Efficient Summarization Based On Categorized Keywords

Christos Bouras | Vassilis Poulopoulos | Vassilis Tsogkas

Associate Professor | Postgraduate Student | Undergraduate Student

Research Academic Computer Technology Institute, Riga Feraiou 61, 26221, Patras, Greece and
Computer Engineering and Informatics Department, University of Patras, 26500, Rion, Greece

bouras@cti.gr | poulop@ceid.upatras.gr | tsogkas@ceid.upatras.gr
+30 2610 960375 | +30 2610 996954 | +30 2610 996954

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

I. 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 that 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 (s)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.

Text searching and summarization are two critical methods for resolving part of the aforementioned problem. 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 [18]. 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 procedure of creating efficient, automatic text summaries begins from the late 1950s with an analytic approach from H.P. Luhn [9]. This classic work is based on analysis of words and sentences. Some techniques [10, 11] 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 [12, 13]. More advanced techniques do not use elements from the “corpus” (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 [14, 15]. Even probabilistic models of term distribution in the documents are researched in order to create summaries of corpuses [4].

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 [19] 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 [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 writers 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. NLP, while not always the best choice, is used frequently, for example the system 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 interaction of the summarization and categorization mechanisms of our system. More specifically, we describe the algorithmic procedure that leads to better results on each mechanism with the supporting of the other. We started from a training set of documents in order to create basic categories of articles. Then we used a set of articles on a daily basis as an input for the mechanism and applied to them summarization and categorization algorithms. During this procedure we tried to estimate how the results of the summarization could affect the categorization procedure and vice versa. Additionally, we found a limit for each of the procedures that produces the most efficient result for both mechanisms. According to the distinction of knowledge-poor and knowledge rich categories for the summarization techniques, our approach could be characterized as knowledge-poor because the basic algorithm for summarization is based on heuristics. Though, the interaction between the categorization and summarization modules enables the summarization to obtain some kind of “knowledge” about the domain of the keywords. This implies that the mechanism introduces an algorithm for a new category of summaries that lies between the knowledge-poor and knowledge-rich categories.

The remaining of the paper is structured as follows. In the next chapter we present the general architecture of our mechanism. In chapter 3 we provide the algorithmic analysis of the system, and in the next chapter we present the experimental results of our wok. Finally, we conclude with some remarks and future work.

II. 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.

The procedure of the mechanism, as depicted in figure 1, is: (a) capture pages from the www and extract the useful text, (b) parse the extracted text, (c) summarize and

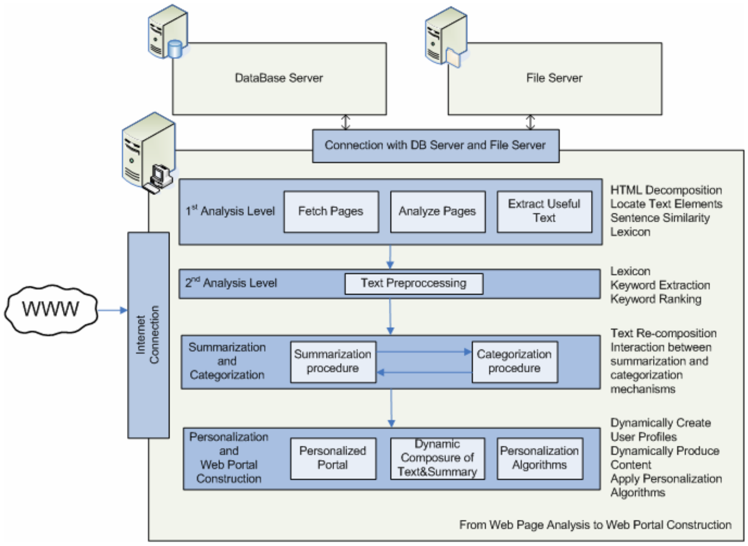
DataBase Server
File Server
Connection with DB Server and File Server
1st Analysis Level
Fetch Pages
Analyze Pages
Extract Useful Text
HTML Decomposition
Locate Text Elements
Sentence Similarity
Lexicon
2nd Analysis Level
Text Preproccessing
Lexicon
Keyword Extraction
Keyword Ranking
WWW
Internet Connection
Summarization and Categorization
Summarization procedure
Categorization procedure
Text Re-composition
Interaction between summarization and categorization mechanisms
Personalization and Web Portal Construction
Personalized Portal
Dynamic Composure of Text&Summary
Personalization Algorithms
Dynamically Create User Profiles
Dynamically Produce Content
Apply Personalization Algorithms
From Web Page Analysis to Web Portal Construction

