Automatic Text Extraction Based on Field Association Terms and Power Links

Mahmoud B. Rokaya
College of Computers and Information Technology
Taif University
Taif, Saudi Arabia
mahmoudrokaya@tu.edu.sa

Sultan Aljahdali
College of Computer and Information Technology
Taif University
Taif, Saudi Arabia
aljahdali@tu.edu.sa

Dalia I. Hemdan
College of Designs and Home Economy
Taif University
Taif, Saudi Arabia
dalia.m@tu.edu.sa

ABSTRACT
The existence of the World Wide Web has caused an information explosion. Readers are overloaded with lengthy text documents where a shorter version would suffice. Text summarization is the process of distilling the most important information from a source to produce an abridged version for a particular user and task. When this is done by means of a computer, i. e. automatically, it is called as Automatic Text Summarization. Automatic text summarization is a technique where the text is input to the computer and it returns the clipped and concise extract of the original text and also sustains the overall meaning and main information content. In this work we introduce a custom summary based on key words or field of interest that a user determine or select. Despite there are as many as methods of summarization but there are very few methods of summarization considered a custom summary that depends on the user need. Even the methods that tried to rely on the user keywords, they followed these key words blindly without any analysis to the actual relations between these key words or other related field association terms. our work is a combination of key phrase extraction for a given corps. Then, we build the connections between these key words. These key words will be the base for the rest of the work. The user enter the initial keywords, the system pick the nearest terms in the map for these field association terms, and use all of them to extract the coherent passages related to them in the document. Then combining coherent passages to bring out the final summary.

1. Introduction.
The existence of the World Wide Web has caused an information explosion. Readers are overloaded with lengthy text documents where a shorter version would suffice. [1]
Text summarization is the process of distilling the most important information from a source to produce an abridged version for a particular user and task. [2]
When this is done by means of a computer, i. e. automatically, it is called as Automatic Text Summarization. Automatic text summarization is a technique where the text is input to the computer and it returns the clipped and concise extract of the original text and also sustains the overall meaning and main information content. [3]
Summarization can be classified into two approaches: extraction and abstraction. Extraction based summaries are produced by concatenating several sentences taken exactly as they appear in the texts being summarized. Abstraction based summaries are written to convey the main information in the input and may reuse phrases or clauses from it. [2]
The state-of-the-art abstractive methods are still quite weak, so most research has focused on extractive methods, and this is what we will cover.
Reza and Gallinari described a system for automatic text summarization that operates by extracting the most relevant sentences from documents with regard to a query.[4]
Automatic text extraction techniques have proved robust, but very often their summaries are not coherent. Constantin proposed a new extraction method which uses local coherence as a means to improve the overall quality of automatic summaries. [5]
Atefeh Farzindar and Guy Lapalme presented their work on the development of a new methodology for automatic summarization of justice decision. they described LetSum (Legal text Summarizer), a prototype system, which determines the thematic structure of a judgment in four themes Introduction, Context, Juridical Analysis and Conclusion. Then it identifies the relevant sentences for each theme [18].
A novel technique was proposed for summarizing text using a combination of Genetic Algorithms (GA) and Genetic Programming (GP) to optimize rule sets and membership functions of fuzzy systems by Kiani and Akbarzadeh. [6]
Nitin Madnani et. al performed multi-document summarization by generating compressed versions of source sentences as summary candidates and using weighted features of these candidates to construct summaries.
They combined a parse-and-trim approach with a novel technique for producing multiple alternative compressions for source sentences. [7]
Oi Mean Foong1,et al. investigated recent techniques and challenges on advances of automatic text summarization. [1]
In our method, the user insert the document which he would like to get a summary plus a key words that are used to pave a road that leads to a custom summary.
The method depends on the concepts of field association terms, co-word analysis, power link algorithm.
Humans can recognize the field by finding the specific terms, these words called Field Association words (FA words). So it is more effective if the search engines could pick these words, FA terms, from the queries and use them as the bases of searching process. [8]
Field association terms (FATs) dictionary concept was studied and tested for English. Arabic language has many differences from the English language so it needs special techniques for preprocessing before applying the power link algorithm.

978-1-880843-93-2 / copyright ISCA, CAINE 2013
September 25-27, 2013, Los Angeles, California, USA

Rokaya and Nahla, depending on available FATs dictionary in English, proposed a multilingual FATs dictionary in English and Arabic. [10]

Despite there are as many as methods of summarization but there is no method of summarization which considered a custom summary that depends on the user need. Even the methods that tried to rely on the user keywords, they followed these key words blindly without any analysis to the actual relations between these key words or other related key words.

