PERSONALIZING TEXT SUMMARIZATION BASED ON SENTENCE WEIGHTING

Christos Bouras
Research Academic Computer Technology Institute and Computer Engineering and Informatics Department, University of Patras
N. Kazantzaki Str., University of Patras, 26500 Rion, Greece
bouras@cti.gr

Vassilis Poulopoulos
Research Academic Computer Technology Institute and Computer Engineering and Informatics Department, University of Patras
N. Kazantzaki Str., University of Patras, 26500 Rion, Greece
poulop@cti.gr

Vassilis Tsogkas
Computer Engineering and Informatics Department, University of Patras
University of Patras, 26500 Rion, Greece
tsogkas@cti.gr

ABSTRACT

The amount of data that exists on the Internet is enormous enough to distract users when trying to find useful information. In addition, the expansive use of small screen devices for browsing the World Wide Web generates huge problems when trying to find and read information. A solution to these problems is to personalize the web and try to reduce algorithmically the amount of text. Many text summarizers have been presented in order to reduce the valueless information that is presented to the users and many web sites, especially news portals, introduce personalization features for the users, though, still these techniques are not often used in combination in order to create more effective results. In this paper a mechanism for creating personalized summaries for the members of a news portal that reproduces articles collected from major portals is presented, together with evaluation both of the summarizer and the personalized summaries. The evaluators of the personalized summaries are members of the news portal. The personalized summary mechanism can also be utilized by users of small screen devices, for easier reading of less but inclusive information.

KEYWORDS

Text summarization, Personalizing the Web, Text Categorization, Sentence Weighting

1. INTRODUCTION

Information that exists on the World Wide Web and the users that have access to it or produce it have reached outrageous numbers. This state is not static, but a dynamic, continuously changing condition, that converts the Internet into a chaotic system. It is estimated that more than ten billion web pages exist at present, while the number of Internet users is uncountable. The consequence of the popularity of the Web as a global information system is that it is flooded with a large amount of data and information, and hence finding useful information on the Web is often a tedious and frustrating experience. The solution to finding information is search engines, but their main problem is that they search every corner of the Web and often the results, even to specific queries, are millions of pages.

We intend to focus on the needs of the Internet users who access news information from major or minor news portals. From a very brief search we found more than thirty major or minor portals that exist only in the USA. This means that if an internet user wants to find information regarding a specific topic he/she will have to search one by one, at least the major portals, and try to locate the news of his preference. A better solution is to access every site and search for a specific topic, if a search field exists in the portal. The problem becomes bigger for someone who would like to track a specific topic daily (or more than once a day).

Many well-known websites try to solve this problem by creating rss feeds or personalized micro-sites where a user can add his own interests and watch the most recent and popular issues on them. But still, the problem of filtering the information is present. Regarding the personalization issue, the attempts that have been made from the major search engines and portals include only the issue of viewing already categorized content according to the

user's interests. This means that the user is not included into the classification procedure. A problem that arises is: "even after the personalization, is the user satisfied with the result?" and "can we satisfy the user with transparent to him/her procedures?". Trying to access personalized portals and through personalized news, the user seems to put the focus on the title and on summaries of articles, if any. This means that it is important to have a representative summary for each article and even more, a representative personalized summary on each user. This is the challenge we are working on in this paper.

Recently, there have been many efforts towards the direction of text summarization together with the many forms it can take, eg. Web page summarization [6],[15], online encyclopedia summarization [17], etc. The work of text summarization starts of with the classic work of H.P. Luhn [9]. The approaches of this kind take into consideration the words in sentences. Some other techniques [10],[11], try to find special words and phrases in the text, others [12] compare patterns of relationships between sentences or take into consideration the length of the sentences or the word case [13]. Using statistics from the corpus itself is very common since it can provide with a similarity measure between the summary and the text itself.

While some summarization techniques try to extract the most important sentences, as far as a certain measure is concerned, others [14], [15] attempt to generate the summary directly using a knowledge-based representation of the content or a statistical model of the text. In [4] the authors explore the use of probabilistic models of term distribution in documents using the negative binomial (k-mixture model).

