ISSN:0975-9646

Efficient Summarization of Text Data Based on Categorized Keywords Ranking

Sailaja Madhu
M.Tech Student
Dept of CSE, AIET

B.Ramesh Babu
Asst. Professor,
Dept of CSE, AIET

Y.Ramesh kumar
Asst. Professor,
Dept of CSE, AIET

Abstract: The information that exists on the World Wide Web is enormous enough in order to distract the users when trying to find useful information. In order to overcome the large amounts of data many personalization and summarization mechanisms have been presented. In this paper we propose a mechanism that applies summarization techniques on articles extracted from the web, based on the categorization procedure (also applied on the same articles). Through extensive experiments we proved that the summarization procedure can affect the categorization mechanism and vice versa. This means that when the results of the summarization mechanism seem to be weak, then the categorization can be used in order to provide a more efficient summary and on the other hand when the categorization procedure becomes too overloaded, the summarized articles can be used in order to categorize the article more efficiently. Moreover this paper introduces that the combination of summarization and categorization can lead to more efficient results not only for both mechanisms but for a personalized portal also. Finally, we propose a complete mechanism that can be used in order to provide the users with helpful tools in order to locate more easily the information they need.

Keywords: summarization algorithms, categorization procedure, data reprocessing, efficient summarization

1. INTRODUCTION

NOWADAYS the internet users have reached outrageous numbers. Additionally, the web pages together with the information that exists in each page create a chaotic condition for the World Wide Web. This condition is not a static, stable condition but a dynamic continuously changing state that feeds daily the entropy of this chaotic system. Many attempts have been made in order to count the pages of the internet and the estimation of more than ten billion web pages existing seems to be conservative. Moreover, each of these pages include from no information at all to thousands of pages full of information, multimedia and articles. The problem that arises from the aforementioned condition is when searching for useful information.

Let us focalize this searching on news and articles from different major news portals. From a brief search we have located more than thirty major and minor news portals existing in America that include worldwide news (concerning probably all the internet users as they are not just local news). This means that whenever a user needs to be informed about an issue he has to search all the web sites on by one. This is what actually happens nowadays from the internet users. This could be considered as a problem of locating useful information among all the news portals especially when a user wants to track a specific topic on a daily basis.

They are two critical methods for solving part of aforementioned problem.

i. Text Searching
ii. Summarization

Text Searching:

The search engines play the role of the filter for the information while text summarizers are utilized as information spotters to help users spot a final set of desired documents. Recently, there have been many efforts towards the direction of text summarization together with the many forms it can take, e.g. Web page summarization, online encyclopedia summarization, etc, this classic work is based on analysis of words and sentences. Some techniques introduce the searching of special words or phrases in the text while others are based on patterns of relationship between sentences or take into consideration the length of the sentences. More advanced techniques do not use elements from the set of document on which summarization is applied itself but try to generate the text directly using a knowledge-based representation of the content or a statistical model of the text.

Summarization:

In general, the summarization techniques can be divided into the aforementioned four major categories:
(a) Heuristics,
(b) TF-IDF,
(c) knowledge-based and
(d) Statistical models.

Another categorization of the summarization techniques is introduced by Mani and Hahn concerning the extent of involvement of domain-knowledge. The two categories include methods that are knowledge-poor and knowledge-rich methods. The first category includes methods that do not take into account any knowledge that has to do with the domain and are easily applied to any domain while knowledge-rich techniques assume that knowing or understanding the meaning of the text will lead to better results. According to this ontology heuristics and TF-IDF are considered to be knowledge-poor while knowledge-based and statistical models are knowledge-rich techniques

Recently, in 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, the writers try to

adopt Web-page summarization to Web-page classification and improve the classification results using summarization methods.

The current mehods having the following advantages

i. The current system introduces that the combination of summarization and categorization.
ii. This can lead to more efficient results.
iii. This complete mechanism that can be used in order to provide the users with helpful tools in order to locate more easily the information they need.
iv. This combination mechanism is mostly used in the personalized portals.

ARCHITECTURE

The mechanism consists of a series of subsystems that produce the desired result. The collaboration between the distributed systems is based on the open standards for input and output that are supported by each part of the system and by communication with a centralized database. Figure 1 depicts the architecture of the complete mechanism.

