Multi-Document Text Summarization using Fuzzy Logic and Association Rule Mining

Assist. Prof. Dr. Suhad Malallah
suhad_malalla@yahoo.com
University of Technology – Department of Computer Science
Zuhair Hussein Ali
zuhair72h@yahoo.com
Al- Mustansiriya University – College of Education
Department of Computer Science

Abstract: *Information is very necessary. The huge quantity of information on the Internet makes text summarization research increase rapidly. Text summarization is the process of choosing significant sentences from one or multi-document without losing the main ideas of the original text. In this paper a new multi-document English text summarization was proposed, which is based on linguistic and statistical features of the sentences. The extracted features fed to the fuzzy logic system, then the Apriori algorithm used for association rule extraction. The proposed model is performed using dataset supplied by the Text Analysis Conference (TAC-2011) for English documents. The results were measured by using Recall-Oriented Understudy for Gisting Evaluation(ROUGE). The obtained results support the effectiveness of the proposed model.*

Keywords: Fuzzy Logic, Support, Frequent Itemset, Apriori, Confidence

1 . Introduction

Information is very necessary, the greater part of it exists on the web. The Internet includes a huge quantity of documents and is developing at a high rate. Instruments that give convenient access to different sources are vital so as to alleviate the data overload people are confronting these worries have led to the care about the growth of automatic text summarization system [1]. The goal of text summarization is to create abstract from a multi document or single document pertinent to the user's question. It must explain a complete content of the document in the least amount of words without losing the main ideas of the original text [2]. The main purpose of text summarization is to help users in discovering information from source documents by gathering the indispensable information and giving its shortened form. In such manner, text summarization can be considered as an arbiter between information included in many documents and users [3].

Text summarization methods are categorized as a abstractive summarization and extractive summarization. Abstractive summarization depends on Natural Language Processing (NLP) strategies for parsing, finding and creating content. Currently, NLP machinery is computationally cost-effective but it has less precision. Conversely, extractive summarization can be defined as the technique for verbatim extraction of the literary components like passages, sentences and so on from the source content. Abstractive summarization is noticed to be complex and consumes more time as compared to extractive summarization [4]. The fundamental objective of document summarization depend on extraction technique is the picking of suitable and pertinent sentence from the input documents. A technique to acquire the suitable sentences is assigned a weight for each sentence which indicates the salience of a sentence for choosing to the summary and then selecting the top ones [5].

Depending on the quantity of documents to be summarized, the summary can be classified to a single document summarization(SDS) and multi-document summarization

(MDS). Just one document can be condensed into a shorter document in a SDS, whereas in MDS a set of documents is condensed into a summary. Without automatic text summarization it's difficult to summarize a great number of documents and another issue confronted is repetition in the information offered by the multiple documents [6]. In this paper, we have proposed an automatic MDS, which use fuzzy logic and Apriori algorithm. Here, we have used seven different features to specify the importance of sentences. We have utilized TAC-2011 dataset to assess the summarized results.

Whatever remains of paper sorted as, The related works is given in section 2, the proposed method in section 3, the dataset and evaluation metrics in section 4, the experimental results in section 5 and conclusion in section 6.

2. Related Works

Automatic text summarization decreases a big corpus of documents or text document in a smaller set of sentences which explain the main ideas of the text. By huge quantity of text and the enormous increase of information on the internet, specialists in NLP are more interested to discover new methods for summarizing and explore a variety of models to come up with perfect summarization. As we mentioned summarization can be categorized into two techniques, text abstraction and text extraction [7]. In this section we investigate some of these methods.

Ramiz M. Aliguliyev at 2010 proposed a discrete particle swarm optimization algorithm which includes a mutation operation obtained from genetic algorithms to fix the clustering trouble in multi-document summarization. The author suggested two weighted clustering techniques, their three collections and a new dissimilarity measurement to optimize different aspects of inter-cluster dissimilarity, intra-cluster similarity and their collections [8].

