. These features are used as inputs for classification and training a neural network. [9] extracted the text features of the web using PSO to select important features. In area of text summarization, [10] introduced a work for feature selection closest to our work proposed here. He exploited five features regarding to text summarization and the PSO is used to train the system to obtain the weights of each features rather than doing selection as other works did. The used features are different in their structures, which are simple and complex. The results show that the complex features obtained high importance compare to simple ones. These weights have been employed in his next work [11] to generate the best summary. The system is compared against MS-Word summarizer and human summary as standard benchmarks. The results shown that, the proposed PSO method generate summaries which are 43% similar to the manually generated summaries, while MS-Word summaries are 37% similar.

Our work is closest to [10] but is different in following manner. [10] used the PSO as an optimization technique, whereas we used a partial concept of (pseudo) genetic algorithm. The number of population is 500 for each document in [10], whereas we employ a small size population which is only 32 per one document or iteration which is generated probabilistically. The PSO is used to train the system in order to differentiate between the features in terms of structures and importance, while we trained our system using a selected group of available proposed features having the same structure.

3. THE SELECTED FEATURES

A few of text summarization's features have been proposed in order to extract salient sentences in the text. For our empirical part, we

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

selected five simple statistical features [12, 13]: Title-Feature (TF) [7], Sentence-Length (SL) [14], Sentence-Position (SP) [6, 7], Numerical-Data (ND) [12], and Thematic-Word (TW) [2, 7, 15].

Title Feature (TF): To generate a summary for news article, sentences include a title words are considered important. Title feature is a percentage of how much the word of the sentence i match words of titles. Title feature can be calculated as Eq. 1.

$$TF(S_i) = \frac{\#\text{ of (Si)'s words matched title words}}{\#\text{ of title words}} \quad (1)$$

Sentence Length (SL): Selecting a short sentence may not represent the topic of the article due to fewer concepts held. As same, selecting a very long sentence may not also be considered as optimal selection. In order to avoid selecting sentences either too short or too long, a division by longest sentence solves this problem.

$$SL(S_i) = \frac{\#\text{ of words in } S_i}{\#\text{ of words in longest sentence}} \quad (2)$$

Sentence Position (SP): The first sentence in the paragraph is considered an important sentence and highly candidate to be included in the summary. The following algorithm is used to compute the SP feature.

$$t = \text{total numberof sentences in document (i)}$$
$$\text{for } i = 0..t \text{ do}$$
$$SP(S_i) = \frac{t - i}{t} \quad (3)$$

Numerical Data: A sentence that contains a numerical data refers to some important information such as date of event, money transaction, damage percentage, etc. The following formula shows how to compute this feature.

$$ND(S_i) = \frac{\#\text{ of numerical data in } S_i}{\text{Sentence Length}} \quad (4)$$

Thematic Words: Thematic words are a list of top n selected terms with the highest frequencies. To compute the thematic words, firstly we count frequencies of all terms in the document. Then specify a threshold in order to sign which terms shall be selected as thematic words. For our case, we select a top ten terms frequency.

$$TW - S_i = \frac{\#\text{ of thematic words in } S_i}{\text{Max number of TW found in a sentence}} \quad (5)$$

4. THE METHODOLOGY

In this section, we will illustrate our proposed methodology starting with the genetic base (Chromosome Encoding), evaluation function, and the proposed PGPSum model followed by training and testing stages.

Chromosome Encoding: Regarding to our target numbers of features that we used, the chromosome is composed of 5 genes. Each gene refers to a feature represented in binary format level. If the gene position (bit) holds a value of 1, it means that the corresponding feature is active and counted in the final score, otherwise, if the bit contains zero, it means that the corresponding feature is inactive and shall not be considered in final score. Figure 1 shows chromosome structure representing features' positions. The first bit refers to first feature; the second bit refers to second feature and so on.