All of the aforementioned summarization techniques are roughly divided into four categories. The first category contains techniques that use some kind of heuristic approach towards the problem. Sentence rating or special weighting of sentences containing title words [10] belong to this category. The second category includes corpus-based methods [13] that frequently use the TF.IDF (term frequency – inverse document frequency) method. The third category includes methods that take into account the text structure. Lexical chains usage is a representative method of this class [16]. Finally there is a category that uses knowledge-rich approaches towards the problem. Summarization methods of this category are the most advanced but are of use only for particular domains. An effort for an online medical encyclopedia is presented in [17].

Recently, in [5] there is an effort to find the dynamic portions of a document and use this to produce good summaries based on the hypothesis that the higher the number of dynamic parts containing a term, the more important this term is for the summary. In [6], the writers try to adopt Web-page summarization to Web-page classification and improve the classification results using summarization methods. Using text categorization to produce good summaries is also faced in [7] where the authors use a self-organizing feature map (SOFM) which learns the salient features of each of the texts and assigns the text in a mnemonic position of the map. Latent semantic analysis [8] is also frequently used for extracting summaries. Natural Language Processing, while not always the best choice, is used frequently, for example in SUMMARIST [3]. These methods tend to operate at word level and miss concept-level generalizations. Marginal Relevance (MMR) holds the idea of balancing novelty and usefulness of terms and focuses on query-based summarization of a static collection of stories. In many of the techniques, the problem is faced as a classic IR problem and solved using precision-recall metrics.

In this paper we focus on the summarization mechanism of our system. More specifically, we describe the algorithmic procedure that leads to personalization of the summary of the articles, based on sentence weighting. Additionally, the sentence weighting procedure which leads to summarization (through selection of the most representative sentences) is influenced by the categorization procedure.

The rest of the paper is structured as follows. In section II we describe the architecture of our mechanism, giving more emphasis on the text summarization procedure. In section III we give some evaluation of the mechanism based on some experiments that we conducted. In section IV we conclude the results of our research and in section V there are our thoughts of possible future work on the field of text summarization and expansion of our mechanism.

2. ARCHITECTURE

The architecture of the system is based on standalone subsystems which collaborate in a sequential manner in order to produce the desired result. The main subsystems that the mechanism consists of are: the information retrieval system, the text pre-processing procedure, the text categorization and summarization mechanisms and finally the personalized portal in order to present the results to the end-user. An important architectural issue is the modularity of the mechanism which means that each subsystem can work as a standalone system without affecting any other procedure of the complete mechanism as it only requires communicating with the central database. In this section we will describe how are these features implemented through the architecture of the system. This paper is focused on the module of the text summarization, though, analysis of the categorization module and the personalization mechanism of the personalized portal is presented in order to cross-connect the features of our system.

2.1 Flow of Information

The flow of the information is represented by the main system's architecture as shown in Figure 1. Beginning from the public internet we retrieve information from major news portal web sites. The information retrieval (IR) mechanism is responsible for extracting the desired information. More specifically after the acquisition of the web page, the html tags must be removed, and the html file should be searched for useful text (UT). UT is specified as the text which represents both the title and the body of an article. Information about the fetching of the pages, the extraction of the useful text that is done by our mechanism and the categorization procedure can be found in [1] and [2].