In our method, the user insert the document which he would like to get a summary plus a key words that are used to pave a road that leads to a custom summary.

The method depends on the concepts of field association terms, co-word analysis, power link algorithm. In this work we introduce a custom summary based on key words or field of interest that a user determine or select.

The remaining parts of this paper is organized as follows, an over view of the concepts of field association terms and power links are presented in sections 2 and 3 respectively. The details proposed method is given in section 4. Experiments and results are given in section 5.

2. Field Association Terms

It is natural for people to identify the field of document when they notice specific words. These specific words are referred as Field-Association terms (FA terms); specifically, they are words that allow us to recognize intuitively a field of text or field –coherent passage. Therefore, FA terms can be used to identify the field of a passage, and can be also used to classify different fields among passages. For these reasons FA terms can be used as a clue to identify a passage field. FA terms can be either words or phrases. [12]

Field association terms (FA terms) are the words that indicate each subject matter category in the classification scheme. [13]

We define a minimum term (or a word), as one which cannot be further divided without losing its semantic meaning, as a single FA TERM (single FA Term). Compound FA terms are defined to consist of two or more single FA terms. Both terms are expressed by enclosing them within quotation mark. A compound FA TERM is regarded as being single if it loses its field information when divided. Compound FA terms (e.g. nuclear weapon, consumption tax or global warming) are considered to be simple FA terms because document field information is easily lost when those compound terms are divided. So here, proper nouns (e.g. Atlanta Braves, Oakland Athletics and South Africa) are considered to be simple FA terms. Personal names (e.g. Sammy Sosa, Carlos-Delgado) are considered to be simple FA terms, but proper noun containing a title (e.g. Coach T. Lasorda) are divided into two single terms: ‘‘Coach’’ and ‘‘T. Lasorda’’, which belong to the same document field Baseball but are on different levels. [14]

Field means a basic and common knowledge that can be used in human communication [15] and for convenience; hierarchical fields are categorized as Sub-Fields and Super-Fields. Therefore, Pitcher can relate to sub-field Baseball of superfield SPORTS and Pitcher may be classified SPORTS/Baseball. Selecting useful FA terms requires consideration of relationships between simple and compound FA terms and field classification. [14].

3. The Power Link

The term power link was proposed by Rokaya and Atlam, 2010, as a method of building a dynamic field association terms dictionary. Power link algorithm presented a new rules to improve the quality of filed association terms (FATs) dictionary in English [9].

The origin of this concept comes from the co-word analysis researches. Co-word analysis considers the dynamics of science as a result of actor strategies. Changes in the content of a subject area are the combined effect of a large number of individual strategies. This technique should allow us in principle to identity the actors and explain the global dynamic [16].

If any two terms t_1 and t_2 belongs to a document D we will say that there is a link between t_1 and t_2. The power of this link will be measured by the function $P_D(t_1,t_2)$ where:

$$P_D(t_1,t_2)=\frac{|D|\times cr(t_1,t_2)}{\underset{i,j}{average}(L(t_{1i},t_{2j}))} \quad (1),$$

where $|D|$ is the number of different terms in the document , $cr(t_1,t_2)$ is the co-occurrence frequency of t_1 and t_2 in the document D and $\underset{i,j}{average}(L(t_{1i},t_{2j}))$ represents the average distance between any instants t_{1i} and t_{2i} of the terms t_1 and t_2 in the document D. For more details see Rokaya and Atlam, [11].

To estimate the power link between two terms t_1 and t_2 over a given corps we define the function $P_C(t_1,t_2)$. This function can be defined as:

$$P_C(t_1,t_2)=\underset{D\in corps}{average}P_D(t_1,t_2) \quad (2)$$

This function states that the terms t_1 and t_2 will tend to appear nearer together if the value of this function reasonably high.

Let $P_s(w,D)=\underset{f\in D}{average}P(w,f), f\in S$ be the power link between the word w and the field $\langle S\rangle$ in the document D, if the set of field association terms that belong to the field $\langle S\rangle$ in the document D is empty then $P_S(w,D)=0$.

The power of link between a word w and the field $\langle S\rangle$ will given by $P(w,\langle S\rangle)=\underset{D\in S}{average}P_S(w,D)$. (3)

4. The proposed method

The summarizer go through the following steps.

