The Forth Scientific Conference of the College of Science University of Kerbala 2016

Journal of University of Kerbala

Automated System To Summarize Texts In Arabic Using Intelligence Techniques

النظام الالي لتلخيص النصوص العربية باستخدام التقنيات الذكية

Prof. Dr. Ahmed Tariq Sadiq
Computer Science Department

University Of Technology
Baghdad, Iraq

Suaad Abed Al-Wahab Ismail
Computer Science Dep./ College of Education

University Of Mustansiriya
Baghdad, Iraq

Abstract

The proposed system was adopted extraction method in summarizing the Arabic text using natural language processing by stripping the text of the Arabic stop words and some precedents and suffixes, as well as through statistical operations of words within the text, and thus get a sentences most important in the text for a summary. Have been adopted more than one method to offer this proposed system where the use two algorithms of the steaming algorithms are Lovins steamer algorithm from cutting methods because it is based on the removal of the longest suffix and this is what was adopted by (storage Luxor word), also used n- grams algorithm of statistical methods in terms of their dependence on the similarity between the words down to the root account and repeat the words in the text to get the summary.The essence of the proposed system is to find the best parameter values of all the features used in all ratios used to summarize.

الخلاصة

اعتمد النظام المقترح طريقة الاستخراج في تلخيص النص العربي باستخدام معالجة اللغة الطبيعية من خلال تجريد النص من كلمات وقف العربية وبعض السوابق واللواحق ،وكذلك من خلال العمليات الإحصائية للكلمات داخل النص، وبالتالي الحصول على الجمل الأكثر أهمية في النص للحصول على الملخص. وقد تم اعتماد أكثر من طريقة لتقديم هذا النظام المقترح حيث تم استخدام اثنين من خوارزميات التجذيع هي خوارزمية لوفينز ستيمرمن طرق القطع لأنها مبنية على إزالة أطول لاحقة وهذا ما تم إقراره من خلال(تخزين الكلمة الأقصر)، وكما تم استخدام خوارزمية ن -غرام من الأساليب الإحصائية من حيث اعتمادها على التشابه بين الكلمات وصولا الى حساب الجذر وتكرار الكلمات في النص للحصول على الملخص.جوهر النظام المقترح هو العثور على أفضل القيم المعلمة من كل الميزات المستخدمة في كل نسب تلخيص المستخدمة.

1. Introduction

Finding the right information that you want to search for them without wasting a lot of time is difficult of the large number of available information, which is increasing day by day.To address this problem, is to prepare a summary of this information, which makes easy to find a document. Still use this technique in many areas.

The text summarization process that takes a very long time and very expensive, this is the reason, finding a method to make your computer perform this task [1]. Automatic text summarization it is the text contains non-redundant information while maintaining the sense of the original text, which was obtained through the use of the computer, for information retrieval and provide recognition at a glance a certain type of information in the text [1]. They can also summarize the text that plays an important role in supporting the user in the search for relevant information. As can be defined Automatic text summarization: is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document. As the problem of information overload has grown, and as the amount of data has increased, so has interest in automatic summarization.

Technologies that can build a coherent summary take into account variables such as distance, dropping a line style and sentence structure. An instance of the role of summarization technology is search engines such as Google.[2]. As more documents lacked of abstracts , because summarize these documents manually take a very long time and very expensive as well as tiresome to set the summary So it had to be to discover the automatic summarization method [3]. The aim of this thesis is to find a way of summarizing that describes the content of the document.

In order to save time and effort in the summary process and the distinction between the main ideas from secondary ideas and dispensing read the original text and Achieve this goal is through automatic text summarization.

2. Related Works

1- In [4] [2012], Iskandar Keskes ,Mohamed Mahdi Boudabous, Lamia Hadrich Belguith, Mohamed Hédi Maâloul, present proposed a comparative study of three methods to automatically summarize the Arabic texts where the first method is based on the symbolic approach and the second method is based on a numerical approach is based on the third hybrid approach. Where the application of these methods respectively by Systems "RIA and Resume and hybrid". The results showed that the symbolic approach is less of performing numerical approaches that combine the two approaches in a hybrid approach showed better results in summarizing texts**. 2-** In [5] [2013], Hanane Froud, Abdel minim Lachkar and Said Alaoui Ouatik, present proposal to the extent of the benefits of the summary using latent semantic analysis model, and compare the results obtained on the basis of summarizing the baseline with full Arabic text of the documents. To take the similarity / distance measures for the three times, off without stopping, and with the