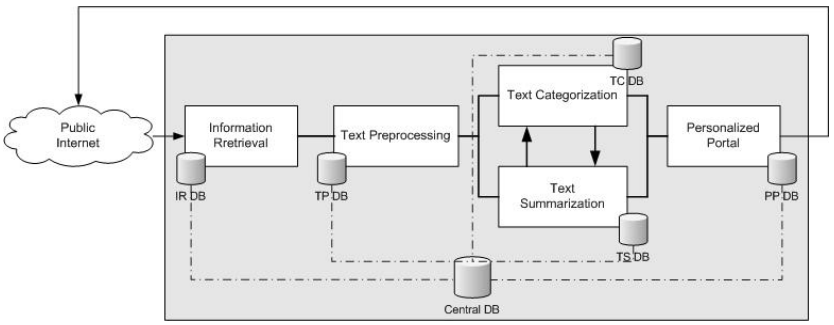

Fig. 1. The main system's architecture

The result of the information retrieval mechanism is tuples - in the local database storage system - which includes the id of the article, the title, the body and the date of fetching. Following is the text preprocessing step which receives as input XML files or XML structured information coming from the DB. Preprocessing algorithms such as punctuation removal, word size limiting, stopword removal and stemming are applied at this step producing XML output that contains: the extracted keywords, their frequency and their position in the text. This output is stored in the Central DB and can be forwarded to the text categorization and summarization subsystems which can respectively label and summarize the text. The communication between these subsystems (as described later) is fundamental for the improvement of both the categorization and the summarization results. The output (XML) of these subsystems is also stored in the DB and then forwarded to the personalized portal in order to be presented in a consistent manner to the end-user.

The modularity of the system is based on the global form of the input on each level. Even though the core of each subsystem is programmed in C++ language for optimal performance using the latest text manipulating libraries, each can be replaced by any other alternative accepting the same input and producing the same output. This means that each mechanism (apart from the IR mechanism which uses as input HTML files) receive as input XML [18] files with specific XML format specified by DTD [18] files and has as output XML files and Database records. Database Records are used for the specific procedure and mechanism only while each module of the system can work as a standalone mechanism.

3. ALGORITHM ANALYSIS

3.1 Summarizer Subsystem

The summarization procedure is based on heuristic methods. This means that the summary is not constructed "from scratch", but it consists of the most representative sentences, in order to find which, a ranking system is deployed. This implies that every sentence should be given a score, and higher ranked sentences are included in the resulting summary. In the proposed mechanism, 6 distinct factors are used in order to rank each sentence of the text: (a) the keyword's frequency (how many times a keyword appears in a sentence), (b) the keyword's appearance in the title, (c) the percentage of keywords in a sentence, (d) the percentage of keywords in the text, (e) the keyword's ability to represent a category and finally (f) the keyword's ability to represent the choices and needs of a unique user or a category of users with the same profile. According to the first two [(a) and (b)] we produce the first and basic equation to begin with a generic scoring of the sentences:

$$Si = \sum wk,i(k1+k2) \quad (1)$$

Where $w_{k,i}$ is the weight (relative frequency) of the k^{th} keyword of sentence i, k1 is a constant that represents the impact of factor (a) and k2 is a constant that represents the impact of factor (b) to the summarization procedure.

In order to normalize the values that derive from equation 1 we propose the use of the factors (c) and (d). The normalization is needed as the big in length sentences tend to score higher than the small in length ones. The first represents the percentage of keywords in a sentence while the second represents the percentage of keywords in the text. More specifically if three keywords are extracted from a sentence which consists of five keywords and the number of extracted keywords is twenty five then factor (c) equals three of five (=3/5) and factor (d) equals three of twenty five (=3/25).

The following example demonstrates some potential problematic situations that are prevented using the aforementioned normalization. Assume that a text has many small sentences and one very large. Additionally, the large sentence consists of 20 keywords and the extracted (useful) are 5, while a small sentence is very representative of the text consists of 4 keywords and all of them are extracted as useful. The total number of useful keywords that are extracted is 30. The big sentence is more likely to score higher according to the aforementioned equation as its length "helps" it to have more keywords. The two factors "normalize" this possible unfairness. The big sentence will have 5/20 and 5/30 respectively, while the second sentence will have 4/4 and 4/30 as c and d factors respectively. In this way, the small in length sentence will be treated as more important than the big sentence. The normalization is applied directly to equation (1) and S'i = SiN, where N is the normalization factor and equals the product of c and d factors.

The factors (e), keywords' ability to represent a category, and (f), keywords' ability to represent the choices of a unique user, are presented thoroughly in the following chapters, as their influence to the procedure is important and promotes the summarization system into a fully personalized mechanism.