Figure 1: Systems Basic Architecture

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 in [1]. 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 [2]. 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.

III. 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$_r$	Popular	10
…		
29	Busines	1

The keywords are ordered in descending order of 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 and (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.

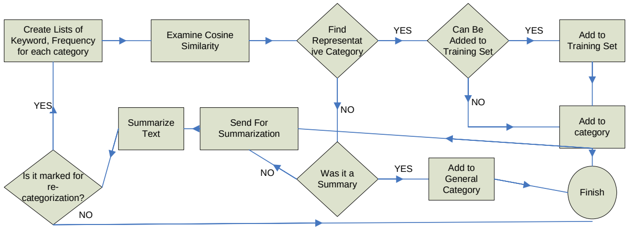
Create Lists of Keyword, Frequency for each category
Examine Cosine Similarity
Find Representat ive Category
YES
Can Be Added to Training Set
YES
Add to Training Set
NO
NO
Add to category
YES
Summarize Text
Send For Summarization
NO
Was it a Summary
YES
Add to General Category
Finish
Is it marked for re-categorization?
NO

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

A. *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:

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

Where wk,i is the frequency of the kth 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

$$k_1 = 1 + 0,1x \tag{2}$$

where x is the times that the keyword is found in the title. Accordingly k_2 derives from the following equation

$$k_2 = 1 + 1,2y \tag{3}$$

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

$$y = \frac{n}{t}\frac{m}{t} = \frac{nm}{t^2} \tag{4}$$

B. *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 into the summarization procedure.

$$k_3 = \begin{cases} A \cdot cw_i & \text{where A>1 and cw the \underline{positive} category weight} \\ -A \cdot cw_i & \text{where A>1 and cw the \underline{negative} category weigh} \\ 1 & \text{for neutral or not ranked by the system keyword s or if A=0} \end{cases} \tag{5}$$

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:

$$S_i = \sum w_{k,i}(k_1 + k_2)k_3 \tag{6}$$

IV. Experimental Procedure

Armed with our summarization and categorization mechanisms we conducted experiments that would reveal the two-sided relationship between categorization and summarization. In order to have a working knowledge base (even a small one) we gathered news articles from some major news portals from the U.K. and the U.S. We defined six distinct news categories: business, entertainment, health, politics, science and sports, organizing our captured texts (around 180 for each category) to them. Afterwards, using our categorization mechanism, we extracted 50% of the keywords of each text and associated each keyword with the text's category using the absolute frequency as a relativity measure.

In particular we carried out three types of experimental procedures.

First of all, we needed to determine the text's keywords percentage we should keep, in order for our categorization module to be the most effective. Towards this direction we modified the keeping percentage from 0.1 (i.e. 10% of the keywords) to 1 (i.e. all the keywords) with a step of 0.1, using a representative text for each of the aforementioned categories and categorizing it. The text that is entered to our categorization module has not been used for the construction of the knowledge base (was not part of the training set). For each keyword percentage we measure the cosine similarity between the text and each category that resides in our knowledge database. We conduct the experiments using a minimum keyword size limit of 5 and 6 for both the knowledge base and for the text that is to be categorized. Following are some charts depicting the results.