1. determine depending on the key words or the field of interest the field association terms. To achieve this the summarizer needs a training phase to learn the field association terms and their power link relationship. This means that the summarizer build the map of field association links in a given corps.
2. The summarizer guess the corresponding English terms and then it gets the required words to begin the field coherent passages extraction.
3. The summarizer give a score for each retrieved passages and arrange them according to their score. If there is a defined number of line that the summary cannot exceed. The summarizer add passages till the number of lines be greater than the predefined threshold then the summarizer delete the last added one. To not harm the logic of the summary the summarizer shows the resulting summary in order according to what appear in the original text.

This approach is different from any other approach. In the beginning, it is far away from direct retrieval since we do not

retrieve the passage that contain the key words or even translated key words. Instead, we retrieve the field association terms that have a strong relation or a high power link connection with respect to the field under interest. Also, this approach is not a blind machine translation for the whole text or even to the passages that contain as many as of the key words. Instead, it retrieve the most connected passages to the field of interest through an intelligent guessing of the most related passages to the field.

Despite there are as many as methods of text extraction to produce a summary or a custom summary but the methods that are developed to produce a custom summary depends on the given terms directly. In the proposed method the custom summary is produced by extracting passages that contains the keywords entered by the user and the FA terms that are strongly related to the user keywords. For example, if the user entered the term "*cryptography*" the system will connect this term to "*Symmetric key*", " *Public key*" " *cipher*", " *encryption*" and "*decryption*". But since some terms related to the given keywords with different levels, the system activates the power link algorithm to pick those terms that strongly related to the given key words. The power link algorithm is dynamic since it depends on the corps. This algorithm is designed to measure the strength of relatedness between the terms in a specific field. In this method, the user insert the document which he would like to get a summary plus a key words that are used to pave a road that leads to a custom summary.

In fact, our work is a combination of key phrase extraction for a given corps. Then, we build the connections between these key words. These key words will be the base for the rest of the work. The user enter the initial keywords, the system picks the nearest terms in the map for these key words and use them to extract the coherent passages related to them. The summarization algorithm give a result depending on these extracted passages.

The following figure illustrates the main steps of the algorithm

The algorithm consists of two main parts. The first part is the production of the FA terms and the second part is production of coherent passages. The last step is a minor step to extract the passages that have a stronger link to the FA terms. In what follows we explore these parts.

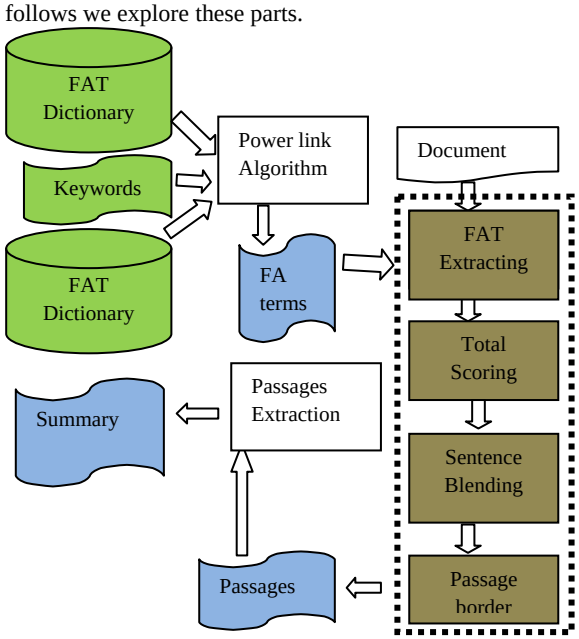

Figure1. System configuration

4.1. FA terms extraction

It is natural for people to identify the field of document or a passage when they see specific words. These specific words is referred as *field-association terms* (FA terms). A minimum unit (or a word), is defined as one which cannot be further divided without losing its semantic meaning, as single FA term. For example the term "*symmetric key*" is a single FA term since if it divided into "*symmetric*" and "*key*" it will directly loss its semantic meaning and its relation to the field of *cryptography.*

For a given keywords, the algorithm computes the power link to each keyword with respect to each field in the corps. The algorithm picks the field with the largest power link for each keyword. In every field the algorithm extracts the FA terms that have a power link greater than α, where α is a given threshold. The union of all extracted FA terms and the keywords forms the terms that will be used in the next steps.

