EVALUATING PERSSONAL: A MEDIUM FOR PERSONALIZED DYNAMICALLY CREATED NEWS FEEDS

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

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

Vassilis Tsogkas
Research Academic Computer Technology Institute, N. Kazantzaki, Panepistimioupoli and
Computer Engineering and Informatics Department, University of Patras
26500 Rion, Patras, Greece
tsogkas@cti.gr

ABSTRACT

In this paper we present the evaluation of PeRSSonal, a system which generates personalized, dynamically created RSS feeds and is focalized on small screen devices. PeRSSonal is a complete system able to consolidate news from major news portals, categorize and summarize them, and finally syndicate them personalized to the end users. In an era where news feeds, become a part of the daily Internet life and small screen devices gain more and more attention by the end users, PeRSSonal seems the panacea that could provide fast and easy access to everyday news. The system is based on algorithms that incorporate the user into the categorization and summarization procedure of news articles, while the results are presented to the user according to his/her interests and end device.

KEYWORDS

Personalized feeds, Data Preprocessing, Categorization, Automated Summarization, Small Screen Devices, RSS feeds

1. INTRODUCTION

During the last years, the advances in technology and the ease of access to information have modified dramatically the status of what we call World Wide Web. Every day, thousands of articles are created by the vastness of information and news articles. This freedom that the Internet is meant to provide has attracted more and more users not only to “check out” in a daily basis the “Internet newspaper”, but also to create their own articles generating thus their own sources of information and articles. The latest trend of blogging, which is far from serving exclusively a personal dairy, acts primarily as a media of information exchange. Besides, research is already underway in order to depict the phenomenon of journalism at the WWW (Baron, 2004).

The aforementioned facts can be considered as an innovation for our world and as a founder of democracy. However, the condition that was described generates a number of repeated and expanding

problems for the users of the Internet who try to access information via their mobile phones, PDAs and generally small screen devices. These kinds of systems that are becoming more and more common already do have the power to run complex interactive applications (Fitzmaurice et al. 1993). The main problem of such type of users is the available space that they have on their monitor in order to track and read news from news portals. Despite the increasing resolution of PDA screens, limitations on the physical size of the screen will prevent these devices from ever reaching parity with desktop computers (Gutwin & Fedak 2004).

The problem of presenting information to the end-user is the one side of the coin. The other side is the fact that users are forced to specific tags of information, while most of the time "(s)he finds it difficult" to locate useful information that he/she is searching for within the news portals. As the news articles that are generated daily from the major news portals are increasing in numbers, the act of locating "useful articles" is becoming extremely demanding. With the term "useful articles" we define the articles that the Internet user is really interested in and actually wants to read.

The rest of the paper is structured as follows. In section 2 we present the related work of our research aspects. The next section presents the architecture of the mechanism while section 4 introduces the algorithmic aspects of the complete mechanism. In section 5 we present some basic results of the evaluation of the mechanism and in section 6 we conclude the paper with possible future directions that can be done in order to further improve the PeRSSonal tool.

2. RELATED WORK

The mechanism that we have created includes, among others, algorithms for categorization and summarization and as the dynamically created personalized RSS feeds relies on the aforementioned algorithms we will present the state of the art of the classification and summarization procedures.

Text classification (categorization) is the process of deciding on the appropriate category choice for a given document. Classification tasks include determining the topic area of an essay; deciding on which folder an email message should be directed; and determining the newsgroup a news article belongs (Google News). The purpose of text categorization (Hayes et al. 1988; Hsu, et al. 1999; Kummamuru, et al. 2004) is to accompany readers to their search of news articles, by creating and maintaining key categories which hold articles related with a specific topic of interest. News articles are categorized to the predefined set of categories using some criteria which vary from one categorization technique to another. The use of predefined categories can be relatively coarse-grained, i.e. only some basic, unrelated to each other, categories are defined, such as business, education, science, etc., or fine-grained where many categories, which are frequently overlapping with each other, are introduced. The latest is the case of our mechanism, since the correspondence of an article to a category is not one-by-one.