Binwahlane in 2010 suggested fuzzy-swarm hybrid diversity A method that merges three models depend on swarm, diversity and fuzzy-swarm. The diversity-based model forms, sentence groups arranged in a binary tree according to their scores. It then executes Maximal Marginal Importance (MMI) to choose the sentences for embedding in the summary. The model based on PSO binary is applied to optimize the weight related to every feature of the objective function. The location of the particle is a string of bits, where one means that the related feature is chosen, otherwise it has a zero. On obtaining the weights, the score is given for every sentence and the sentences that have higher score are chosen to be incorporated in the summary. In the model depend on fuzzy logic and swarm, the sentence score is calculated by a system of inference in the fuzzy algorithm, starting with the weights creates with PSO. The sentences are sorted depending on the score and the summary is acquired [9].

Ramiz M. Aliguliyev at 2011 applied a mathematical method which consists of two phases and modeled as a discrete optimization problem. The first phase introduces topics discovery through clustering of document collection sentences by applying k-means algorithm. In the second phase, the related sentences from every cluster are extracted and repetition is avoided using sentence and cluster and sentence-to-sentence relations. The differential evolution algorithm used to resolve the trouble of discrete optimization problem [10].

Abuobieda et in 2012 utilized sentence position, sentence length, numerical information, thematic words and title feature as features for scoring sentences. They applied genetic algorithms to the last model of the feature space [11]

3. The Proposed Method

The aim of automatic multi-document summarization is to extract the summary related to multiple documents that are written about the similar events. The mechanical procedures for the creation of the single document summary have been started in 1950's but still it has been getting significant interest among

the researchers. Because it's very, rapidly increasing on the web content, there is a strong require to summarize a great set of documents in a few period of time. In this research, we have developed a new method depend on Fuzzy logic and the rules generated by Apriori algorithms. Figure (1) shows the proposed automatic summarization model. There are four main steps in the proposed system for the MDS

1. Preposessing
2. Feature Extraction
3. Fuzzy logic, scoring
4. Generation of rules using A priori algorithms

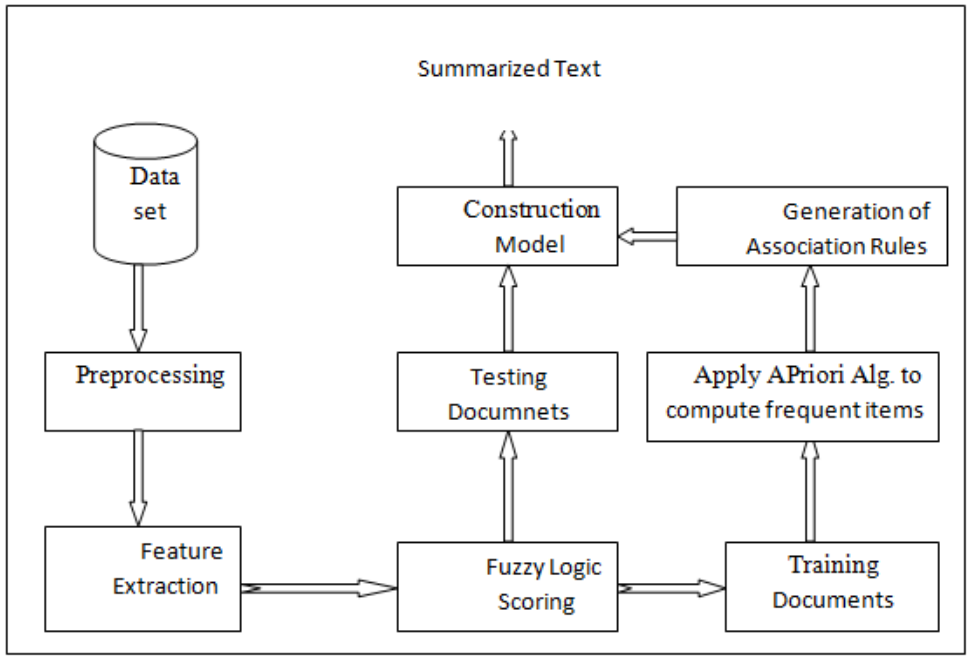

Figure (1) block diagram for the Proposed Automatic Text Summarization Model

3.1 Preprocessing

There are four steps for preparing the data, these steps are:

A. Sentence segmentation: Sentence segmentation is an important approach in text processing such as machine translation, information extraction, plagiarism detection, text summarization

and syntactic parsing. It can be done by separating sentences according to the dot "." between sentences.