3.2 Categorization Subsystem

The categorization subsystem is based on the cosine similarity measure, dot products and term weighing calculations. More specifically, the system is initialized with a training set of articles collected from major news portals. The articles are pre-categorized – by humans – and are presented categorized into the news portals. Our training set consists of these pre-categorized articles. The categorization module receives as input the extract of the pre-processing mechanism. This is (a) an XML file containing stemmed keywords, their absolute frequency and their relative frequency in the article and (b) the XML file containing the article (information about the article includes id, type, title and body). After the initialization of the training set, the categorization module creates lists of keywords that are representative of a unique category, consisting of keywords with high frequency in a specific category and small or zero frequency for the other categories. The creation of the lists is helpful for categorizing newly arriving articles but we can prove that can be helpful for summarization also.

As the summarization procedure of our module is based on the selection of the most representative sentences which are selected by weighting them appropriately, the categorization outcomes can be helpful for adjusting more effectively the weighting of the sentences. Common sense implies that a keyword that has very high frequency for a specific category should give more weight to the sentence that it appears into while a keyword that has small or zero frequency for a category, could add less to the weight of a sentence. Moreover a keyword that is included into the extracted keywords of an article that is representative of another category, than the one that the article is, would give negative weight to the sentence. The following equation is used for calculating the impact of the categorization into the summarization procedure:

$$k_3 = \begin{cases} A \cdot cw_i & \text{where A>1 and cw the positive category weight} \\ -A \cdot cw_i & \text{where A>1 and cw the negative category weigh} \\ 1 & \text{for neutral or not ranked by the system keyword s or if A=0} \end{cases} \quad \textbf{(2)}$$

Parameter A must be greater than 1 and it is used in order to add a weight for the k3 variable. If we want the summarization procedure to be based mainly on k3, then high values for A are used, but if the summarization should be equally based on all the "k" variables, then A should not be greater than the values that are assigned to k1 and k2. The parameter cw depicts the relative frequency of the keyword in the category. The relative frequency of a keyword in a category can provide us with evidence about how important is the keyword for the category.

With the use of equation 2, equation 1 is formed as shown below:

$$S_i^{'} = \sum w_{k,i}(k_1 + k_2)k_3 \quad \textbf{(3)}$$

3.3 User's Role in Summarization

The personalization procedure of the portal that is supported as a medium of communication between all the procedures with the users can be used in order to personalize the summarization on each user. We believe that the user should be able to see a summarization of the articles that match his/her criteria and not a generic summarization that derives from a simple algorithmic procedure.

According to the algorithmic procedures of the personalized portal, the system creates lists of keywords for each user that represent his selection while browsing the news portal. More specifically the keywords form two types of lists: a "positive" list with keywords that seem to suit the character of the user or a group of users and a "negative" list with keywords that are out of interest for a user or a group of users. These lists derive from the selections of the user (which articles the user selected to read and which did not, in which articles the user spends more time to read and in which does not, etc.). Our intention is to rank higher the sentences which include "positive" keywords and to lessen the rank of sentences that include "negative" keywords for the user. In this scope we add another "k" variable, k4, which will act as the personalization factor.

The personalization variable is used like the variable that derives from categorization, and is given by the following equation:

$$k_4 = \begin{cases} B \cdot uw_i & \text{where B>1 and uw the positive user's weight} \\ -B \cdot uw_i & \text{where B>1 and uw the negative user's weight} \\ 1 & \text{for neutral or not ranked by the user keywords or if B=0} \end{cases} \quad (4)$$

The parameter uw depicts the relative frequency of the keyword for the user. The relative frequency of a keyword in a category can provide us with evidence about how important is the keyword for the user.

This variable is added as a product to equation 3 which is formed as follows:

$$S_i^{'} = \sum w_{k,i}(k_1 + k_2)k_3 k_4 \quad (5)$$

The variables A and B in equations (2) and (4) respectively are used in combination to each other. If we do not intend to use the categorization factor (k3), we may set A=0, and accordingly if we do not intend to use the personalization factor (k4) then we may set the B variable to 0. If we want to focalize mainly on the personalization factor and less on the categorization then we can set B=2 and A=1. This means that k_4 factor will have twice impact than k_3. The following table shows the impact of (e) and (f) factors according to values of A and B.