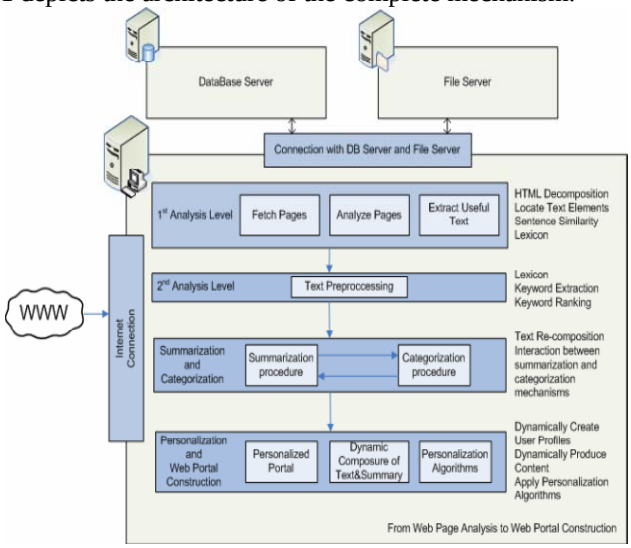


Figure1:architecture of complete mechanism

The procedure of the mechanism, as depicted in figure , is:
(a) Capture pages from the www and extract the categories of articles.
(b) Parse the extracted text,
(c) Summarize and categorize the text and
(d) Present the personalized results to the end user.

In order to capture the pages, a simple crawler is used. The addresses that are used as input to the crawler are extracted from RSS feeds. The RSS feeds point directly to pages where articles exist.

The crawler stores the html pages without any other element of the web page (images, CSS, JavaScript are omitted). By storing only the html page, the database is filled with pages that are ready for input to the 1st level of analysis. During the 1st analysis level our system isolates the "useful text" from the html page. The useful text can be defined as the title and the main body of the article. Information about this procedure can be found. The second analysis level receives as input XML files that include the title and body of articles. Its main scope is to apply on this text pre-processing algorithms and provide as output keywords, their location into the text and their frequency of appearance in the text. These results are necessary in order to proceed to the third analysis level. Information about our preprocessing mechanism can be found in . The core of our mechanism is located in the third analysis level, where the summarization and categorization sub-systems are located. Their main scope is to characterize the article with a label (category) and produce a summary of it. All these results are then presented back to the end users of our personalized portal. The role of the portal is to feed each user only with articles that the user "wants" to face according to his dynamically created profile.

2. METHODOLOGY

ALGORITHMIC ANALYSIS

In order to analyze how each algorithm is applied on the texts we will present the algorithm of execution of each step. We start by trying to categorize the article. In order to label (categorize) the article, we create a list of the representative keywords (stemmed) of the text together with their frequency (Table 1).

TABLE I
KEYWORDS WITH FREQUENCIES

ID	Keyword	Frequency[a]
1	Intern	19
2	Compan	17
3	Fire	12
4	Lead	12
5	Integr	11
6	Popular	10
...		
29	Busines	1

Table1: The keywords are ordered in descending orderof their frequencies

Next, we create identical lists for all the categories that we own. These lists consist of the same keywords followed by the frequency of them into the category. We examine the cosine similarity of these lists in order to determine the category of the text (Table 2).

TABLE II
SIMILARITY BETWEEN TEXT AND CATEGORY

Keyword	Frequency[a]
business	0,742862
entertainment	0,449297
health	0,532352
politics	0,418447
Integr	0,596509
science	0,526925
sports	0,642862

From the outcomes we can have three different results:
(a) The text is very representative of a category and can be added to the dynamically changing training set, (b) The text can be labeled as it is very similar to a category compared to others

(c) The text cannot be labeled clearly.

If the text cannot be labeled clearly then we forward it to the summarization mechanism and check if the summarized text is able to be labeled. A text is supposed to be labeled whenever the cosine similarity is over a threshold and additionally the difference between the cosine similarity of the higher category and the others is more than a threshold. This will be explained thoroughly in the next chapter. Finally, if the cosine similarity between the text and the representative category is very high and the difference between the similarities of the other categories is enormous, then the text is added to the dynamically changing training set. The aforementioned procedure is expressed in figure 2.