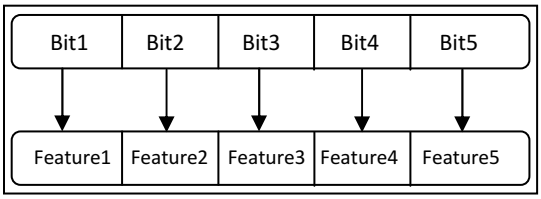

Fig. 1: shows chromosome structure for features

Evaluation Function: The evaluation or fitness function is a unit measure in optimization techniques. These techniques use this function in order to determine which chromosome obtains the best solution among a large number of chromosomes generated and the chromosome is then used in the next generation of the new population. In our case, we will generate only a probability of 2^n chromosomes for each input document. We let our system assign

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

for each chromosome a recall value of its generated summaries as a fitness value. The top chromosome with the highest recall value will be selected and represents the corresponded document in the dataset. W used ROUGE-1 Eq. (6) as a fitness function [16].

$$\frac{\sum_{S \in \{\text{ReferenceSummaries}\}[\sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}}{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)} \quad (6)$$

Where n is the length of the n-gram and $\text{Count}_{\text{match}}$ is the most possible number of n-grams shared between a system generated summary and a set reference summaries.

Pseudo Genetic Probabilistic-based Summarization (PGPSum) Model: In our proposed model, each sentence's features scores are tuned using the weights resulting in the training stage of this PGPSum model, as in (7). To train our PGPSum model, 100 document were selected from Document Understanding Conference [17] dataset 2002. This collection is used for both: training and testing in our PGPSum model. The score of each sentence is as follows.

$$\text{Scr}(s) = \sum_{i=1}^{5} \text{ScrF}_i(s) \times \text{wf}_i \quad (7)$$

Where $\text{Scr}(s)$ refers to the score of sentence s, $i = 1 \text{ to } 5$, $\text{ScrF}_i(s)$ refers to score of feature i, and wf_i refers to the weight of feature i generated by our PGPSum model.

The Training Procedure: We divided our training and testing procedures into two phases. Phase 1 focuses in working with issue of feature selection process, and the outputs of this phase are taken as input for next phase. Phase 2 deals with a text summarization problem. The purpose of including phase 2 in our experiments is to validate the selected features by our PGPSum model.

In both phases, we used the Document Understanding Conference dataset [17] 2002. Each document has two 100-word human experts' summaries. We have selected 10 random clusters (topic): D075b, D077b, D082a, D087d, D089d, D090d, D092c, D095c, and D096c. Each cluster has 10 related documents inside comprising 100 documents. Firstly, we preprocessed all documents: sentence segmentation, stop words removal, and stemming using porter stemmer [18]. Secondly, the text features are computed and the score of each sentence is represented as vector.

Phase 1: For each document, we made the chromosomes representation for feature selection. The number of genes for each chromosome is representing the number of the features. In this way, we obtained 2^5probable chromosomes/summaries. For each chromosome, we generate a summary based on its active and inactive features, see Eq (8).

$$\text{Scr}(s_i) = \sum_{i=1}^{5} \text{ScrF}_j(s) \times \text{VoGbP}(i) \quad (8)$$

Where $\text{Scr}(s_i)$ refers to the score of sentence s_i, $\text{ScrF}_i(s)$ is the score of the j^{th} feature, and $\text{VoGbP}(i)$ refers to the value of the gene bit position (either 0 or 1).

This is done by computing the features scores for each sentence. For each chromosome, among total of 32 generated chromosomes, our system will assign a recall value used to identify which chromosome should be selected into final computation using ROUGE-1. Once the system finishes selecting the fittest chromosome, this process will be repeated for all documents in the selected dataset.

Afterward, our system will compute the average of all 100 selected chromosomes in bit cases. The averages obtained represent the weights of each five features. The objective (outputs) of this phase is to obtain features weights.

Phase 2: for each document, we created 32 features' scores vectors. The scores of each sentence features are presented as a vector. Then each vector is passed as an input for the PGPSum model scoring function as shown in (7).

For each document, and base of (7), we created 32 features' scores vectors. For each vector we employed the features weights obtained from phase 1 as in (7). The system will then rank the scored sentences in a descending order. Then the top n sentences are selected as

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

summary, where n is refers to the total length (compression rate) of the summary. In this experiment, we used 20% as a compression rate for our generated summaries. Lastly, we used ROUGE [16] for assessing those generated summaries and to assign for each summary (chromosome) a fitness function value using (1).

The Testing Procedure: For each document, we created 32 features scores vectors. For each vector we employed the features weights obtained from phase 1 as in (7). The system will then rank the scored sentences in a descending order. Then the top n sentences are selected as summary, where n is refers to the total length (compression rate) of the summary. In this experiment, we used 20% as a compression rate and ROUGE [16] for assessing those generated summaries.