4.2. Total Scoring and sentence blending

The first characteristic of a topic flowing is defined as *continuity* and the second as *transition.* Passages with different field theme are delimited and the field duplication of passages is prevented. Every sentence is supposed to have one subject or less. The field that a sentence presents is called a theme field, which is denoted by F_{theme} [13].

F_{theme} is maintained by *continuity* or changed by *transition* through sentences. For a given sentence S that contains FA terms $(FA_1, FA_2, FA_3, ..., FA_n)$ the power link between sentence S and all fields is computed according to the formula $P(S,F_j)=\sum_{i=1}^{n} P(FA_i,F_j)$ and the F_{theme} for the sentence S is determined by the field that gives $\max_j P(S,F_j)$. If the current sentence has two or more fields with the following property $\left|P(S,F_j)-P(S,F_k)\right|<\lambda$, where λ is small enough, then S is said to have no F_{theme}. Also if the set of FA terms contained in S is empty, S is said to have no field If the current sentence S has the same F_{theme} as the previous sentence, or has no F_{theme}, or has no field the current sentence is appended to the same passage. And if the current sentence S has a different F_{theme} from the previous sentence, then S is delimited and a new passage begins. According to the previous rules the passages borders are determined and the passage extraction can be done.

4.3. Passages extraction

For every document we have two sets. Extracted FA terms, denoted by SFA, and the document as a set of delimited passages. For a passage PS that belongs to a specific F_{theme} the score of the passage PS is computed according to the formula $P(PS)=\sum_{FA\in SFA} P(FA,F_{theme})$. The passages are ranked according to their scores and passage with the highest score is chosen as the base summary. If the length of this passage is less than the desired summary length the next passage, in the ranked list, is appended and again the total length of the two passages is less than to the desired length a third passage is appended. The process is repeated until a

desired length is achieved. The final summary shows the passages in the order that they had in the original document.

5. Experiments and Evaluation

There are two sets of experiments should be done here

(a) Experiments to evaluate the efficiency of delimiting the passages.

(b) Experiments to evaluate the efficiency of the extractor

5.1 Experiments to evaluate the efficiency of delimiting the passages.

To estimate the quality of the presented passage delimitation method, fifty articles composed artificially of some fields. each articles length were between one KB and 3 KB. Now we know the real field that each passage should belong. Two measures are used here, namely, *precision* and *recall*. *Precision* and *recall* are defined as follows.

$$\Pr ecision = \frac{P_{accord}}{P_{output}} \text{ and } \operatorname{Re} call = \frac{P_{accord}}{P_{answer}}$$

where P_{output} is the number of characters in passages that the system produces, P_{answer} is the number of characters in correct answer passage human decides, and P_{accord} are those occurred in both in the system output and in the correct answer passages. Figure 2 presents the results of retrieved passages for each measure. Figure 3 presents the results of *recall* and *precision*. The average value of *precision* is 0.78% and the average value of recall is 0.74%. These results proof that the proposed system gives a high accuracy.

To ensure the strong of the results, F is also calculated using the formula.

$$F = \frac{2 * \Pr ecision * \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call}$$

The average value of F is 75% which reflects a high performance of the algorithm.

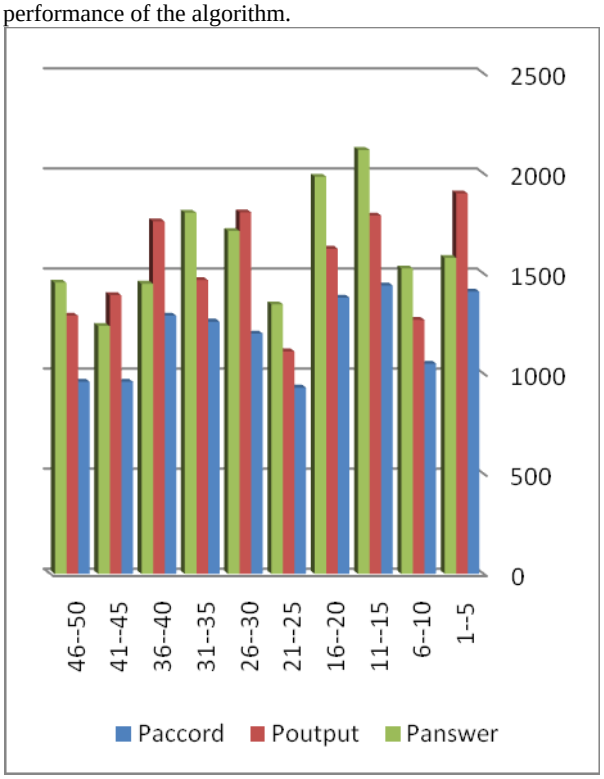

Figure2 Results for Paccord, Poutput and Panswer

5.2 Experiments to evaluate the efficiency of the extractor

Since the goal of summarization schemes is to automate a process that has traditionally been done manually, a comparison of automatically generated extracts with those produced by humans would provide a reasonable evaluation of these methods. [17]