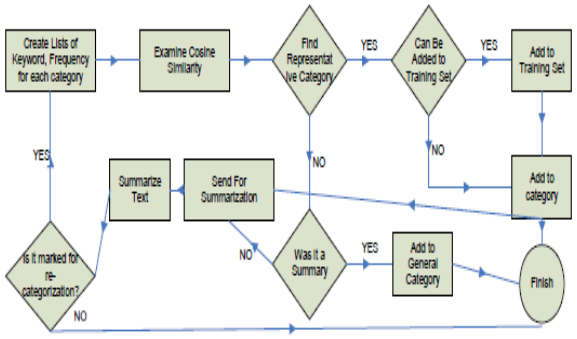


Figure 2: the block diagram of the system's procedures

SUMMARIZATION

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. This implies that every sentence should be given a score which leads to the construction of the summary. In the proposed mechanism, 5 distinct factors are used in order to create the summary and achieve the interaction with the categorization mechanism:

(a) the keywords' frequency (how many times a keyword appears in a sentence),

(b) the keywords' appearance in the title, and finally

(c) The keywords' ability to represent a category which is the factor that the interaction is based. According to the first two [(a) and (b)] we produce the first and basic equation to begin with a generic scoring of the sentences:

$$S_i = \sum w_{k,i}(k_1 + k_2) \qquad (1)$$

Where w, k1 is the frequency of the 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.

Through experimental procedure we have resulted in values for k1 and k2. k1 derives from the following equation

$$k_1 = 1 + 0{,}1x \qquad (2)$$

where x is the times that the keyword is found in the title. Accordingly k2 derives from the following equation

$$k_2 = 1 + 1{,}2y \qquad (3)$$

Where y is the possibility that the keyword is found n times in the sentence. Assuming a sentence with length m (m keywords), a text with length t the possibility of finding n times a specific keyword in a sentence is

$$y = \frac{n}{t}\frac{m}{t} = \frac{nm}{t^2}$$

CATEGORIZATION

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. Equation (5) is used for calculating the impact of the categorization

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 height 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:

3. EXPERIMENTAL PROCEDURE

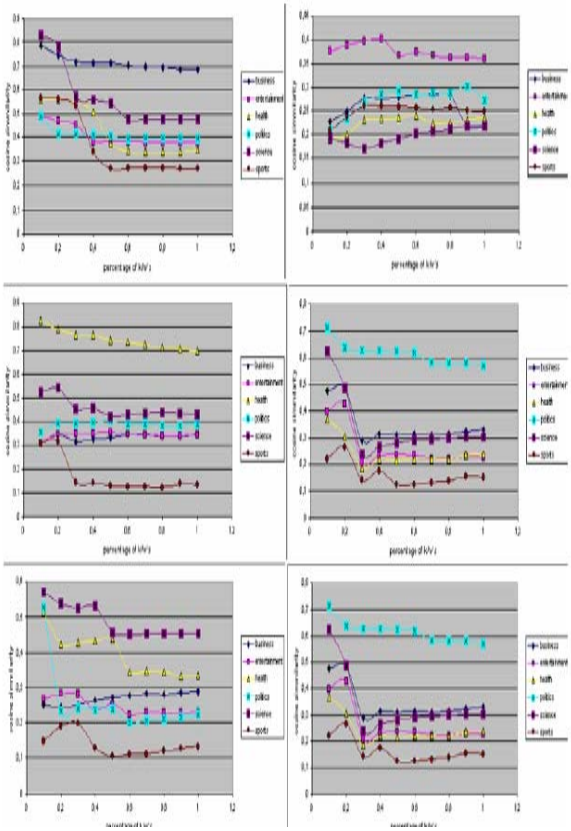


Figure 3: Cosine similarity of texts compared to categories. Training set is constructed with 50% of the keywords kept (pre-processing procedure).