The goal of summarization as explained by Radev et al (2001 & 2005), is to generate a summary out of one or more articles, usually related to each other (Wasson 1998), easing the user from the tedious task of reading large texts. Most of the times the summarization approaches are based upon a "sentence level" where each sentence is rated according to some criteria (e.g. important keywords, lexical chains, etc.). Some techniques (Ferragina & Gulli 2005; Goldstein et al. 1999), try to find special words and phrases in the text, others (Hayes et al. 1988) compare patterns of relationships between sentences or take into consideration the length of the sentences or the word case. In our research we utilize a "sentence level" approach for sentence selection escorted with the ability of detecting special words in it. These words may express high category significance or user preference. In this paper we present a novel Internet service whose main scope is to support Internet users that are interested in reading, on a daily basis, specific news articles and we focalize mainly on users with small screen devices. The challenge is big, as we have to locate the news articles that the user wants to read and at the same time present them in such a way that the user will be able to read even a fraction of them that will be representative of the article. Within these limitations we present a mechanism based on personalized RSS feeds utilizing dynamic creation of summaries. Despite the fact that it is not a difficult procedure to create dynamic RSS feeds, the major and minor news portals are mis-utilizing this service. A simple example is the RSS feeds of CNN where the user is only provided with the title of the latest articles while no information about the body is provided.

A Similar approach to PeRSSonal for document summarization is Columbia's Multi-Document Summarization (McKeown et al., 2001). However the creators focus mostly on the summarization

characteristics of the system lacking categorization features. PeRSSonal on the other hand, incorporates many different preprocessing and retrieval techniques to implement a competitive indexing service that will serve personalized content to its users, especially those with small screen devices.

3. ARCHITECTURE

The mechanism consists of standalone subsystems that work together in order to produce the personalized RSS feeds. The collaboration between the distributed systems is based on open standards for input and output which are supported by each part of the system and by communication with a centralized database. Figure 1 depicts the architecture of the complete mechanism.

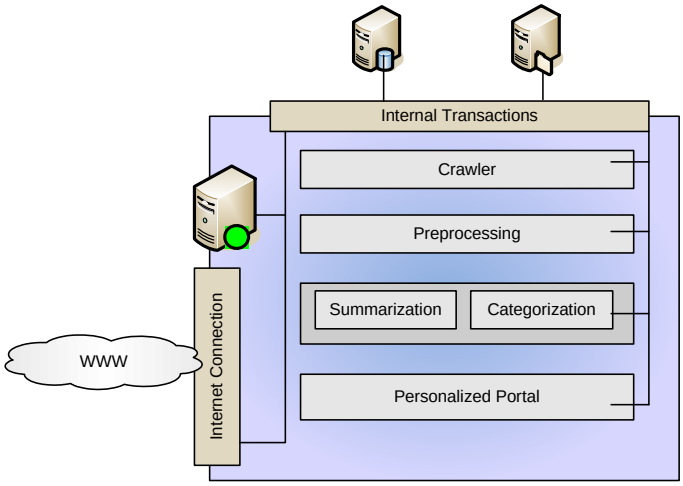

Figure 1: General Architecture of the Mechanism

The procedure of the mechanism as depicted in Figure 1 is: (a) capture pages from the Internet and extract the useful text, (b) parse the extracted text and preprocess it (keyword extraction), (c) summarize and categorize the articles and (d) personalize the results and present them to the end users.

In order to capture the pages, a focused web crawler is utilized. More information about this procedure can be found in Bouras et al. (2005). The required information from a Web Page is the useful text (body of the article), as well as some other page meta-data, such as URL and capture date. In this manner, the database is populated with news articles and well-shaped information about them, which is ready to be forwarded to the text preprocessing subsystem for further analysis.