cessation of the use of Khoja Steamer, Wal Larkey steamer. Has been used as an LSA technique of techniques that remove noise from the documents and selection sentences of the most prominent to reflect the original documents. Through experiments observed that the Euclidean distance and cosine similarity and Jaccard have effectiveness similar measures to produce group a more cohesive of Pearson correlation averaged k L difference whether with or without stemming **. 3-** In [6][2014], Elham Mahdipour, present compared between Parsing system and Phoenix to summarize texts automatically, where Phoenix employs a hybrid algorithm (SA, GA) to determine the sentences for the summary and then Summary evaluation of the product. The results showed that the hybrid algorithm (SA, GA) system more quality where Phoenix got 64.35% vs. 59.86% of the Parsing system**. 4-** In [7] [2014] Iskandar Kskas, Malak Allhiwi, Farah Ben Amara and Lamia Hadrich Belguith: present a new approach in automatically texts summarized based on the fragmented structure of theory (SDRT). The first phase and the rhetorical structure :This method consists of two phases are created by extracting the rhetorical relations between units of the text and then drawing SDRT scheme which represents the rhetorical structure of the text, wherein the second phase used this SDRT scheme to get a summary and through the reductase of SDRT planned remove rhetorical relations unwanted in the summary selected has been the inclusion of this method within the SDRTResume system for evaluation. As has been selected this system on code of accuracy rate was 56%.

3. Summarization Types

There are two major types of text summarizing: Abstract and Extract.

a) Extract Summarization

What extraction techniques only copy information considered relevant by the system for the summary (e.g., major items, sentences or paragraphs) [8].

Since the basic version of the summary is based on the extraction is to determine the characteristics of sentences with, and put them together in summary [9].

The text summarized extracts text from the original text on the basis of statistical methods or by applying a heuristic or mixing the two together. [10].

b) Abstract Summarizes [9]

Abstraction is at least some of the articles that do not exist in the original document (such as rewriting or viewpoint of the document, etc.). Or is the rendering of the original text, which carries connotations level of the original text and involve linguistic processing at the point of what representation [11].

Production process as it requires rewriting the original text in the shortest by replacing wordy concepts with other shorter version, which necessitates the ability to various problems challenging AI management. For instance, the phrase "was eating apples, bananas and oranges," can be resumed as "eat the fruit." [12].

4. **The Propose System**

Have been adopted more than one method to offer this proposed system where the use two algorithms of the steaming algorithms are Lovins steamer algorithm from cutting methods because it is based on the removal of the longest suffix and this is what was adopted by (storage Luxor word), also used n grams algorithm of statistical methods in terms of their dependence on the similarity between the words down to the root account and repeat the words in the text to get the summary.The essence of the proposed system is to find the best parameter values of all the features used in all ratios used to summarize.

Figure (1) shows the structure of the proposed system

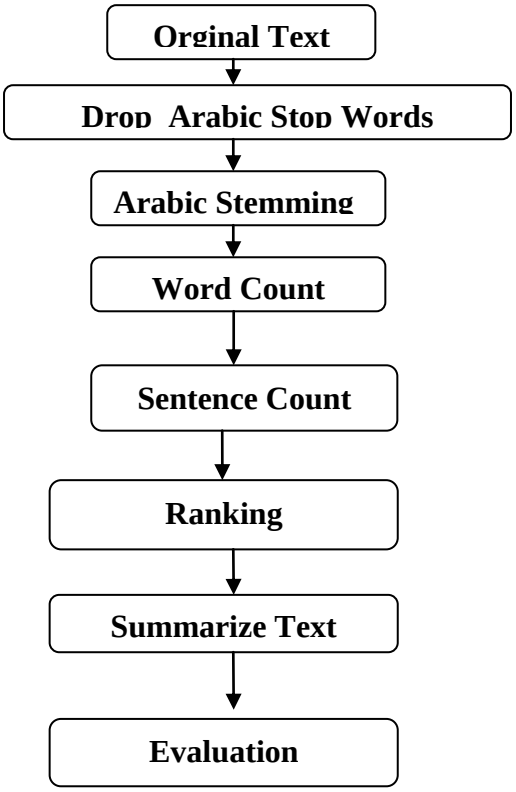

Figure (1) Block Diagram Of The Proposed System

4.1 Original Texts

That meant the original text is the text before the summary were a number of texts on various subject collections (politics, news, cultural stories, medical, scientific and other) different lengths for the application of the proposed system and access to the summary where the number was more than thirty text .