B. Tokenization: Is the process of splitting sentence into word.

C. Stop word removal: is the third step in preprocessing steps, where words which don't give the necessary information for identifying significant meaning of the document content and appear frequently are removed. There are a variety of methods used for specifying of such stop words list. Presently, a number of English stop word list is usually used to help text summarization process. Regardless of its repetition and having no effect to the meaning, these words contribute an important percentage of the overall documents. Removing of such words can increase the efficiency and effectiveness of information retrieval process. The document size can be minimized without affecting its meaning, less memory and time consumed.

D. Word Stemming : Is the process of producing base or root of the word in This paper, performed words stemming by removing suffixes proposed by Porter's stemming algorithm [12].

3.2 Features Extraction

It's an important part the text summary, which includes computing of features score for every sentence. These features include sentence position, sentence length, numerical data, Thematic word, title word, proper noun, and centroid value. All these features from [13]

1. Sentence positions: Where the higher score will give to the first sentence in the document, and the score decreases according to the sentence position in the document. This feature can be computed according to equation (1).

$$F1 = \frac{N - P + 1}{N} \quad (1)$$

Where:

N: total number of sentences in the document
P: current position of the sentence

2. Sentence length: This feature is helpful for deleting the short sentences, the short sentences are the new article which includes writer name, datelines which is not necessary to be included in the summary. This feature is computed as the ratio between sentence length and the longest sentence in the document as in equation (2).

$$F2 = \frac{sentence\ length}{longest\ sentence\ length\ in\ the\ document} \quad (2)$$

3. Numerical data: This feature has important information and it would more probably incorporate into the summary [14]. This feature is calculated according to the following equation (3).

$$F3 = \frac{Number\ of\ numerical\ data\ in\ the\ sentence}{Total\ sentence\ length} \quad (3)$$

4. Thematic Words: is the term that appears most frequently in the document. This feature can be calculated by computing the repetitions of all terms in the document, then top (n) terms with the highest frequency is selected, in this research, we used top (5), a ratio is given between the number of thematic words in the sentence and the maximum thematic words in the document as explained in equation (4).

$$F4 = \frac{Number\ \ of\ thematic\ words\ in\ the\ sentence}{Max\ umber\ of\ Thematic} \quad (4)$$

5. Title word: This feature is important when summarizing the document. The score is given as the ratio between number of title words in the sentence and the number of words in the title as follows (5).

$$F5 = \frac{Number\ \ of\ title\ word\ in\ the\ sentence}{Total\ number\ of\ word\ in\ the\ title} \quad (5)$$

6. Proper noun: The sentence is important if it includes the maximum number of proper nouns [15]. As in equation (6)

$$F6 = \frac{Number\ of\ proper\ nouns\ in\ the\ sentence}{Sentence\ length} \quad (6)$$

7. Centroid value: Is a feature used to specify salient sentences in the multiple documents. This feature can be calculated as follows

$$F7 = \sum_{i=1}^{n} C_{wi} \quad (7)$$

$$C_{wi} = TF * IDF \quad (8)$$

$$IDF = log[\frac{Total\ NO.\ of\ documents}{NO.\ of\ documents\ containing\ the\ given\ word}\ ] \quad (9)$$

Where:

C_w is the centroid of the words

TF is the term frequency which represents the frequency of a given term in the document.

IDF is the inverse term frequency computed by division of the total number of documents in the dataset and the number of documents including the given term [13].