The main scope of the second analysis level is to apply text pre-processing algorithms on the article, providing as output keywords, their location into the text and their frequency of appearance in it. These results are necessary in order to proceed to the third analysis level. Information about the preprocessing mechanism can be found in Bouras et al. (2006).

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 led to the personalization module of our mechanism, where personalization algorithms are applied, and are finally presented back to the end users in the requested form (i.e. RSS feed). The role of the personalization layer is to feed each user only with articles that he/she “wants” to face according to his dynamically created profile.

4. ALGORITHM ANALYSIS

In order to analyze how each algorithm is applied on the texts we will present the procedure that is followed in each step but we will focalize on the creation of the personalized summaries. The complete flow of information of our system is depicted in Figure 2.

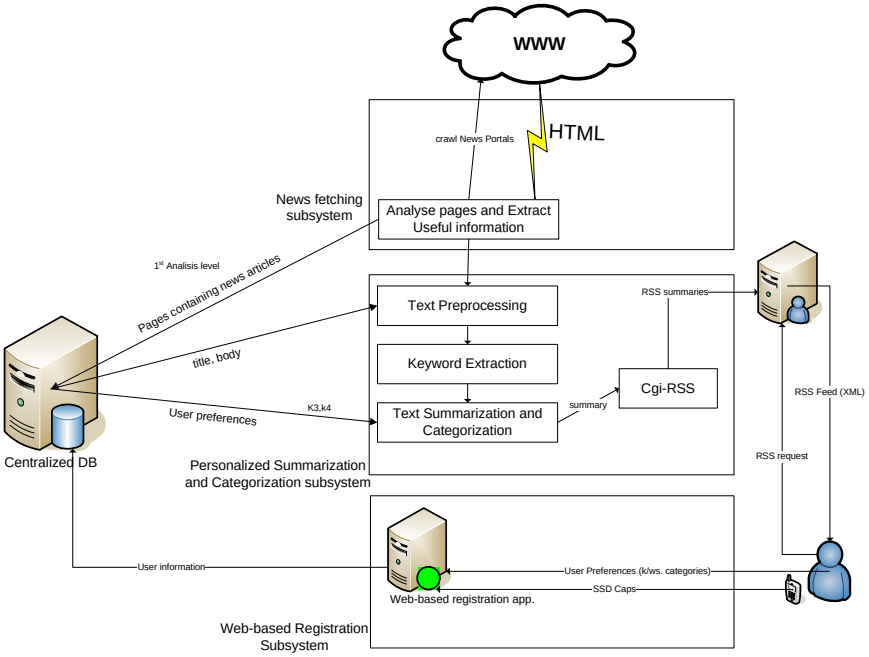

Figure 2: Flow of information

The procedures of fetching the articles from the WWW, preprocessing them, categorize and automatically summarize them are analyzed superficially as the intention is to focalize on the web interface and the personalization factors of both the Web Portal and the personalized RSS that is offered to the visitors of the Portal.

4.1 Fetching Articles, HTML analysis, Categorization and Summarization

The algorithm that fetches the article is very simple and is based on the fact that every web portal includes a series of RSS feeds that are offered to the end user. Opposed to having to visit every page of the many news portals that exist on the WWW, we fetch their RSS and more specifically the one that includes the daily “top stories”. From the XML structure of the feeds we can obtain the most important articles that are published to each news portals, together with information that have to do with the title of the article, its exact URL and the date of publication. After the articles' URLs are extracted from each feed, the focused crawler, currently working as a wrapper, changes its functionality to a simple crawler and visits every singe URL extracted from the RSS feeds in order to obtain the HTML pages that include the latest articles of the RSS.

The next procedure includes the steps of analyzing the HTML files, extracting the useful text from them and preprocessing the useful text in order to extract the article's keywords. The useful text extraction is based on the fact that HTML pages can be depicted as a tree with every HTML tag holding a node on the tree, while every leaf includes pure text. In order to extract the useful text we utilize a clipping extraction technique described in Bouras et al. (2005). The preprocessing techniques and the keyword extraction techniques (Bouras et al., 2006) follow.