For example:

ميكوموتو

ميكوموتو رجل ياباني، قروي عادي جدًا، ولد في قرية " توبا" كان والده رجل فقير يبيع الأرز المسلوق.. ومنذ طفولته كان يساعد والده ويقضي نهاره في دفع عربة صغيرة لبيع الأرز.. وفي سن الثامنة عشرة عمل بصيد الأسماك والغوص وصيد اللؤلؤ وبيع الأصداف وكان يهوى جمع النادر منها.

4.2 Drop Arabic Stop Words

It is the process of deleting letters and words meaningless with high repetition in the text. Amounted Number of stop words to more than 450 words where which includes all of the question tools and the of connectivity tools and the names of the signal and others. Table (1) shows some of the stop words.

Table (1)
Samples Of Arabic Stop Words

من	ماذا	مما	ممن	أين	أينما	حيثما	كيفما	مهما	تلك
ذلكم	ذلكما	ذلكن	ذوا	ذواتا	كذلك	هكذا	هنا	هناك	هنالك
تجاه	جميع	حسب	حيث	سبحان	شبه	كل	لما	مثل	مع
أقل	أكثر	دونك	إياهما	إياهم	إياه	إياكن	التي	الذي	الذين
اللتين	اللذان	اللذين	اللواتي	ذا	ذات	أب	أخ	قبل	ذو

Here it must be noted that after the application to delete the words stop algorithm. The previous text output was a text without the words Stop as is evident below.

ميكوموتو

ميكوموتو رجل ياباني، قروي عادي جدًا، ولد قرية "توبا" والده رجل فقير يبيع الأرز المسلوق . . طفولته يساعد والده ويقضي نهاره دفع عربة صغيرة لبيع الأرز . سن الثامنة عشرة عمل بصيد الأسماك والغوص وصيد اللؤلؤ وبيع الأصداف يهوى جمع النادر.

Stages generate summary is a standard, so the focus will be on stage Arabic stemming with count words.

4.3 Arabic Stemming With Count Words

Phase Arab stemming a very important stage of the text mining and where words are returned to the assets without the use of the dictionary where the removal of part of precedents and suffixes and some letters of non-original until you find the words that have the same trunk when you search for any one of them precedents is (ال،فال،بال،وال،لل) and suffixes are (ون،ان،ت،وا،ا،ة،كن،كم،كما،ك،ن) and then store the resulting words in the buffer and then compare each word with the rest of the words

to search for words like them and find out how often each word in the buffer, as shown in the algorithm (1) below:

The algorithm (1) illustrates stemming with count word

```
Algorithm
Input: Text Without Arabic Stop Words
Output: Pure Words With Frequency
Begin
    While not end of text do
        Get current word
    If any element in prefix found in the current word then
      Deleted prefix
    If any element in suffix found in the current word then
      Deleted suffix
    If the current word found in the buffer then
      Increase the frequency by one.
    Else If the current word is different from any word in the buffer two letters at the beginning or at the end and in the precedents of the letters or suffixes then
      Store the shortest word
      Increase the frequency by one.
    If the current word is different from any word in the buffer with a single character in the beginning of a word or the end of the word then
      Store the shortest word
      Increase the frequency by one.
    Else
      Store the new word in the buffer
      Let the frequency of the new word equal one
    End while
End algorithm
```

When applying the stemming algorithm with count words in the text without stop words results are as shown in the table (2).

Table (2)
Word and Count

Word	Count	Word	Count
ميكوموتو	2	دفع	1
رجل	2	عربة	1
يبيع	2	صغيرة	1
ارز	2	لبيع	1
نهاره	1	ثامنة	1
عادي	1	عشرة	1
جدا	1	عمل	1
ولد	1	بصيد	1
"توبا"	1	اسماك	1
فقير	1	غوص	1
ياباني	1	وصيد	1
قروي	1	لؤلؤ	1
مسلوق	1	أصداف	1
طفولته	1	يهوى	1
يساعد	1	جمع	1
ويقضي	1	نادر	1

4.4 Sentence Count

At this stage, you return to the original text and dividing it into sentences where it ends every sentence (.,؟ ,!,،,؛).