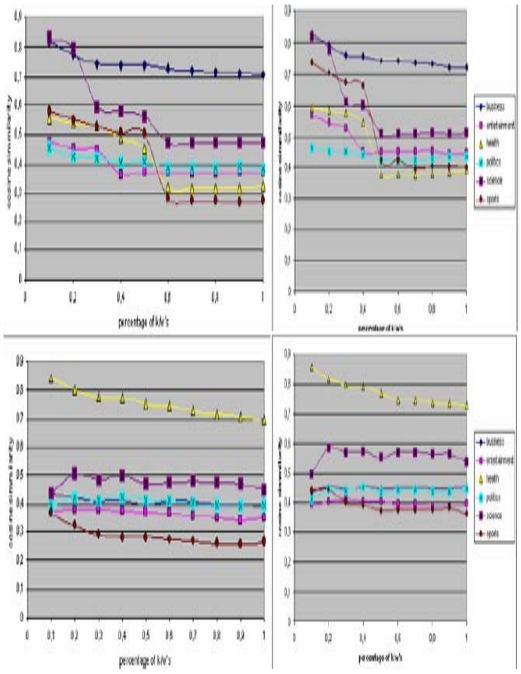
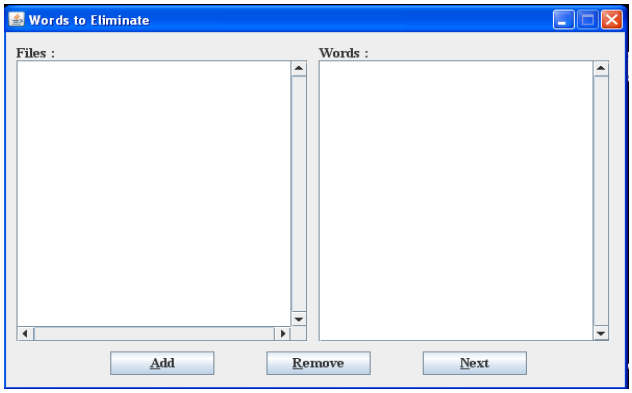


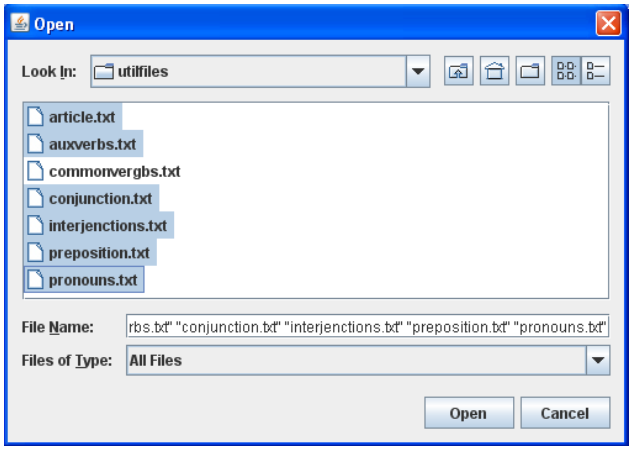
Figure 4: The first column depicts the cosine similarity measured by utilizing the 50% of the keywords from the training set and the second column is the same cosine similarity measured by utilizing the 100% of the keywords from the training set.