If the categorization procedure fails to categorize a text, meaning that the levels of similarities between the article and the categories provide an obscure choice, then a simple assumption is used in order to retry the categorization step. A well summarized text that includes only the important information of a text and thus, only the important keywords of it, has a higher possibility to be categorized than the original text. After a failure of the categorization procedure, the system summarizes the text and attempts a second categorization. If the categorization fails again the text is labeled as uncategorized or general.

The summarization procedure is based on heuristics. During the attempt of summarizing a text we utilize two metrics: (a) the existence of the keyword in the title and (b) the frequency of a keyword. These are the factors k_1 and k_2 and will be referred afterwards by their abbreviation. A keyword with very high frequency in a text is considered to be representative of it and thus, any sentence that includes this keyword

can be thought as representative of the text. Additionally, any text keyword that also exists in the title, is marked as an important one and the sentences that include this keyword are considered more representative for the text's meaning.

Based on these heuristics we create the summary of the article which consists of its most representative sentences. In order to find them, we create a score for each sentence according to the factors k_1 and k_2. Assuming that the text T has s sentences where i = [1..s] and f keywords where k = [1..f] then each sentence is assigned a score according to the following equation.

$$W_i = \sum (1 + rel(\, fr(kw_{k,i}\,)))(k_1 + k_2) \quad (1)$$

where k_1 and k_2 are the factors explained in the previous paragraph and rel(fr($kw_{k,i}$)) is the relative frequency of the keyword k in sentence i.

As long as the text is categorized we can utilize this fact in order to create a more efficient summary. The theory that we are relying on, is that if the text is categorized, then there exist some keywords in the text that are representative of the text's category. This information can lead us to the use of another factor, k3 that represents the ability of the keyword to represent a category. Assuming that the relative frequency of a keyword within a category is cf then k3 can be computed as:

$$k_3 = A \cdot (1 + cf_i) \quad (2)$$

where A is the "special weight of k_3" and is added in order to represent how much the computation of the sentence weighting will be relied on factor k_3 . Typically A takes values between 1 and 3, with the later meaning that sentence weighting relies greatly on the categorization information. If a text keyword does not belong to the category of the text then k_3 is set to 1. k_3 factor is added to the overall weighting equation as a product, as shown in the following equation.

$$W_i = \sum (1 + rel(\, fr(kw_{k,i}\,)))(k_1 + k_2)k_3 \quad (3)$$

4.2 Web Interface

The Web-based registration and user's interface subsystem represents the initial interface between the whole mechanism and the end user. A user registers to the system providing information about i) his small screen device (device capabilities) and ii) his categories' preferences. This information is stored in the centralized database and is used afterwards at the personalized summarization procedure. While registering, each user is prompted with the categories that exist in the mechanism and is asked to assign a score to each of them according to his/her preference. Relying on these selections, we can create a simple user profile. At first, we create a list of the categories that the user likes and the ones that (s)he does not like. This assists us to a first article cleanup when selecting which news articles the user is interested in. The user is not just prompted to select the "likes" and "dislikes" but is induced to select a weight for each category. By utilizing these data, we are able to create a more detailed profile of the user which consists of a list of keywords that indicate the ones that the user likes and the one that (s)he dislikes followed by a relative frequency depicting the preference of the user towards the specific. The creation of the profile is constructed with the help of the following algorithm.

```
For each (selection s) {
 If (s!=0) {
    Keyword_name_usr = select 20*s keywords from category keywords
    Keyword_weight_usr = select (2*s*relative frequency) from category keywords    }
else {
    Keyword_name_usr = select 10 keywords from category keywords
    Keyword_weight_usr = select relative_frequency from category. Keywords    }
    Insert into user profile keyword_name_usr, keyword_weight_usr
    If exists
        Update user profile set keyword_weight += keyword_weight_usr where keyword_name = keyword_name_usr    }
```