Then examine each sentence to see the most number of words frequently included in each sentence and thus know Rank each phrase and then Order Rank each sentence Order Descending. After applying this algorithm to the output of the previous stage we get what is shown in table(3) below:

Table (3)

Sentences Rank

Word count	Rank	Sentence
ميكوموتو(2)،رجل(2)ياباني(1).	5	ميكوموتو رجل ياباني
قروي(1)عادي(1)جدا(1)قر(2).	5	قروي عادي جدا
ولد(1)قر(2)"توبا"(1)يده(2)فقير(1)يبيع(2)أرز(2)مسلو ق(1)رجل(2).	14	ولد في قرية "توبا"كان والده رجل فقير يبيع الارز المسلوق
طفولته(1)يساعد(1)ده(2)ويقضي(1)نهاره(1)دفع(1)عرب ة(1)صغيرة(1)لبيع(1)ارز(2).	12	ومنذ طفولته كان يساعد والده ويقضي نهاره في دفع عربة صغيرة لبيع الأرز
ثامنة(1)عشرة(1)عمل(1)بصيد(1)اسماك(1)غوص(1)و صيد(1)لؤلؤ(1)اصداف(1)يهوى(1) جمع(1)نادر(1)من(1).	13	وفي سن الثامنة عشرة عمل بصيد الأسماك والغوص وصيد اللؤلؤ وبيع الأصداف وكان يهوى جمع النادر منها

4.5 Ranking

The rank is the sum of the occurrences of the most words repeatedly in wholesale and represents the power of wholesale or how important sentence in the text and on the basis of the value of the rank be pulling strings of text, according to the ratio that is selected by the user.

4.6 Summarize

It is the final step of the proposed system to generate a summary in this step is pulling strings with high importance (most Rank) of the original text as given in the original text is done by percentages that are selected by the user of the (WN, SN) would receive a summary commensurate with selected ratio and excellent speed. For example (this summary by 50% (WN) and 50% (SN).

ميكوموتو

ميكوموتو رجل ياباني، ولد في قرية " توبا" كان والده رجل فقير يبيع الأرز المسلوق. ومنذ طفولته كان يساعد والده ويقضي نهاره في دفع عربة صغيرة لبيع الأرز.

5. Experimental Results

It has been collecting more than 30 text various subjects and disciplines (cultural - social -scientific - religious - political - Medical - stories).

These texts were different length has been a special extraction texts experiment summarized (the application of the proposed system on the texts for texts summarized) and rates 40% -50% -60%, respectively, for the number of words and the number of sentences and at rates ranging between 40% -50% - 60 % of the sentences that will be withdrawn from the original text to create a text summary.

Then after that has been summarized texts (30 text) to several people for the purpose of evaluation of the Human out a realistic assessment includes the content and concept together through a questionnaire. This questionnaire included 30 text summaries of rates of 40% - 50% -60% for each of the number of words The number of sentences. This questionnaire was presented to the people with a bachelor degree or above and the various terms of reference and the table(4) below shows the details of where the first column represents sequence the text and the second column containing text title and the third column that contains a proportion of the summary 40% fourth column contains the ratio of 50% and fifth Column contains a ratio 60% was the highest evaluation drawer got it all proportion.

Table (4)

The Results Of The Questionnaire Were Presented To The People With A Bachelor's Degree Or Above

ID	Text Title	Summarization Ratio		
		40%	50%	60%
1	الإعلام الدعائي	Good	Very good	Very good
2	اربيل	Excellent	Very good	Excellent
3	أنت الملاك الطاهر	Very good	High-good	Excellent
4	حدث جدلي(عمران العبيدي)	High-good	High-good	Very good
5	ميكوموتو	Good	Very good	Excellent
6	أنت الحنان ياأمي	Very good	Very good	Very good
7	قصة واقعية	Medium	Very good	Excellent
8	الخمر أم الخبائث	Good	Very good	Very good
9	مرحلة مابعد النتائج(عمران العبيدي)	Very good	High Medium	Medium High
10	حكاية المنديل السحري	Good	Very good	Very good High
11	داعش قرش البحر	Very good	Excellent	Very good
12	طموحات الشباب والبرلمان القادم	Very good	Excellent	Excellent
13	الفساد المالي خطورته وسبل مكافحته	Excellent	Excellent	Excellent
14	صنع في العراق	Very good	Very good	Excellent
15	التحالف الوطني وأخوة يوسف	Excellent	Excellent	Excellent
16	قراءة في قانون البنى التحتية	Excellent	Excellent	Excellent
17	بدانة الطفل أم بدانة المجتمع	Very good	Excellent	Excellent
18	عصر تدفق المعلومة	Very good	Excellent	Excellent
19	الصبر دواء لكل داء	Excellent	Excellent	Excellent
20	الأيمان في قيادة العقل	Excellent	Very good	Excellent
21	الإعلام الثقافي الذي نريد	Good	Medium	Very good
22	ديمقراطيون ولكن	Medium	Medium	Good
23	سيناريو التشكيك	Medium	Good	Excellent
24	النباتات العطرية والطبية	Good	Very good	Excellent
25	أهمية تعليم اللغة العربية	Very good	Excellent	Excellent
26	الصداقة الحقيقية بين الحقيقة والخيال	Good	Good	Excellent