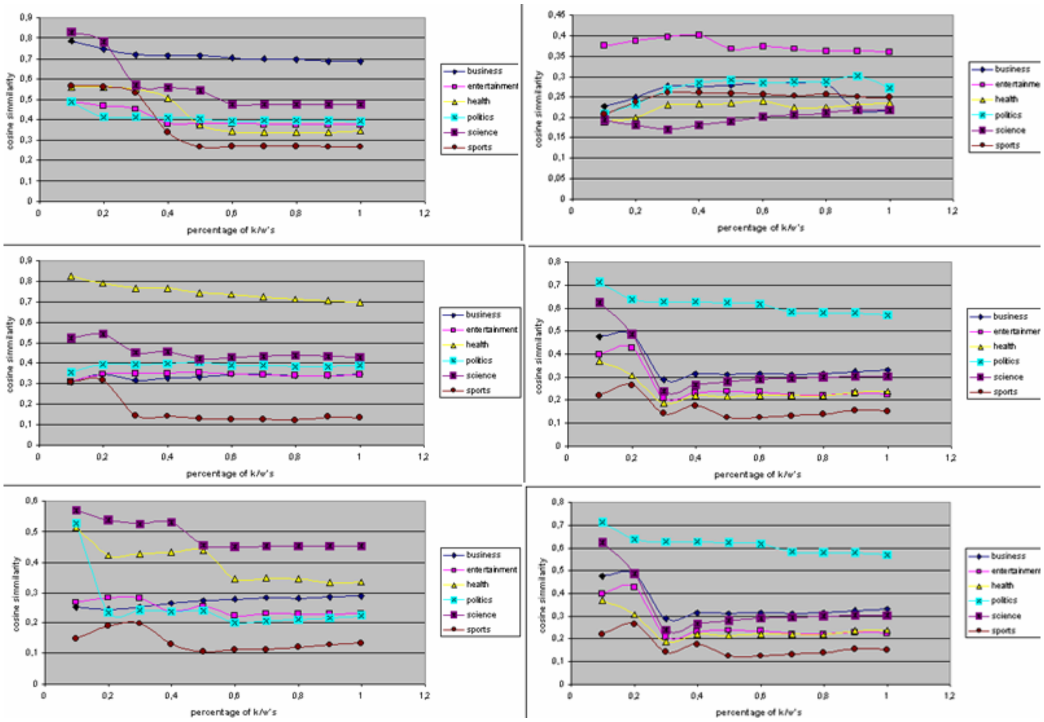

Figure 3: Cosine similarity of texts compared to categories. Training set is constructed with 50% of the keywords kept (pre-processing procedure).