By executing the aforementioned algorithm we result to the selection of what is really needed for the personalization procedure: a) Many keywords from the categories that the user has selected with high score (either positive or negative), and few keywords from the categories that the user has selected with low score. b) High positive value for the relative frequencies of the keywords belonging to categories that the user has

selected with high preference, and low negative value for the frequencies of the keywords belonging to categories that the user has selected with negative preference. By further analyzing this information we can achieve the following: a) select texts from the categories that the user likes and do not belong to a category that the user dislikes b) refine the outcomes of the summaries by adding the personalization factor.

The above procedure, including the creation of the keywords list, was facilitated in order to be able to add another factor which personalizes the summaries. The factor utilized is called k_4, expresses the effect that personalized keywords have on the keyword extraction procedure and can be used as a product to equation (1) or (3). Assuming that for the user A we have constructed a list of keywords followed by their relative frequency (preference of the user), k_4 derives from the following equation:

$$k_4 = B \cdot (1 + uf_i) \qquad (4)$$

where uf_i is the user's preference for the keyword i and B is the "special weight" of k_4 and defines how much will k_4 affect the result of the sentence weighting. After experimental procedure, we have concluded to the value 1,8 for B. When knowledge of the article's category is available, we apply the k_4 factor on equation (3), while the application of the factor on equation (1) is done when the categorization subsystem is unable to provide us with evidence of the exact category.

The two-step refinement of the articles is very helpful, firstly to decide on which articles to present to the end-user and secondly, how to present the articles to the specific device of the user. The aforementioned refinement of the articles provides us with two unique features for the mechanism. The first one is the ability to select which articles to present to the user relying on his/her preferences, personalizing in that way a dynamically created RSS feed for each user. The second feature is the ability to personalize the dynamically created RSS feed (summary) on the device of the end-user presenting only the amount of data that can be viewable/browsable within one or two pages of the specific small screen device.

5. SYSTEM EVALUATION

In this section we provide some experimentation of the PeRSSonal system as well as some screenshots of the responses the system sends back to the users who utilize small screen devices. In order to evaluate our mechanism we conducted three sets of experiments. During the first test we created 10 user profiles with specific preferences concerning the categories. We assumed that these people were receiving daily to their RSS reader the feeds from 10 portals and the feed of our portal (which also collects news from these 10 portals). We examined the amount of the articles collected from the news portals that were of some interest to the user and how many articles originating from our RSS feed were of interest to the user. The results are depicted in Figures 3 and 4.

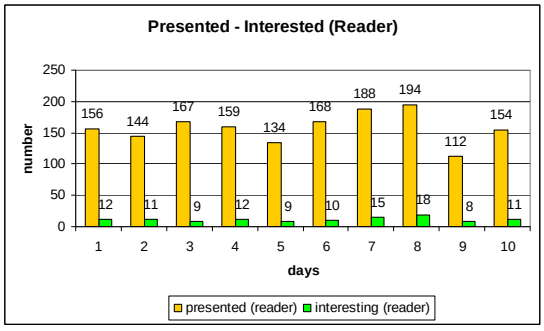

Figure 3 Presented and Interesting articles directly from all news portals

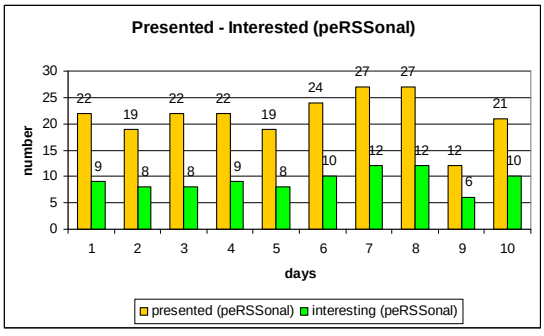

Figure 4 Presented and Interesting articles from personal

As we may observe, PeRSSonal presents on average 85% less articles daily but the percentage of articles that the users seem to be interested in is more than 40% of the presented, while in the second occasion the users are interested in reading only the 7% of the articles presented. This means that the mechanism can achieve better cleaning of the feeds providing articles that the user is really interested to read.