Assuming that a human would be able to identify the passages beginnings and ends and also would be able to identify the most important passages effectively. If the set of passages selected by an automatic extraction has a high overlap with the human generated extract, the automatic method should regarded as effective. In all experiments, two manually extracts are generated. In both cases the user is asked to extract a specific number of paragraphs, delimited using our method, based on a given keywords for each document.

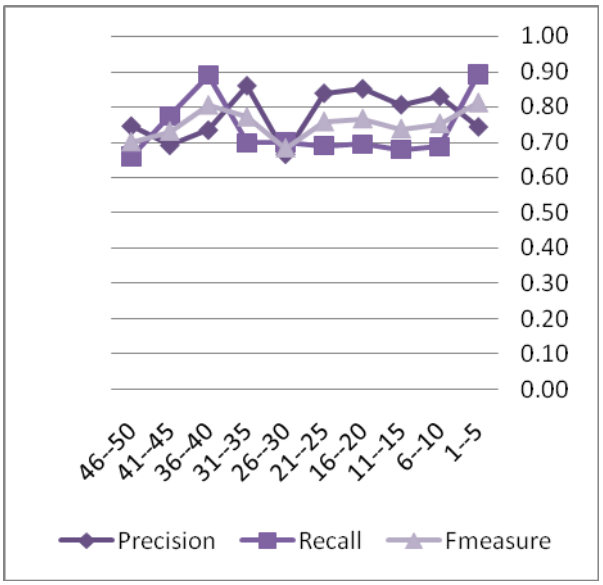

Figure3: *Precision* and *Recall* results

The evaluation takes the following form,

- A user walks to the system and picks a document for a custom summary based on a given keywords
- the summary based on the same keywords and presents it to the user. The user compares this summary to his own one.
- In the second case, the system generates the summary using our proposed method to get the automatic one and the user compares this summary to his own.

In both cases, the user satisfaction is measured by the number of common passages between his own summary and the target summary (manually or automatically extracted).

For more strict comparison, two other automatic summarization are considered:

a) A random summarization is done by randomly extracting the required number of passages from all passages that contains at least one of the given keywords. Any extraction algorithm must do better than the random method.

b) Another method depends on extracting the first passage that contains at least one of the given keywords, then we go through the passages to pick the second passage that contains at least one of the given keywords. This process is repeated until the required number of passages are extracted.

Fifty articles were chosen for the evaluation process. It is known that the overlapping between two manual free extracts,

free extracts are those that do not depend on any given keywords, is lower than 50 %. [17].

Table1: Evaluation measures for automatic extraction

Overlap between manual extracts: 85.7 %			
Algorithm	S1	S2	Average
Power link	87.92	75.26	81.59
Random	57.74	48.36	53.05
Initial	80.42	74.67	77.54

This can be understood as follows, each human has interests that may differ from the other human. So the lower overlapping between the two manual extracted summaries reflects the lower overlapping between their needs. But if the needs were unified, it is expected the percentage of overlapping will be increased. The overlapping between the two manual extracts were 85.7 % which were expected. Table 1, summarizes the results of the experiments. The overlapping between the automatic extracts and the manual extracts. From this table the automatic summary do better than the random and the initial methods. But the results of the initial summary were extremely unexpected. The overlapping between the initial method and the manual extracts were 80.42 % and 74.67%. This high result were very strange. But the nature of the human is to not go deeply through the documents to get the required summary. In most cases the user will add the passage whenever he found some of the keywords. He will repeat this process to get the required number of passages more than interesting to get most of the important passage related to the keywords. Also since he was advised to complete the reading to the end of each documents, he might found some more important passage and will replace them in the place of earlier added passages. Finally the results for the power link method were 87.92% and 75.26%. The average is 81.59%. These results proof that the proposed method is fine and gave a considered performance than the initial method.

6. Conclusion

The results showed that the automatic extraction based on power links do better than the random and the initial lead methods. But the size of the extracts paragraphs were some long, in many cases, it were more than 40 lines. This is due to appending many non field or non specific field in each passage. In the future work, it will be intended to develop a methods to eliminate some of these sentences without reduce or losing the context value of the summary.