4. DATA ANALYSIS

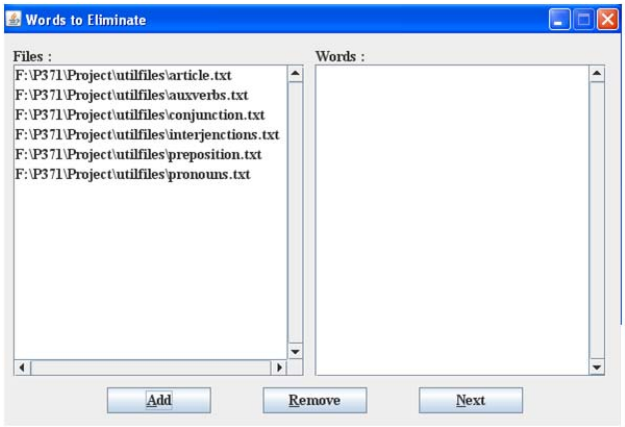
ADDING THE WORDS TO BE ELIMINATED



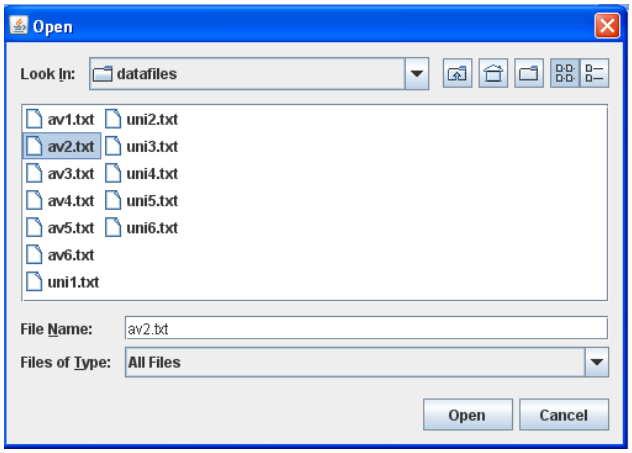
SELECTING THE WORDS TO BE ELIMINATED



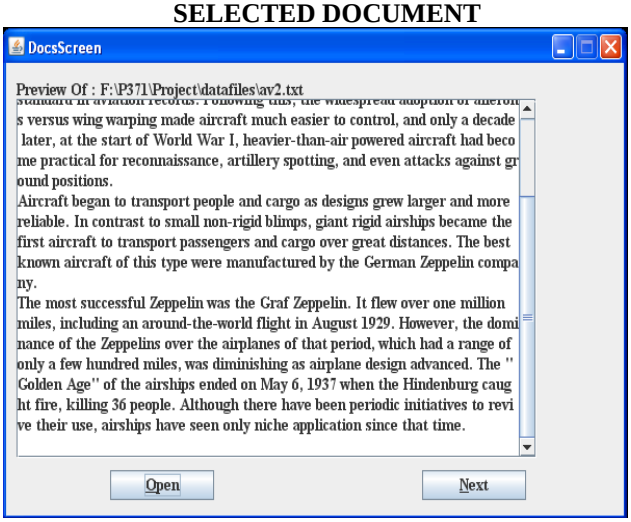
WORDS TO BE ELIMINATED



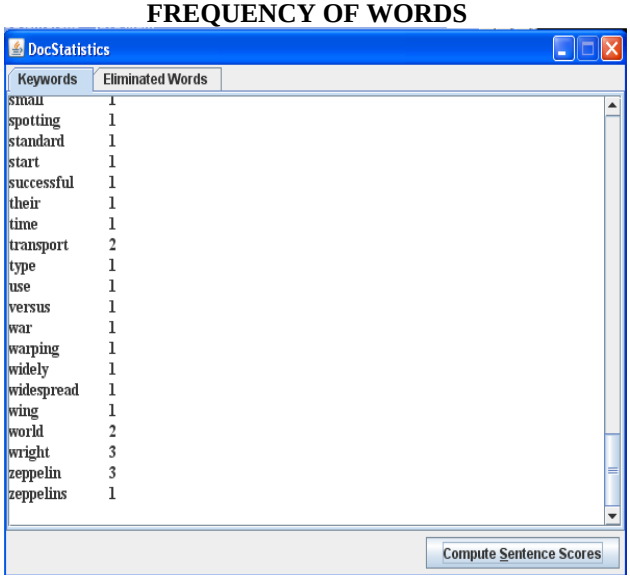
CHOOSING THE DOCUMENT



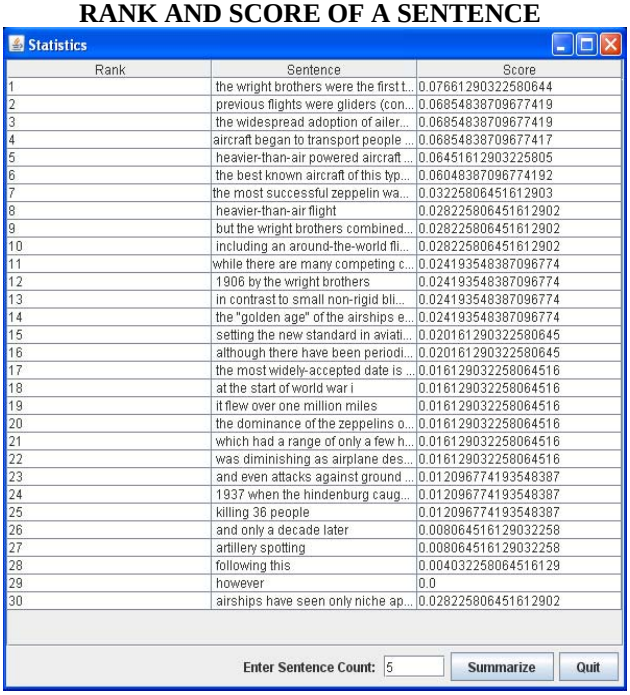
SELECTED DOCUMENT



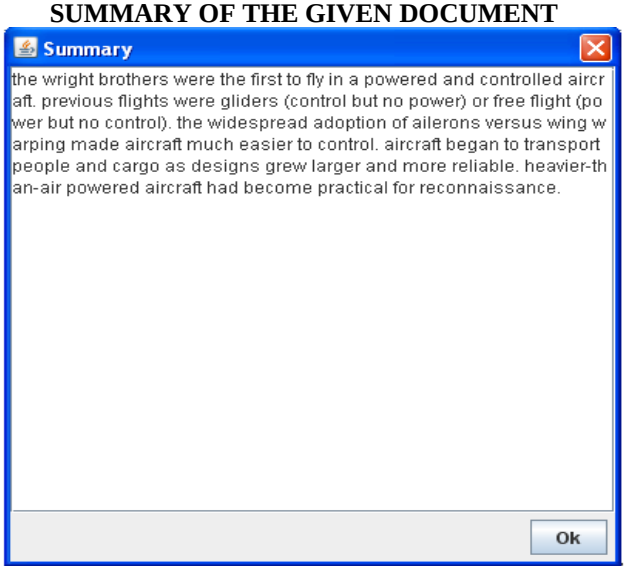
FREQUENCY OF WORDS



RANK AND SCORE OF A SENTENCE



SUMMARY OF THE GIVEN DOCUMENT



5.CONCLUSION

In this we have presented a mechanism that its main scope is to combine summarization and categorization techniques in order to produce more efficient results for both the aforementioned mechanisms. The ultimate scope of the mechanism is to apply real time, efficient summarization and categorization which is proved to be achieved through the interaction of these subsystems. As a major problem of today's Internet and more specifically of today's news and articles streaming is the burst mode that they are created in the Web our intention is to collect as many of them for the users, refine them and present them back in a more humanistic manner. Our paper focalized on the core of the mechanism that we are creating which is the categorization and the summarization sub-systems.

We have proved that by using the outcomes of categorization we can achieve better results on summarization and vice versa. The algorithms used for the summarization procedure are based on heuristics while the algorithm used for categorization is cosine similarity. The labeling of the articles achieves over 95% accuracy which is: achieving to categorize correctly almost all the articles into the prototype categories, while the results from the summarization mechanism are comparable to human created summaries. A major advantage of the system is that it manages to complete the whole procedure – from the fetching of the pages to the regeneration of the article to our portal – in less than 20 seconds per article. This means that the system is able to achieve real-time regeneration of the articles.