A second experiment that is focused on the adaptability of the mechanism proves that the mechanism is able to produce even more effective results by establishing the exact profile of the user. The RSS feeds include the original URL of the HTML page that includes the article. The URL of our RSS includes a URL to our portal that redirects the user to the original HTML page. This is done in order to record the articles that the user selected to read. By monitoring the activity of the user, we are able to update the list of keywords that represent his/her profile, updating thus constantly his/her profile. As a result, we observed that in less than 7 weeks time, the system is adapted completely to the user's true preferences. The following graphs depict the adaptation of the mechanism to the user during the flow of time.

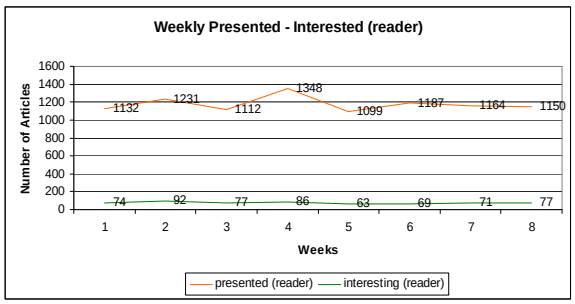

Figure 5 Weekly presentation of articles from the RSS reader

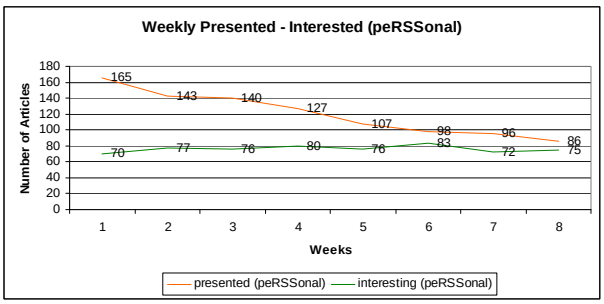

Figure 6 Weekly adaptation of personal to the profile of the user

As observed from the aforementioned figures, given a period of time (relatively short), the mechanism is able to adapt almost completely to the user's profile. What we want to achieve is the converging of the "presented" line to the interesting line as the number of articles that seem to be interesting to the user does not change through the time. The aforementioned graphs prove two basic functionalities of the mechanism. Firstly, the mechanism is able to act as a filter, so that articles that are of low interest to the user are not presented to him/her, and secondly, the mechanism is able to dynamically create the profile of the end user and in less than 7 weeks time presenting only information that is important to the user.

At a third experimentation phase we examined the ability of the mechanism to adapt its content to the user's end device so that the users would be able to view clearly the RSS from their small screen device. When an unregistered user requests an RSS feed, a default RSS response, which contains the default summaries, is sent back. On the other hand, if the user is registered, he is fed with a personalized summary corresponding to his/her profile.

Figure 7: Default RSS feed

Figure 8: Personalized RSS feed

Figure 9: RSS response for user A about a specific article

Figure 10: RSS response for user B about the same article

An important factor to take in mind is that different users receive different RSS responses which vary in terms of news': length, ordering, amount, and categories. It is possible that two users receive the same articles but different summaries, as depicted in Figures 9 and 10.

The mechanism is such that extensive experiments can be carried out in order to observe every aspect of it and every possible state. Briefly, the mechanism is able to create personalized, dynamically created RSS feeds with variant summary size and content according to the user's profile and end device. The system is able to adapt to the specific user needs and act as a complete personalized micro-site or personalized RSS feed in order to cover the needs even of the sternest users.