5. RESULTS AND DISCUSSION

The main purpose of this experiment is to study the behavior of a selected group of available features proposed and used in summarization methods. Therefore, we obtained two kinds of results, one for feature selection model, and the other for the summarization model. Using our simple proposed PGPSum model, it can identify which features are important than others. Figure 2 shows the calculated weights of features used in this study.

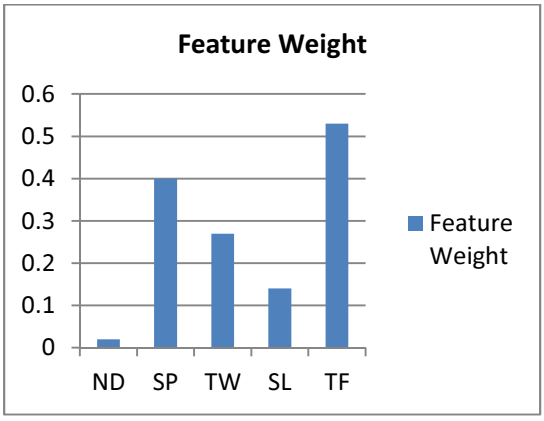

Fig. 2: Features Weights

From the result we observe that, TF (Title-Feature), SP (Sentence Position), and TW (Thematic Word) got the weights higher than the other two features which are SL (Sentence Length), and ND (Numerical Data) features.

TF selects sentences that contain the title words, which are considered, related to the text topic. In text analysis, the sentences located at unique positions may have some importance. The sentences located at the early positions in the paragraph contain the topical issue, while the coming sentences discuss the issue [6, 19]. Regarding to the way we compute the TW feature, this feature identifies a group of sentences that are considered important. SL feature shows that the lengths of the sentences have importance in text summarization. ND feature obtained lowest weight among the other features because most sentences do not contain numerical values.

In order to whether the features selection has effect on summarization, we employed the obtained weights on text summarization problem. The DUC 2002 dataset models two 100-word summaries for each article generated by human-experts. To evaluate our model we setup one human summary as a reference summary, while the other one has settled as a benchmark. In order to assess the quality of our proposed model, we count the similarity measure of PGPSum model, (Ms-Word Summarizer and Copernic Summarizer) [20], and benchmark against the reference summary.

Table 1 shows the final comparison between the proposed PGPSum model, Ms-Word summarizer model, Copernic Summarizer and the benchmark summary against the reference summary. The evaluation is based on the average recall generated by ROUGE-1,-2, and –L. The reason behind selecting these measures is due to their suitability in evaluating single document summarization [16]. Figures 3, 4, and 5 visualize the same results obtained in Table 1.

Table 1: The PGPSum, Ms-Word Summarizer and Benchmark Comparison. Avg Recall Using ROUGE-(1,2, and L).

ROUGE	Model Used	Avg-Recall	95%-Confidence Interval
1	Human	0.51642	0.49620 - 0.53910
	PGPSum	**0.456**	**0.42872**

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

	(Proposed)	**34**	**- 0.48273**
	Ms-Word	0.39653	0.37504 - 0.41751
	Copernic	0.18045	0.17041 - 0.19003
2	Human	0.23394	0.21280 - 0.25791
	PGPSum(Proposed)	**0.24506**	**0.22059 - 0.27010**
	Ms-Word	0.17441	0.15547 - 0.19509
	Copernic	0.01399	0.01137 - 0.01685
L	Human	0.48389	0.46495 - 0.50601
	PGPSum (Proposed)	**0.42186**	**0.39607 - 0.44728**
	Ms-Word	0.36368	0.34262 - 0.38337
	Copernic	0.16814	0.15912 - 0.17732

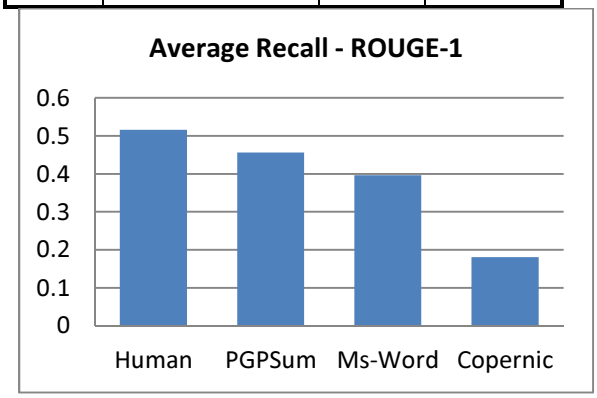

Fig 3: The Benchmark Model, PGPSum model, Ms-Word Summarizer, and Copernic Comparison: Average Recall using ROUGE-1.