27	**حمى التسقيط السياسي**	**Very good**	**Excellent**	**Excellent**
28	**أثار المعاصي**	**Excellent**	**Excellent**	**Excellent**
29	**أوربا والإسلام في العصور الوسطى**	**Medium**	**Good**	**Very good**
30	**قصة ذات عبرة**	**Very good**	**Very good**	**Very good**

It was destined for the ratio of the proposed system according to the questionnaire above are excellent.

Ratio of summarization

A- 40% got a very good the evaluation.

B- 50% got between the evaluation is very good and excellent.

C-60% got an excellent evaluation.

It was also presented the same questionnaire to experts in the Arabic language and the number was four .

Where the rate ratio of the proposed system according to this questionnaire is very good. As it is clear in the table (5).

Ratio of summarization

A- 40% got a very good the evaluation.

B- 50% got a very good the evaluation .

C-60% got a very good the evaluation.

Table (5)
The Results Of The Questionnaire Submitted To The Arabic Language Experts

ID	**Text Title**	**Summarization Ratio**		
		40%	**50%**	**60%**
1	**الإعلام الدعائي**	**Very good**	**Very good**	**Very good**
2	**اربيل**	**Very good**	**Very good**	**Very good**
3	**أنت الملاك الطاهر**	**Very good**	**Very good**	**Very good**
4	**حدث جدلي(عمران العبيدي)**	**Very good**	**Very good**	**Good**
5	**ميكوموتو**	**Very good**	**Very good**	**Very good**
6	**أنت الحنان ياأمي**	**Very good**	**Good**	**Very good**
7	**قصة واقعية**	**Very good**	**Very good**	**Very good**
8	**الخمر أم الخبائث**	**Good**	**Very good**	**Very good**
9	**مرحلة مابعد النتائج(عمران العبيدي)**	**Very good**	**Very good**	**Good**
10	**حكاية المنديل السحري**	**Good**	**Good**	**Good**
11	**داعش قرش البحر**	**Very good**	**Very good**	**Good**
12	**طموحات الشباب والبرلمان القادم**	**Very good**	**Very good**	**Good**
13	**الفساد المالي خطورته وسبل مكافحته**	**Very good**	**Very good**	**Very good**
14	**صنع في العراق**	**Very good**	**Very good**	**Good**
15	**التحالف الوطني وأخوة يوسف**	**Good**	**Good**	**Very good**
16	**قراءة في قانون البنى التحتية**	**Very good**	**Very good**	**Good**
17	**بدانة الطفل أم بدانة المجتمع**	**Good**	**Very good**	**Very good**

18	**عصر تدفق المعلومة**	**Very good**	**Very good**	**Good**
19	**الصبر دواء لكل داء**	**Very good**	**Very good**	**Very good**
20	**الأيمان في قيادة العقل**	**Very good**	**Very good**	**Very good**
21	**الإعلام الثقافي الذي نريد**	**Good**	**Good**	**Good**
22	**ديمقراطيون ولكن**	**Very good**	**Very good**	**Very good**
23	**سيناريو التشكيك**	**Very good**	**Very good**	**Very good**
24	**النباتات العطرية والطبية**	**Good**	**Good**	**Good**
25	**أهمية تعليم اللغة العربية**	**Very good**	**Good**	**Medium**
26	**الصداقة الحقيقية بين الحقيقة والخيال**	**Good**	**Good**	**Good**
27	**حمى التسقيط السياسي**	**Very good**	**Very good**	**Good**
28	**أثار المعاصي**	**Very good**	**Very good**	**Good**
29	**أوربا والإسلام في العصور الوسطى**	**Very good**	**Good**	**Very good**
30	**قصة ذات عبرة**	**Very good**	**Very good**	**Very good**

6. Conclusions

In this thesis, The proposed system in summarizing the Arabic text was adopted extraction method using the Arabic language processing by deleting the Arabic Stop words and some precedents and suffixes of the text as well as through statistical operations of words within the text and thus infer the most important sentences in the text to get the summary.