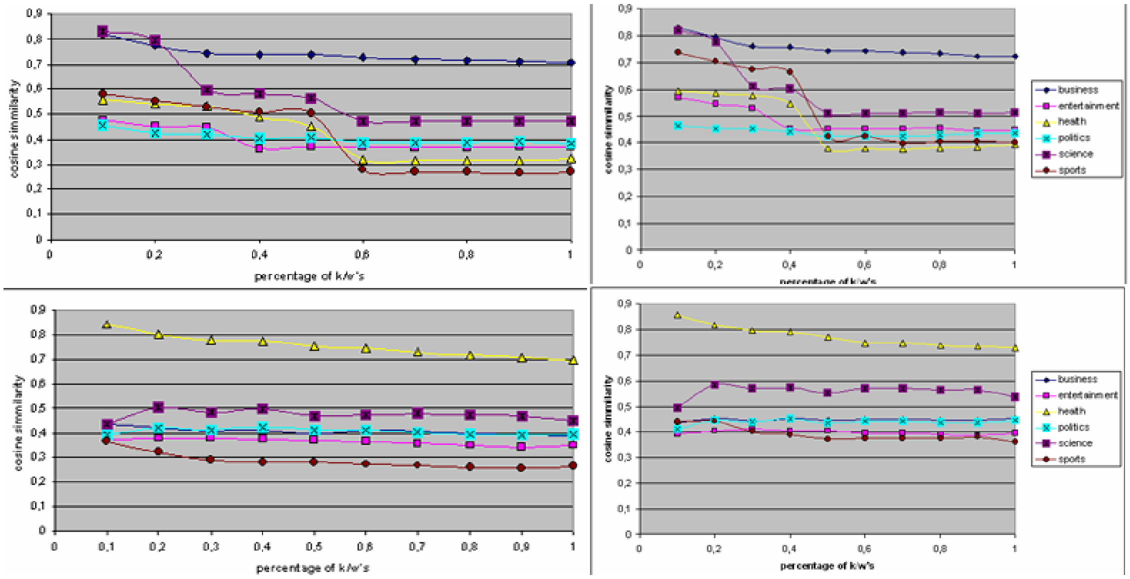

Figure 4: The first column depicts the cosine similarity measured by utilizing the 50% of the keywords from the training set and the second column is the same cosine similarity measured by utilizing the 100% of the keywords from the training set.

From figure 3 (categorization procedure results), it is concluded that a percentage of 30% of the text's keywords should be kept for our categorization procedure to be optimal. Even though that a lower percentage might be sufficient to decide on the text's category, we are keeping a percentage of 30% because, firstly it gives us almost always the right category decision and secondly, it provides us with a stronger distinction percentage between the correct category and all the others. This difference of similarity is, in our opinion, the most important factor for a categorization mechanism since, it can provide us, even with expanding knowledge databases, with correct category answers. For example it is possible, when our database has many categories, some of which similar to each other, the similarity of an input text to be relatively high for more than one category. In this case, the difference of similarity can be a better measure of categorizing, rather that an absolute similarity threshold.

It is clearly depicted in figure 4 that a text can achieve better scoring using a minimum keyword size limit of 5 letters and keeping 50% of the resulting keywords (from the training set). This way the knowledge base is more refined while no category-important keywords are left out of the procedure.

In the next step of our experimentation, we wanted to examine the influence that the summarization procedure has on the categorization stage. In order to achieve this, we first summarized some humanly pre-categorized texts and then inserted them into the categorization procedure. Finally we compared the output of the categorization module (which in this way gives the summarized text's relativity with each registered category), with the predefined category of the text.

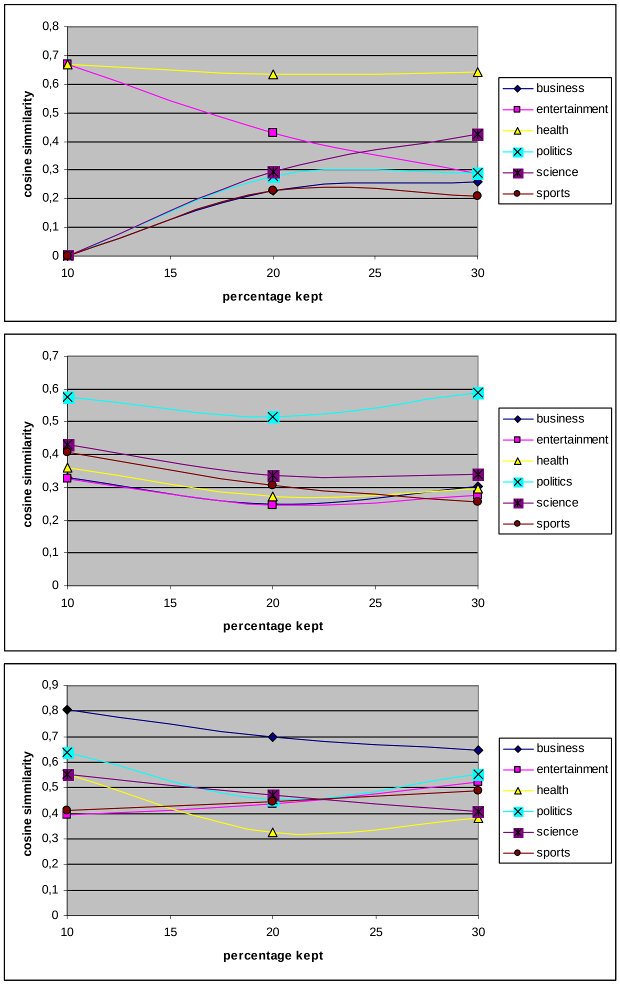

Figure 5: Cosine similarity measured for categorizing summaries by keeping different percentages for creating the summaries

We used multiple summarization sizes in order to see the effect that they have on the categorization of the summary. Following some sample charts of this experimentation that took place using texts belonging to different categories reveal the ideal percentage of sentences that could form a summary.