3.3 Fuzzy Logic Scoring

The obtained features from the previous section are used as inputs to the fuzzy logic. The fuzzy logic system uses the Bell membership function to partition the score of each feature into one of three values that are high, medium and low [16]. We use the Bell membership function which is identified by three parameters (a,b,c)as in the following equation (10)

$$Bell(x,a,b,c)=1/(1+|(x-c)/a|^{2b}) \quad (10)$$

Where x is the input feature to the equation to be fuzzified, the parameters a and c are used to specify the width and the center of the membership function. Then b is used to control the crossover points slopes [17].

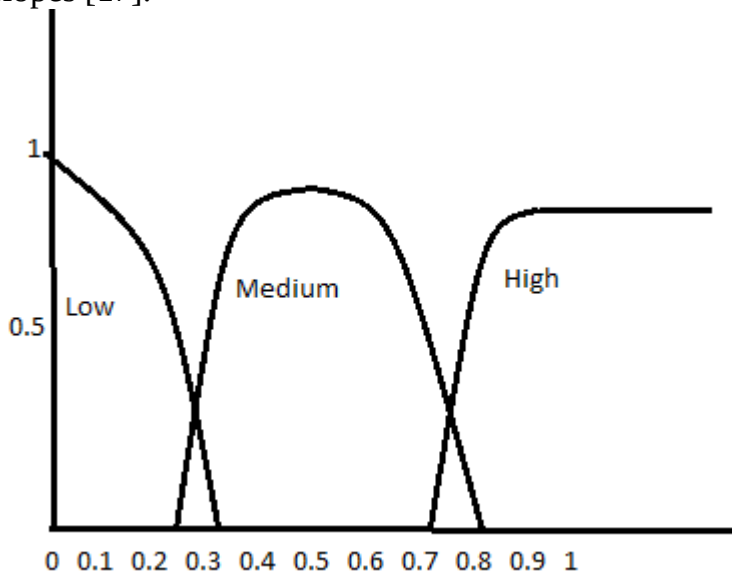

Figure (2) Bell membership function

Figure (2) shows how the Bell membership function assigns values for input variable sentence length. Where an input variable (X) assigns one of the values(low,medium,high) according to equation (10).

3.4 Generation of Association Rule Using Apriori Algorithm

The goal of association rule is to explore the relationship between a group of items in a database. In 1993 Association rule was first introduced [18]. There are two parts of an association rule, an antecedent (if) is an item appear in the data and a consequent (then) an item that appear in combination with the antecedent. Generation of association rules can be done by analyzing the data for frequent if/then patterns and applying the criteria support and confidence to specify the most significant relationships. Support refers to how the item appears frequently in the database. Confidence refers to how many the if/then statements have been appearing to be true [19]. There are two modes in our proposed automatic MDS training mode and testing mode.

Algorithm1 Aprior Algorithm first phase

C_k :Candidate item set of size k

L_k:Frequent itemset of size k

L1={Frequent items}

For (k=1;k!=empty; k++) do bgin

C_{k+1}=Candidate generated from L_k

for each transaction t in database do

increment the count of all candidates in C_{k+1} that are contained in t

L_{k+1} = candidates in C_{k+1} with min_support

End

return $\cup_k L_k$;

1) **Training mode**: Where the extracted features from (70) manually summarized English documents fed to the system to obtain the constructed model. As illustrated in figure (1) Apriori Algorithm applied to the fuzzified features. Apriori algorithm is an important algorithm for mining frequent itemsets for Boolean association rules. It was first introduced in 1994 by Agrawal and R.Srikant [20] and is the first algorithm that control the exponential increase of candidate itemset. The support is used for pruning candidate itemset [21]. Apriori algorithm is composed of two main phases, the creating frequent itemsets and discovering the association rules from the created frequent itemsets. At the first step, it finds the frequent itemsets. The frequent itemset is an itemset whose support is larger than some specified minimum support. The following algorithms illustrate the first phase of Aprior algorithm**.**

At the second phase of Apriori algorithm the association rules are chosen by applying the minimum confidence support. The following algorithm illustrates the generation of association rules.