Table 1. Impact of A and B to sentence weighting

A	B	Result
0	0	Personalization and Categorization factors not computed to the result
1,2	1,8	We are focusing mainly on the Personalization factor rather than the Categorization.
1	2	Personalization factor has twice the impact of the categorization factor to the result
1	1	The same impact for personalization and categorization factor

As observed from equation (5) some "special" occasions may occur from the negations that are introduced by the variables k3 and k4. The following table shows the reaction of the algorithm to the four different states.

Table 2. Reaction of Summarization Algorithm to Variables k3 and k4

Variable k3	Variable k4	Result
Positive	Positive	Positive
Positive	Negative	Negative
Negative	Positive	Positive (k3 not computed to the result)
Negative	Negative	Negative

One "special" occasion occurs when the categorization variable is negative and the personalization variable is positive. In this occasion we assume that the user, despite the fact that the keyword is not concerned as a representative of the category, has selected the specific keyword as a representative of his interests and thus the personalization variable overrides the categorization variable. Additionally when both variables are negative the result remains negative, as the negations in our situation mean even lower score for the sentence.

4. EVALUATION

Each of the aforementioned equations for sentence weighting was tested on some pre-summarized (by humans) texts. The results of our mechanism seem to be adequate compared to already existing mechanisms. Our main aim is to focalize on the personalized summary and thus the summaries that derive from equations 1 and 2 may be less effective than already existing algorithms. The personalization procedure into the summary cannot be evaluated by any prototype human created summary, despite the fact that any human created summary implies the subjective human factor. The only evaluator of the system is the end-user that receives the summaries. We tested our summarization algorithm compared to MEAD summarizer algorithm and the summarizer that is used by Microsoft's Word. The personalization summaries are ranked by five test users who use the personalized portal.

4.1 Evaluating the Automatic Summarization Mechanism

In order to ensure that the procedure before embedding the personalization factor produces adequate results for summaries we evaluated our mechanism in comparison to results from Microsoft Word's summarizer. The results are compared to extracts from MEAD summarizer[1] onto 30 articles from major USA and UK portals. The metrics that were used in order to calculate the results were precision and recall.

Table 3. Comparison of Summarization Algorithm to MS Word summarizer (Results compared from outcomes of the MEAD Summarizer)

	MS Word		Proposed Mechanism	
	Precision	Recall	Precision	Recall
Article 1	0,33	0,12	0,66	0,75
Article 2	0,12	0,25	0,75	0,66
Article 3	0,25	0,12	0,5	0,66
Article 4	0,25	0,12	0,75	0,5
Article 5	0,33	0,5	0,66	1
Article 6	0,33	0,25	0,66	0,75
Article 7	0,25	0,33	0,75	0,66

From the results derives that the summarization mechanism produces adequate results compared to tests that have been done with MEAD summarizer and obviously better results than the ones extracted by MS Word. By adding the categorization factor to the summarization mechanism we manage to get slightly better results. We observe an overall increase of about ten percent to the previous results concerning the metrics of precision and recall. The difference derives from the categorization procedure and more specifically from the addition of k_3 factor to the summarization equation. This factor enables the higher ranking of sentences which include keywords that are representative of the category that the article belongs to. If an article does not include many keywords from the category that it belongs, no changes occur. In this occasion, it is remarkable to note that after some time, when more keywords are inserted in the system, when someone tries to access the summary of the specific article it will be updated and the metrics of precision and recall will be measured higher than the first time of summarization. In the following table the metrics of precision and recall are presented for a specific article and how they change when new articles are categorized and more representative keywords for the category are inserted into the mechanism. The articles "arrive" in our system every four to six hours as the major news portals update very often their data.

From the statistics shown in Table 4, derives that the system is not static, but is able to dynamically change and update the summaries that are extracted. Moreover it is expected that after the publishing of an important news event, many articles on this issue will occur and will be published. This means for example in the occasion presented in Table 4, that in the next 103 articles of the category that are captured by the mechanism within the next 78 hours, at least one of them will be similar to the first article either as an update or as a complement. This derives also from the functionality of the modern news portals which include the "related articles" feature.