From this kind of experimentation we noticed that when keeping a plausible amount of the initial sentences, around 20%, for producing the text's summary, we could categorize the summary correctly to the text's category thus saving a tremendous amount of work on the categorization side, since the summary is only a small portion of the text. This result is of huge importance for a fast responding, real time categorization system.

Another field that our experimentation investigated concerns the effect of the categorization to the summarization procedure. In order to discover the potential relationship, we constructed our summarization mechanism incorporating the categorization feature. For example, when we know a-priori the text's category, this information is taken into consideration from the summarization module and each sentence rating is adjusted accordingly. For example, should a sentence contain many keywords irrelevant to the text's category (a priori knowledge), it's rate will be much lower, or even negative, than when we don't know the text's category.

Using corpus texts, we first produced the text's summary without the use of the categorization factor (i.e. k_3=1) and afterwards we used this extra information to produce the summary and compared both of the results with the text's summary, which came with the corpus and was formatted by humans. The results are quite positive since we discovered that the categorization feature improved our summarization results by a factor of 10% or even more, meaning that the sentences which our summarization mechanism kept after the use of the categorization information are closer to the “optimal”.

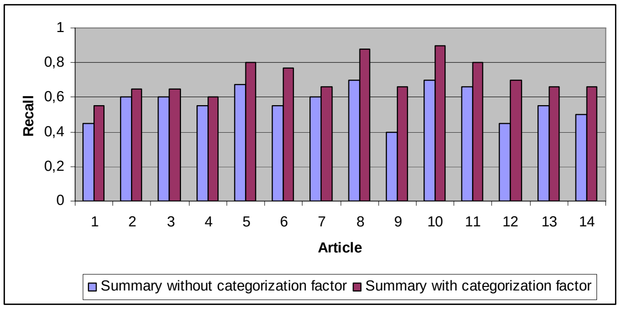

Figure 6: comparison of recall from summaries extracted with and without categorization factor

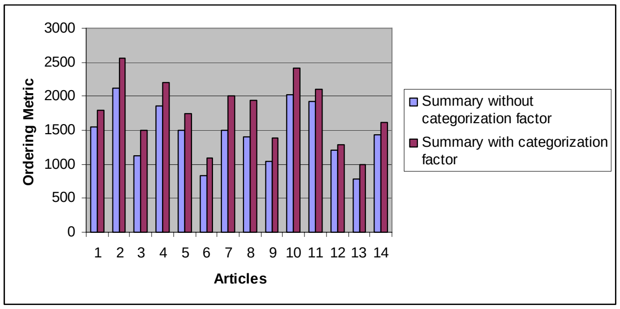

Figure 7: comparison of ordering metric from summaries extracted with and without categorization factor

In order to compare the results from both the cases (using the categorization information and not) we used the recall metric, i.e. how many of the sentences of the human-formed summarization where recovered by each procedure, and a sentence ordering metric. The latest was used to indicate the importance that the order of the sentences has in a summary. For example, it is possible that both of the summarization techniques achieve the same recall scoring but the ordering of the sentences is better in one of them. In fact, we observed that the summarization technique which utilizes the categorization information produces not only better recall scoring, but also higher sentence ordering score.

V. Conclusion and Future Work

In this paper 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", 3rd International Conference on Universal Access in Human – Computer Interaction, Las Vegas, Nevada, USA, 22 - 27 July 2005

[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] 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

[17] Y. Gong and X. Liu. Generic Text Summarization using relevance measure and latent semantic analysis. In Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'01), pp. 19 – 25, 2001, New Orleans, USA.

[18] I. Mani and M.T. Maybury. Advances in automatic text summarization. Cambridge, MA: The MIT Press, 1999

[19] Hahn, U. and Mani, I. The Challenges of Automatic Summarization. IEEE Computer 33, 11, 29-36.