6. CONCLUSION AND FUTURE WORK

In this paper we have presented a mechanism that is able to complete a procedure of collecting news from major news portals and present them personalized back to the end-users. This mechanism is extremely helpful for Internet users who are spending a lot of time trying to find news of their interest through major or minor news portals or even through RSS feeds. Even though the procedure of accessing all the news portals in order to collect useful information is part of our everyday life, information that is shown on the screen of the end user includes almost 80% of not needed or even "trash" information.

As a future work for our mechanism we are thinking of a news tracker system which will be able to track the changes that are done on news articles. As, more and more, articles about a specific theme are published on several news portals or even on the same news portal we should be able to collect all the similar news and present them as one to the end user, providing also with the several links that the articles derive from and let the user make the best choice on which link to follow. Additionally, the automated procedure of collecting news articles must be empowered by a more effective focused crawler in order to avoid collecting unwanted data, putting the focus only on information that is needed as a feed for our mechanism. Experimentation for tuning the various factors is already underway and is expected to lead us to better results and future versions of the mechanism.

Finally, as the system is able to work at a very high speed, creating dynamically the RSS for the user in real time, we are thinking of creating an add-on for the news portals that will enable the real-time creation of personalized RSS feeds for the end-user directly through them.

REFERENCES

Baron D. 2004. Persistent Media Bias. Stanford University, *Graduate School of Business Research Papers*: No. 1845.

Bouras C., Poulopoulos, V. & Thanou A., 2005 Creating a Polite Adaptive and Selective Incremental Crawler. *IADIS International Conference WWW/INTERNET*, Lisbon, Portugal, Volume I, C., pp. 307 – 314

Bouras C., Poulopoulos V.& Tsogkas V., 2006 The importance of the difference in text types to keyword extraction: Evaluating a mechanism. *7th International Conference on Internet Computing (ICOMP 2006)*, Las Vegas, Nevada. pp. 43-49

Ferragina P. & Gulli A. 2005 A Personalized Search Engine Based on Web-Snippet Hierarchical Clustering. Software: Practice and Experience Journal, *Proceedings of the WWW Conference* Vol. 38 Issue 2. pp. 189 – 225.

Fitzmaurice, G., Zhai, S., & Chignell, M., 1993. Virtual Reality for Palmtop Computers, *ACM ToIS*, 11,3, , pp.197-218.

Goldstein J., et al. 1999. Summarizing Text Documents: Sentence Selection and Evaluation Metrics. *In Proceedings of ACM SIGIR Conference*. pp. 121–128.

Google News, http://news.google.com. [accessed 5 July 2008].

Gutwin C. & Fedak C., 2004. Interacting with big interfaces on small screens: a comparison of fisheye, zoom, and panning techniques. *Proceedings of the 2004 conference on Graphics interface*, London, Ontario, Canada, pp. 145-152.

Hayes P. J., et al. 1988. A News Story Categorization System. *In Proceedings of the second Conference on Applied Natural Language Processing*, pp. 9-17.

Hsu W.-L., et al. 1999. Classification Algorithms for NETNEWS Articles. *In Proceedings of CIKM*, pp. 114–121.

Kummamuru K., et al. 2004 A Hierarchical Monothetic Document Clustering Algorithm for Summarization and Browsing Search Results. *In Proceedings of WWW Conference*. pp. 658-665.

McKeown K. Barzilay R. Evans D., 2001, Columbia multi-document summarization: Approach and evaluation, *Proceedings of the Workshop on Text Summarization, ACM SIGIR Conference 2001*

Radev D. R., et al. 2005. NewsInEssence: Summarizing Online News Topics. *Communications of the ACM* Vol. 48, No. 10 pp. 96-98.

Radev D. R., et al. 2001. Interactive, Domain-Independent Identification and Summarization of Topically Related News Articles. *In Proceedings of ECDL* pp. 225-238.

Wasson M. 1998. Using Leading Text for News Summaries: Evaluation Results and Implications for Commercial Summarization Applications. *In Proceedings of ICCL*. pp. 1364–1368.