Summarizing the results were as follows:

1- Adoption the proposed system two of stemming algorithms (Lovins Steamer algorithm which is one of the Truncating Methods, N-gram algorithm of statistical methods). To produce a new algorithm and This new algorithm is excellent in terms of speed and performance.

2-Have been texts summarized display this system through a questionnaire and rates 40%, 50%, 60% for people with a bachelor's degree or above and the various disciplines and the percentage rate according to the questionnaire is excellent. Where the results ratios of summarized as follows:

A- 40% got a very good the evaluation.

B - 50% got between the evaluation is very good and excellent.

C- 60% got an excellent evaluation.

3-Then display of text questionnaire summarized the experts in the Arabic language and the number (Four experts) The result of the questionnaire is (very good).

Where the results ratios of summarized as follows:

A - 40% got Rating (very good).

B - 50% got Rating(very good).

C - 60% got Rating (very good).

4 -The proportion of errors in the proposed the Stemming is a maximum of 4%, where the percentage of error in most of the texts 3% and in some of the texts were 4%. But it is dropped, because the error occurred in low frequency word.

References

[1] Dr. Ahmed Tariq Sadiq, Dr. Saran Akram Chawishly, Kanar Shukr Muhammad**, "Text Summarization Using Hybrid Methods"**, Proceedings of the First Scientific Conference of the Iraqi Association for Information Technology, pages: 199-212, 2009.

[2] Dr. Ahmed Tariq Sadiq, Noor Amjed Hassan**, "Learning- Based Text Summarization Approach Using Association Rules and Statistical Measurements"**, Journal of Iraqi Association for Information Technology, pages: 1-13, 2008.

[3] Susanne Viestan, "**Three Methods for Keyword Extraction**", MSc. Thesis. Department of Linguistics, Uppsala University, (2000).

[4]Iskandar Keskes, Mohamed Mahdi Boudabous, Mohamed Hédi Maaloul1, Lamia Hadrich Belguith**, "Comparative study of three approaches to automatic summarization of Arabic documents**", Actes de la conférence conjoint JEP-TALN-RECITAL, Volume 2: TALN, pages 225–238,2012.

[5] Hanane Froud, Abdelmonaime Lachkar and Said Alaoui Ouatik, "**Arabic Text Summarization Based On Latent Semantic Analysis To Enhance Arabic Documents Clustering**", researchgate, pages: 1-15, Aug2013.

[6] Elham Mahdipour and Masoumeh Bagheri,"**Automatic Persian Text Summarizer Using Simulated Annealing and Genetic Algorithm**",International Journal of Intelligent Information Systems, Special Issue: Research and Practices in Information Systems and Technologies in Developing Countries, Vol. 3, No. 6-1, pages: 84-90, 2014.

[7] Iskandar Kskas, Malak Allhiwi, Farah Ben Amara and Lamia Hadrich Belguith, **"Summarizing automated Arabic texts based on the fragmented structure of rhetorical theory "SDRT""**, International Computing Confrence in Arabic (ICCA), Riadh-Saoudi Arabia, pages: 1-17, 30 MAY 2014.

[8] Hongyan Jing and Kathleen R. McKeown, "**Cut and Paste Based Text Summarization**", Department of Computer Science, Columbia University, New York, NY 10027USA, (2001).

[9] Dragomir R. Radev, Eduard H. Hovy, Kathleen McKeown,. "**Introduction to the Special Issue on Summarization"**. Computational Linguistics 28 (4),Pages: 399-408, (2002).

[10] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, "**Tapping the Power of Text Mining**", Communications ACM, Volume 49, Issue 9, pages: 76 – 82, New York, USA, 2006.

[11] Joel Larocca Neto, Alex A. Freitas, Celso A. A. Kaestner, "**Automatic Text Summarization using a Machine Learning Approach",** Pontifical Catholic University of Parana (PUCPR) Rua Imaculada Conceicao, pages: 1-10, (2002).

[12] J. Y. Delort, B. Bouchon - Meunier, and M. Rifqi, "**Enhanced Web Document Summarization Using Hyperlinks**", Proc. 14th ACM Conference on Hypertext and Hypermedia (HT'03), pages:208–215, United Kingdom, 2003.