For the future versions of the core mechanism we will try to add a more complex algorithm for the creation of the summaries Another factor that is tested lately for our system is the personalization factor. We are intending to include the end user even to the categorization procedure by using its profile. Finally, as the core mechanism described is only a part of the system we should be aware of the results from the sub-systems that are executed prior to the core mechanism in order to obtain "clearer" data for the summaries and the categorization procedure.

REFERENCES

1. C. Bouras, G. Kounenis, I. Misedakis, V. Poulopoulos. “A Web Clipping Services Information Extraction Mechanism”,
2. C. Bouras, C. Dimitriou, V. Poulopoulos, V. Tsogkas. “The importance of the difference in text types to keyword extraction: Evaluating a mechanism”,.
3. Eduard Hovy and Chin Yew Lin. “Automated Text Summarization in SUMMARIST”,
4. M. Saravanan, P.C. Raghu Raj and S. Raman. “Summarization and Categorization of Text Data in High-Level Data Cleaning For Information Retrieval”,
5. Adam Jatowt and Mitsuru Ishizuka. “Web Page Summarization Using Dynamic Content”
6. Dou Shen, Zheng Chen, Qiang Yang, Hua-Yun Zeng, Benyu Zhang, Yuchan Lu and Wei-Ying Ma. “Web-page Classification though Summarization”,
7. Khurshid Ahmad, Bogdan Vrusias and Paulo C F de Oliveira. “Summary Evaluation and Text categorization”,.
8. Josef Steinberger and Karel Jezek. “Using Latent Semantic Analysis in Text Summarization and Summary Evaluation.
9. H. Luhn. “The automatic creation of literature abstracts”,
10. H. P. Edmundson. “New methods in automatic extracting”.
11. J. Pollock and A. Zamora. “Automatic abstracting research at chemical abstracts service”