References

[1] Oi Mean Foong1, Alan Oxley1 And Suziah Sulaiman1,(2010) Challenges And Trends Of Automatic Text Summarization, International Journal Of Information And Telecommunication Technology, Vol. 1, Issue 1, Pp 34-39.

[2] Vishal Gupta And Gurpreet Singh Lehal, (2010) "A Survey Of Text Summarization Extractive Techniques", Journal Of Emerging Technologies In Web Intelligence, Vol. 2, No. 3, August.

[3] Manisha Prabhakar*, Nidhi Chandra,(2012) Automatic Text Summarization Based On Pragmatic Analysis, International Journal Of Scientific And Research Publications, Volume 2, Issue 5, May

[4] Massih-Reza Amini, Patrick Gallinari, (2001) Self-Supervised Learning For Automatic Text, Summarization By Text-Span Extraction 23rd Bcs European Annual Colloquium On Information Retrieval, Pp 1-9,

[5] Constantin Or˘Asan.(2003), An Evolutionary Approach For Improving The Quality Of Automatic Summaries,Multisumqa '03 Proceedings Of The Acl 2003 Workshop On Multilingual Summarization And Question Answering - Volume 12, Pages 37-45,

[6] Arman Kiani -B, M. R. Akbarzadeh –T. (2006) Automatic Text Summarization Using: Hybrid Fuzzy Ga-Gp, 2006 Ieee International Conference On Fuzzy Systems Sheraton Vancouver Wall Centre Hotel, Vancouver, Bc, Canada July 16-21, 977-983

[7] Nitin Madnani, David Zajic, Bonnie Dorr, Necip Fazil Ayan & Jimmy Lin. (2007) Multiple Alternative Sentence Compressions For Automatic Text Summarization, In Proceedings Of The 2007 Document Understanding Conference (Duc-2007) At Nlt/Naacl 2007, April 2007, Rochester, New York,

[9] Mahmoud Rokaya, Elsayed Atlam, Masao Fuketa, Tshering C. Dorji And Jun-Ichi Aoe, (2008). Ranking Of Field Association Terms Using Co-Word Analysis, Information Processing & Management, Volume 44, Issue 2, March, Pages 738-755.

[10] Mahmoud Rokaya And Abdallah Nahla. (2011) Building A Multi-Lingual Field Association Terms Dictionary, International Journal Of Computer Science And Network Security, Vol. 11 No. 3 Pp. 208-213.

[11] Mahmoud Rokaya And Atlam, E-S. (2010) 'Building Of Field Association Terms Based On Links', Int. J. Computer Applications In Technology, Vol. 38, No. 4, Pp.298–305.

[12] Uddin, S. Elmarhomy G., Atlam, E., Fuketa, M., Morita K., Aoe J. (2007), Improvement Of Automatic Building Field Association Term Dictionary Using Passage Retrieval. Information Processing & Management Journal, Vol. 43, Pp. 1793–1807

[13] Lee, S., Shishibori, M., Sumitomo, S., Aoe, J., (2002) Extraction Of Field-Coherent Passages. Information Processing And Management, Vol. 38, Pp. 173-207.

[14] Atlam E., Morita, K., Fuketa, M., & Aoe, J. (2002), A New Method For Selecting English Field Association Terms Of Compound Words And Its Knowledge Representation. Information Processing And Management, Vol. 38, Pp. 807-821.

[15] Kawabe, K., & Matsumoto, Y., (1998) Acquisition Of Normal Lexical Knowledge Based On Basic Level Category. Information Processing Society Of Japan, Sig Note, Vol. 125(9), Pp. 87–92

[16] Callon, M., Courtid J., & Ladle, F., (1991) Co-Word Analysis As A Tool For Describing The Network Of Interactions Between Basic And Technological Research: The Case Of Polymer Chemistry. Science Metrics, Vol. 22(1), Pp. 155-205.

[17] Mandar Mitra, Amit Singhal, Chris Buckley , (1997) Automatic Text Summarization By Paragraph Extraction, Workshop On Intelligent Scalable Text Summarization,

[18] Atefeh Farzindar And Guy Lapalme. Letsum, (2004), A Text Summarization System In Law Field. The Face Of Text Conference (Computer Assisted Text Analysis In The Humanities), P. 27-36, Mcmaster University, Hamilton, Ontario, Canada, Nov