Table 4. Changes in precision and recall for the summary of article 1 after the addition of more representative keywords for the category that the article belongs.

Time (after arrival)	Articles added to category (sum)	Proposed Mechanism

1 http://www.summarization.com/mead/ -Mead summarization mechanism (Last Accessed: December 2006)

		Precision	Recall
10 min	0	0,5	0,66
8 hours	8	0,5	0,66
24 hours	31	0,66	0,5
48 hours	59	0,66	0,66
78 hours	103	0,75	0,8

4.2 Evaluating the Personalized Summarization Mechanism

The evaluation of a dynamically created personalized summary is not a procedure that can be completed comparatively. The measure that is used in order to evaluate the extracted personalized summaries is the relation between the summary and the article observed by the users of the mechanism. The procedure that was used in order to evaluate the results of the algorithmic procedure was: (a) provide the users with the full text of the article, (b) provide the users with both of the summaries created by using equation (3) and equation (5) and (c) let them choose which summary they believe as more representative of what they read. The reverse procedure was also tested, which means first provide the users with both of the summaries, then the article and finally let them decide which summary they believe represents the most suitable for the full article they read. In both occasions the answers were the same.

The outcomes of the user's opinions can be separated into three groups: (a) new users of the system, (b) old users of the system but with little action (which means few data for personalization) and (c) advanced users of the system with high daily action (which means a lot of data for personalization). According to these categories, three different states were observed. The novice users noticed that the summaries were identical, which is a logical observation, as the system does not have enough information for the personalization procedure and thus, the sentence weighting for summarization is not affected by factor k_4 (used for personalizing the summary). The users of the second group selected in more than 80% of the occasions the summary extracted from equation 3 (without the personalization factor). This was expected as the dynamically created profile of such users (with low participation) was not complete and it included many keywords that were of low importance both for the article and its category. The most important results derive from the users of the third group. This group of users is considered to be advanced for the system with almost stable profiles after long time of system usage. The stability and completeness of the profile empowers the personalization procedure of the summaries. This group of users selected in more than 90% of the occasions the personalized summary as the most representative of the article according to their opinion and only 3% of the summaries were reported to be identical. It is important to note that most of the remaining 7% of the articles were reported to the categorization procedure of the mechanism as: "belonging to a specific category but with weak connection". This means that these were articles that added to the specific category with the "note" that the system had not managed to enclose them into a specific category but the category that are inserted in is the most likely to hold these articles.

5. CONCLUSION AND FUTURE WORK

We have presented an algorithmic procedure that can be used in order to produce effectively personalized summaries. In an era of chaotic conditions in the web, personalization cannot be considered as a panacea but it can be very useful and helpful for advanced and novice users. In this scope we proposed a mechanism that is able to dynamically create summaries for texts or branches of text and users are able to view a summary that is fully personalized in their characteristic of browsing. This requires training for the mechanism which is based upon the selections and rejections of the users in the area of a web page and the time that he/she remained looking a specific web page.

The system that was described is generic and it is designed and constructed as a module. This implies that it can be embedded into software and mechanisms in order to extend them in order to support summarization procedures. Our main aim is to efficiently produce summaries for RSS readers and small screen devices. The last remark seems to be interesting and important as the usage of small screen devices for daily activities has reached a quite big number nowadays.

In the future versions of the current mechanism we will try to include more complex algorithms for the summarization procedure in order to make it even more accurate and efficient, though, the results received are more than encouraging. Additionally, what was observed was that, despite the fact that balancing factors were used, still, the greater in length sentences were gaining more weight than the shorter ones. Accordingly this implies that some short but inclusive sentence may be omitted. Furthermore, in order to globalize the system more lexica should be included in order to make the preprocessing and summarization mechanism available for more languages

than English. Finally, a crucial part of the mechanism is the implementation of the procedures for small screen devices. The ultimate goal is to use the mechanism in order to make PDAs, and generally small screen devices, more user-friendly and available for daily tasks like reading mails, reading RSS feeds, and understanding the meaning of large amounts of text through a personalized summary. This mechanism could provide small branches of text to the users and let them choose easier which articles they are really interested in. Also users could select the length of the summary they desire defining either a maximum of character or a maximum of words.