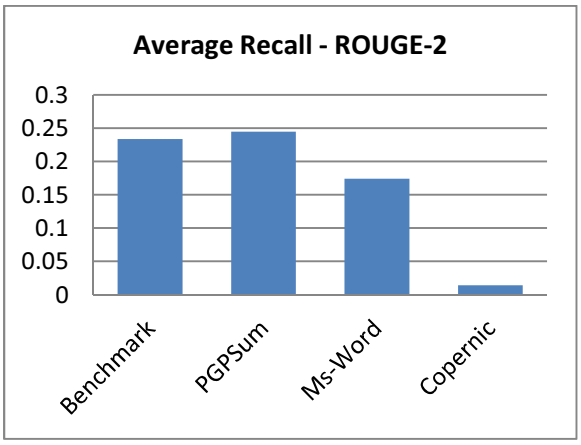

Fig 4: The Benchmark Model, PGPSum model, Ms-Word Summarizer, and Copernic Comparison: Average Recall using ROUGE-2.

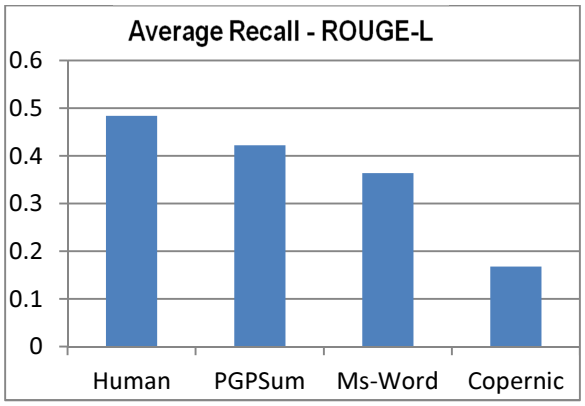

Fig 5: The Benchmark Model, PGPSum model, Ms-Word Summarizer, and Copernic Comparison: Average Recall using ROUGE-L.

From the results above, our proposed model outperforms Ms-Word Summarizer and Copernic Summarizer models as follow. The PGPSum summarizer obtained summaries that are 45% similar to the human summary (Benchmark), whereas Ms-Word and Copernic Summarizers obtained summaries that are only 39% and 18% similarity with the benchmark respectively. The human to human summary is 51% similar to each other.

6. CONCLUSION

In this paper, we presented a simple hybrid approach for feature selection using a concept of (pseudo) genetic based and probabilistic theory extractive-base single document summarization. The features are represented and encoded using the structure of binary genes, while their

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

appearance is governed using probability. Our experiment used DUC2002 data set which consists of 100 document collection. Among our selected features and used dataset, we found that TF, SP, and TW are very important than the other features. The proposed model outperforms the Ms-Word and Copernic summarizers. It scored a ratio of 45% similarity, while Ms-Word and Copernic Summarizers scored ratios of 39% and 18% similarities, respectively, when compared against human summary (reference). The similarity ratio of the Benchmark against the reference summaries is 51%, which is not so far from our PGPSum model. For our future work, we plan to use more features and try other optimization algorithms.

REFERENCES

[1] E. Hovy and C.-Y. Lin, "Automated text summarization and the SUMMARIST system," presented at the Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998, Baltimore, Maryland, 1998.

[2] H. P. Luhn, "The automatic creation of literature abstracts," *IBM J. Res. Dev.*, vol. 2, pp. 159-165, 1958.

[3] H. Xingshi, *et al.*, "Feature Selection with Discrete Binary Differential Evolution," in *Artificial Intelligence and Computational Intelligence, 2009. AICI '09. International Conference on*, 2009, pp. 327-330.

[4] R. N. Khushaba, *et al.*, "Differential evolution based feature subset selection," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, 2008, pp. 1-4.

[5] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*: Addison-Wesley, 1989.

[6] P. B. Baxendale, "Machine-made index for technical literature: an experiment," *IBM J. Res. Dev.*, vol. 2, pp. 354-361, 1958.

[7] H. P. Edmundson, "New Methods in Automatic Extracting," *J. ACM*, vol. 16, pp. 264-285, 1969.