Algorithm2 : A priori Algoritm for generation of association rules

- For all frequent itemset *A*,
  For all proper nonempty subset *X* of *A*,
  - Let *Y* = A- *X*
  - $X \rightarrow Y$ is an association rule if
    - Confidence($X \rightarrow Y$) $\geq$ minconf,

  support($X \rightarrow Y$) = support($X \cup Y$) = support(A)
  confidence($X \rightarrow Y$) = support($X \cup Y$) / support(X)

Where minconf is the minimum confidence in our proposed automatic text MDS we use minconf equal to 75%.All rules with confidence is less than 75%will be ignored. The final results of this phase will be rules which have the following form:
If (sentence_position=high and sentence_length=medium and numerical_data=high and thematic_word=medium and title_word=low and proper_noun=high and

centroid_value=medium ⇒ Sentence important) That we discover about (100) rules such(that have the above form) that specify important sentence.

2) **Testing Mode:** in this mode only the documents that produced from the fuzzy logic is fed to the constructed rules that generated from the previous mode. There is a (30) documents in this mode.

4. Dataset and Evaluation metrics

The dataset used in our research is the Text Analysis Conference (TAC-2011) which consist of seven languages (English, Arabic, Greek, Czech, French, Hindi, Hebrew). There are10 topics, each of 10 documents for each language that there are 100 documents for each language. The dataset was started with English language, then translated to each other language. There are three models summaries for each document set created by fluent speakers, each summary size about 240-250 words. Also, there are numbers of peer summaries provided by the creator for evaluation [22].

ROUGE [23] was utilized to evaluate the proposed system. These measures depend on the comparison of n-grams between the several summaries and candidate summary. The comparison is done by calculating the number of overlapping units. There are many variants of ROUGE, in our proposed system ROUGE-N which is calculated according to eq. (11)

$$ROUGE-N = \frac{\sum_S(s \in reference\ summaries) \sum_{n-gram \in S} count_{match}(N-gram)}{\sum_S(s \in reference\ summaries) \sum_{n-gram \in S} count\ (N-gram)} \quad (11)$$

Where N represents the N-gram length. $Count_{match}$ (N-gram) is the largest number of overlapping between a candidate summary and the number of reference summaries. Count(N-gram) is the number of N- gram in the reference summaries. Here we use ROUGE-1 which compute unigram between the candidate summary and the reference summaries. The output of ROUGE package is three

numbers which represent, precision (P) is the set of sentences overlapped between the human summary and system summary divided by the set of sentences in the system summary. As in eq. (12).

$$P = \frac{S\epsilon human\ summary \cap S\epsilon system\ summary}{S\epsilon system\ summary} \quad (12)$$

The second output of ROUGE is Recall (R) which is the set of sentences overlapped between human summary and system summary divided by the human summary. As in eq. (13)

$$R = \frac{S\epsilon human\ summary \cap S\epsilon system\ summary}{S\epsilon human\ summary} \quad (13)$$

The Third one is F−score measure which is used to balance system performance between "precision" and "recall" measures as in eq. (14)

$$F = \frac{(1+\beta^2)R * P}{R + \beta^2 P} \quad (14)$$

Where $\beta = \frac{P}{R}$

5. Experimental Results

The TAC-20111 dataset was used in our proposed system for multi-document English text summarization. As illustrated in the previous section the dataset consists of 10 topics, each of the10 documents. Here we divided the dataset into two sets, the first set for training and a second set for testing. Each of them consists of (70) documents where (7) documents from each topic is selected as a train data and all the (100) documents as the test data. The results show a good performance of the proposed systems as appear in the table (1) which represents the output of ROUGE-1.