REFERENCES

1. C. Bouras, V. Poulopoulos, A. Thanou. “Creating a Polite Adaptive and Selective Incremental Crawler”, IADIS International Conference WWW/INTERNET 2005, Lisbon, Portugal, Volume I, 19 - 22 October 2005, pp. 307 - 314
2. C. Bouras, C. Dimitriou, V. Poulopoulos, V. Tsogkas. “The importance of the difference in text types to keyword extraction: Evaluating a mechanism”, 7th International Conference on Internet Computing 2006 (ICOMP 2006), Las Vegas, Nevada, USA, , 26 - 29 June 2006, pp. 43 - 49
3. Eduard Hovy and ChinYew Lin. “Automated Text Summarization in SUMMARIST”, Workshop on held at Baltimore, Maryland: October 13-15, 1998
4. M. Saravanan, P.C. Reghu Raj and S. Raman. “Summarization and Categorization of Text Data in High-Level Data Cleaning For Information Retrieval”, Proc. First Intl. Workshop on Data Cleaning and Preprocessing, pp. 119-130, ICDM 2002, Maebashi, Japan (Dec 9-12).
5. Adam Jatowt and Mitsuru Ishizuka. “Web Page Summarization Using Dynamic Content”, Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters table of contents New York, NY, USA, 2004.
6. Dou Shen, Zheng Chen, Qiang Yang, Hua-Yun Zeng, Benyu Zhang, Yuchan Lu and Wei-Ying Ma. “Web-page Classification though Summarization”, Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval,Sheffield, South Yorkshire, UK. July 25-29, 2004.
7. Khurshid Ahmad, Bogdan Vrusias and Paulo C F de Oliveira. “Summary Evaluation and Text categorization”, Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, Pages: 443 - 444, 2003.
8. Josef Steinberger and Karel Jezek. “Using Latent Semantic Analysis in Text Summarization and Summary Evaluation”. Proceedings of the 5th International Conference on Information Systems Implementation and Modelling, pp. 93-100, MARQ Ostrava, April 2004.
9. H. Luhn. “The automatic creation of literature abstracts”, Presented at IRE National Convention, New York, March 24, 1958.
10. H. P. Edmundson. “New methods in automatic extracting”. Journal of the Association for Computing Machinery, Volume 16 , Issue 2, April 1969.
11. J. Pollock and A. Zamora. “Automatic abstracting research at chemical abstracts service”, Presented before the Division of Chemical Information, 169th National Meeting of the American Chemical Society, Philadelphia, Pa., April 8, 1975.
12. G. Salton, A. Singhal, M. Mitra and C. Buckley. “Automatic text structuring and summarization”. In I. Mani, M. Maybury (Eds.), advances in automatic text summarization. MIT Press, 1999.
13. J. Kupiec, J. Pedersen and F. Chen. “A trainable document summarizer”, Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 68-73, Seattle, Washington, United States 1995.
14. M. Witbrock and V. Mittal. Ulta-summarization : “A statistical approach to generating highly condensed non-extractive summaries”. In Proceedings of SIGIR, pages 315-316, 1999.
15. A. Berker and V. Mittal. “OCELOT: a system for summarizing web pages”, Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, Athens, Greece Pages: 144 – 151, 2000.
16. R. Barzilay, M. Elhadad: “Using Lexical Chains for Text Summarization”, Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS’97), ACL Madrid, Spain 1997.
17. M. Fiszman, T.C. Rindflesch, H. Kilicoglu: “Summarization of an Online Medical Encyclopedia”, MEDINFO 2004, M. Fieschi et al. (Eds), Amsterdam: IOS Press 2004 IMIA.
18. XML Specification. http://www.w3.org/XML/ (Last Accessed:8 December 2006)
19. DTD Information. http://en.wikipedia.org/wiki/Document_Type_Definition (Last Accessed:8 December 2006)