[8] C.-j. Tu, *et al.*, "Feature Selection using PSO-SVM," *IAENG International Journal of Computer Science*, vol. 33, pp. 138-143, 2006.

[9] Y. Liu, *et al.*, "Feature Selection with Particle Swarms," in *Computational and Information Science*. vol. 3314, J. Zhang, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2005, pp. 425-430.

[10] M. S. Binwahlan, Salim, N., & Suanmali, L., "Swarm based features selection for text summarization," *International Journal of Computer Science and Network Security IJCSNS*, vol. 9, pp. 175-179, 2009b.

[11] M. S. Binwahlan, *et al.*, "Swarm Based Text Summarization," in *Computer Science and Information Technology - Spring Conference, 2009. IACSITSC '09. International Association of*, 2009, pp. 145-150.

[12] M. A. Fattah and F. Ren, "GA, MR, FFNN, PNN and GMM based models for automatic text summarization," *Computer Speech and Language*, vol. 23, pp. 126-144, Jan 2009.

[13] M. Litvak, *et al.*, "A new approach to improving multilingual summarization using a genetic algorithm," presented at the Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 2010.

[14] S. S. C. N. Satoshi, M. Murata, K. Uchimoto, M. Utiyama, and H. Isahara, "Sentence Extraction System Assembling Multiple Evidence," in *2nd NTCIR Workshop*, 2001, pp. 319--324.

[15] J. L. Neto, *et al.*, "Generating text summaries through the relative importance of topics," *Advances in Artificial Intelligence*, vol. 1952, pp. 300-309, 2000.

[16] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of summaries," presented at the Proc. ACL workshop on Text Summarization Branches Out, 2004.

[17] DUC, "The Document Understanding Conference (DUC).".

[18] M. F. Porter, "An algorithm for suffix stripping," in *Readings in information retrieval*, ed: Morgan Kaufmann Publishers Inc., 1997, pp. 313-316.

[19] J. Kupiec, *et al.*, "A trainable document summarizer," presented at the Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, Washington, United States, 1995.

[20] A. Kiani and M. R. Akbarzadeh, "Automatic Text Summarization Using Hybrid Fuzzy GA-GP," in *Fuzzy Systems, 2006 IEEE International Conference on*, 2006, pp. 977-983.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AUTHOR PROFILES:

Albaraa Abuobieda Mohammed Ali is currently pursuing his PhD from Univeristi Teknologi Malaysia in the area of Text Summarization. He received his B.Sc in Computer Science from the International University of Africa, Sudan, in 2004. He earned M.Sc in Computer Science from Sudan University of Science and Technology in 2008. His current areas of research include text summarization, plagiarism detection, Ontology, network and network security.

Prof. Dr. Naomie Salim is a professor. She received her bachelor degree in Computer Science from Universiti Teknologi Malaysia in 1989. She received her master degree in Computer Science from University of West Michigan in 1992. In 2002, she received her Ph.D(Computational Informatics)from University of Sheffield, United Kingdom. Her current research interest includes Information Retrieval, Distributed Database and Chemoinformatic.

Mohammed Salem Binwahlan graduated in Computer Science from Hadhramout University of Science & Technology (HUST), Mukalla, Yemen (1996-2000). He worked as a Teaching Assistant at HUST (2000-2004). He obtained the MSc degree in Computer Science from University Technology Malaysia (Universiti Teknologi Malaysia-UTM), Johor, Malaysia (2004-2006). He worked as a lecturer in the Department of Computer Science in Faculty of Applied Sciences at HUST (2006-2007). He obtained the Ph.D degree in Computer Science (Automatic Text Summarization) from UTM (2008-2011).

Ladda Suanmali is a Ph.D. candidate in computer science in the Faculty of Computer Science and Information Systems at Universiti Teknologi Malaysia. She graduated a bachelor degree in computer science from Suan Dusit Rajabhat University, Thailand in 1998. She graduated a master degree in information technology from King Mongkut's University of Technology Thonburi, Thailand in 2003. Since 2003, she has been working as a lecturer in the Faculty of Science and Technology, Suan Dusit Rajabhat University. Her current research interests include text summarization, data mining, and soft computing.

Ahmed Hamza Osman is a PhD student in Universiti Teknologi Malaysia. He has a bachelor degree in Computer Science from IUA in 2004. He received a master degree in Computer Science from Sudan University of science and technology (SUST) in 2008. His current research interest includes Plagiarism Detection, Information Retrieval, Text Processing and Data mining.