Table (1) The scores of ROUGE-1 produced by the proposed system

ID Number	Precision	Recall	F-Score
ID1	0.44264	0.43121	0.43670
ID2	0.51121	0.50132	0.50612
ID3	0.49112	0.48091	0.48585
ID4	0.48282	0.51212	0.49617
ID5	0.51123	0.50013	0.50549
ID6	0.51342	0.41012	0.44499
ID7	0.41252	0.41235	0.41243
ID8	0.41123	0.41202	0.41162
ID9	0.38812	0.40212	0.39474
ID10	0.57325	0.57336	0.57330

While table (2) shows the output of Rouge-1 as appear in the TAC-2011 dataset.

Table (2) the scores of ROUGE-1 for the system summary

ID Number	Precision	Recall	F-Score
ID1	0.41253	0.40524	0.40776
ID2	0.45655	0.46481	0.46062
ID3	0.47909	0.43169	0.45404
ID4	0.44966	0.44423	0.44691
ID5	0.43513	0.41092	0.42243
ID6	0.45122	0.35471	0.39617
ID7	0.3953	0.39586	0.39547
ID8	0.39265	0.38714	0.38985
ID9	0.37726	0.38105	0.3791
ID10	0.51806	0.52488	0.52141

We can see from table1 that our results for each of Precision, Recall and F-Score is better than the table2 which represent the result of peer summary for the TAC-2011 dataset. These results show the effect of the selected features and the effect of the rule representation method to get good results.

6 .Conclusion

The extraction of information from a multi-document is very necessary. MDS is the choosing of important sentences from the original text with keeping the main ideas for that summarized documents. In this paper a new method for MDS is introduced which depends on linguistic and statistical features of the sentences. Apriori algorithm applied for extraction of association rules. Two important criteria play main role in getting good results, confidence and the number of training data. Where many confidences applied until reaching 75%, which gives balance between the number of generated rules and the importance of these rules. The number of training dat set is also very important such that, less efficient results obtained when number of trained documents less than(60).

References

[1] R. Kumar and D. Chandrakal" A survey on text summarization using optimization algorithm," ELK Asia Pacific Journals vol. 2, no. 1, 2016.

[2] Y. K. Meena and D. Gopalani, "Evolutionary Algorithms for Extractive Automatic Text Summarization," *Procedia Comput. Sci.*, vol. 48, no. Iccc, pp. 244–249, 2015.

[3] K. Duraiswamy and G. Padma Priya, "An Approach for Text Summarization Using Deep Learning Algorithm," J. Comput. Sci., vol. 10, no. 1, pp. 1–9, 2014.

[4] R. He, J. Tang, P. Gong, Q. Hu, and B. Wang, "Multi-document summarization via group sparse learning," Inf. Sci. (Ny)., vol. 349–350, pp. 12–24, 2016.

[5] A. John and D. M. Wilscy, "Random Forest Classifier Based Multi-Document Summarization System," IEEE Recent Adv. Intell. Comput. Syst. RANDOM, pp. 31–36, 2013.

[6] S. A. Babar and P. D. Patil, "Improving Performance of Text Summarization," Procedia Comput. Sci., vol. 46, no. Icict 2014, pp. 354–363, 2015.

[7] B. . Samei, M. . Eshtiagh, F. . Keshtkar, and S. . Hashemi, "Multi-document summarization using graph-based iterative ranking algorithms and information theoretical distortion measures," Proc. 27th Int. Florida Artif. Intell. Res. Soc. Conf. FLAIRS 2014, pp. 214–218, 2014.

[8] R. M. Aliguliyev, "Clustering techniques and discrete particle swarm optimization algorithm for multi-document summarization," Comput. Intell., vol. 26, no. 4, pp. 420–448, 2010.

[9] M. S. Binwahlan, N. Salim, and L. Suanmali, "Fuzzy swarm diversity hybrid model for text summarization," Inf. Process. Manag., vol. 46, no. 5, pp. 571–588, 2010.

[10] R. Aliguliyev, "A new similarity measure and mathematical model for text summarization," no. January 2015.

[11] A. Abuobieda, N. Salim, A. T. Albaham, A. H. Osman, and Y. J. Kumar, "Text summarization features selection method using pseudo genetic-based model," Proc. - 2012 Int. Conf. Inf. Retr. Knowl. Manag. CAMP'12, pp. 193–197, 2012.

[12] Porter stemming algorithm: http://www.tartarus.org/martin/PorterStemmer

[13] ANSAMMA JOHN, "Multi-Document Summarization System: Using Fuzzy Logic and Genetic Algorithm," Int. J. Adv. Res. Eng. Technol., vol. 7, no. 1, pp. 30 – 40 , 2016.

[14] M. A. Fattah and F. Ren, "GA, MR, FFNN, PNN and GMM based models for automatic text summarization," Comput. Speech Lang., vol. 23, no. 1, pp. 126–144, 2009.

[15] C. N. Satoshi, S. Satoshi, M. Murata, K. Uchimoto, M. Utiyama, and H. Isahara, "Sentence Extraction System Assembling Multiple Evidence," Proc. 2nd NTCIR Work., pp. 319–324, 2001.

[16] M. Patil and N. Kulkarni, "Text Summarization Using Fuzzy Logic," Paragraph, vol. 1, no. 3, pp. 42–45, 2014.

[17] A. K. Kiani and M. R. Akbarzadeh-T, " Automatic Text Summarization Using Hybrid Fuzzy GA-GP," 2006 IEEE Int. Conf. Fuzzy Syst., pp. 977–983, 2006

[18] R. Agrawal, T. Imielinski, and A. Swami, " Mining association rules between sets of items in large databases," Proc. 1993 ACM SIGMOD Int. Conf. Manag. data, vol. 22, no. May, pp. 207–216, 1993.

[19] E. Duneja, "A Survey on Frequent Itemset Mining with Association," International Journal of Computer Applications vol. 46, no. 23, pp. 18–24, 2012.

[20] J. Han and M. Kamber, Data Mining: Concepts and Techniques, vol. 54, no. Second Edition. 2006.

[21] Y. Q. Wei, R. H. Yang, and P. Y. Liu, "An improved apriori algorithm for association rules of mining," ITME2009 - Proc. 2009 IEEE Int. Symp. IT Med. Educ., vol. 3, no. 1, pp. 942–946, 2009.

[22] G. Giannakopoulos, M. El-Haj, B. Favre, M. Litvak, J. Steinberger, and V. Varma, "TAC 2011 MultiLing Pilot Overview," no. November, 2011.

[23] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," Proc. Work. text Summ. branches out (WAS 2004), no. 1, pp. 25–26, 2004.

تلخيص النصوص المتعددة باستخدام المنطق الضبابي مع تعدين القوانين المشتركه

أ.م.د سهاد مال الله
suhad_malalla@yahoo.com
الجامعة التكنولوجية – قسم علوم الحاسوب
م.زهير حسين علي
zuhair72h@yahoo.com
الجامعة المستنصرية – كلية التربية – قسم علوم الحاسوب

المستخلص

بالنظر لكثرة المعلومات الموجودة في شبكة الاتصالات وأهميتها أزداد ألاهتمام بالبحوث التي تتناول تلخيص النصوص. عملية التلخيص أما أن تكون لنص واحد أو لمجموعة من النصوص بحيث تحافظ عملية التلخيص على الافكار الاساسية للنصوص الملخصة. في هذا البحث تم أقتراح طريقة تعتمد على أستخلاص الخواص اللغوية والاحصائية للجمل ومن ثم تقدم هذة لاعطاها تصنيفات بعدها يتم تقديمها الخصائص للمنطق الضبابي (Fuzzy logic) لخوارزمية Apriori لغرض أستخراج القوانين الخاصة بتصنيف الجمل التي تكون اما مهمة أو غير مهمه. كما تم حساب النتائج باستخدام (TAC-2011)تم تطبيق البرنامج على قاعده بيانات (ROUGE)

Recall-Oriented Understudy for Gisting Evaluation

الكلمات الرئيسية: المنطق الضبابي، التدعيم، مجموعة العناصر المتكررة، خوارزمية ابرايوري، الثقة.