FASTNEWS: SELECTIVE CLIPPING OF WEB INFORMATION

Rodrigo Branco Kickhöfel, Gilnei Barroco Farias
Catholic University of Pelotas - UCPEL, Brasil

Stanley Loh
Catholic University of Pelotas – UCPEL, Brasil
Lutheran University of Brazil – ULBRA, Brasil

Keywords: text mining, clipping, information filtering, information extraction

Abstract: This work presents a software system for selective clipping of web information. The system allows users to register queries, expressing their information needs, and monitors information sources (Web sites), in order to find new information and to push it to the users. The difference from traditional Web clipping systems is that FastNews only retrieves information relevant to the user's need, that is, it has an intelligent engine that extracts only information parts according to the interest of the user. Currently, the system allows watching news, currency conversion and weather forecasting. An additional functionality is to allow users to enter an URL (Web site) to monitor, against the traditional use of predefined sources.

1 INTRODUCTION

Web is a great repository of information. However, the huge volume of Web pages generates information overload. This problem occurs when people have too much information so that they can not find what is desired, leading to lost users (CHEN, 1994).

Information Retrieval (IR) systems help people to find documents whose content may contain relevant information (SPARCK-JONES & WILLET, 1997). The traditional search engines like AltaVista, Google, AlltheWeb and Teoma use IR techniques for Web documents and pages.

However, IR systems need online interaction between the user and the system. Sometimes, this is a burden. If someone likes to read news everyday in the Web, she/he needs to go to a Web site everyday and to select the desired news only reading the headlines.

To minimize problems like that, there are Information Filtering systems. According to BLOEDORN ET AL. (1996), this model consists of applying filters to dynamic bases, informing users when new relevant documents arrive in the base. The relevancy evaluation is made using queries or profiles, assuming that the interest persists for a certain time period.

The difference of Filtering from IR is that in the former the user specifies his/her interests one time and the system is responsible for monitoring the information sources and selecting relevant documents.

Filtering of news published in the Web is popularly called Clipping. Web sites offer clipping of their news through newsletter service. However, there is no selection of news. All the news published are sent to all users registered in the service. Sometimes, news are classified in subjects, but the sending still includes irrelevant information. The consequence is again the information overload.

This paper presents FastNews, a system for selective filtering of Web information. The main difference from other clipping systems is that there is selection of news according to the user's profile. This minimizes the information overload since users receive only news about their interests.

In this system, users define their information needs or interests as queries and the system periodically monitors information sources (Web sites), selecting textual parts relevant to the user. This alleviates the user from having to visit Web sites everyday, to examine headlines or more than

Branco Kickhöfel R., Barroco Farias G. and Loh S. (2004).
FASTNEWS: SELECTIVE CLIPPING OF WEB INFORMATION.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 451-454
DOI: 10.5220/0002655404510454

this and then to select the relevant news to read. Other advantage is that users can store the filtered news into an e-mail software for later reading or for organizing the news according to subjects.

The paper is structured as follows. Section 2 presents some concepts and works about Information Filtering and section 3 presents the same for Information Extraction. Section 4 describes the FastNews system in details. Section 5 discusses concluding remarks and future work.

2 INFORMATION FILTERING

Information filtering is a kind of information retrieval. The goal is to find relevant information in a collection of documents.

However, the difference is that filtering reduces the collection (or information base) to a narrower subset with greater probabilities of having relevant information (KORFHAGE, 1997).

In the filtering model, the query is predefined, reducing the cognitive charge for the user: he/she does not need to formulate new queries, neither to search different sources for the desired information.

The query is executed automatically when new information arrives to the collection or base, without the user having to stimulate the system.

Filtering systems work as an human intermediary in the retrieval process, collecting information from many sources and producing an organized summary for the users (OARD & MARCHIONINI, 1996).

Queries may be defined by users using languages to express their needs or a profile may be used by the system to find information related to the user's interest. Profiles work as classes that represent the information need or interest of the user or a group (MOSTAFA ET AL., 1997). One alternative is to use keywords to express interesting subjects of a user.

Profiles may be automatically created using machine learning methods (OARD & MARCHIONINI, 1996) or they may be manually specified by the own user or by a domain expert.

According to OARD & MARCHIONINI (1996), there are 2 kinds of information filtering. One is the content-based filtering, where attributes of the user are matched against attributes of the information source (for example, a document or Web page). The second kind is the social or collaborative filtering, where items are cross-sent to users with similar attributes or interest.

FastNews uses the content-based approach, comparing user's profile against information sources.

3 INFORMATION EXTRACTION

Information extraction (IE) is responsible by analyzing relevant parts within a text and identifying specific data. It works converting unstructured data into attributes of a structured database (COWIE & LEHNERT, 1996).

The difference of IE from IR is that the latter only retrieves documents, but the former has to retrieve pairs attribute-value.

According to COWIE & LEHNERT (1996), the importance of the IE systems is to minimize the effort for information acquisition, freeing people from having to read or/and analyze texts and look for the desired information.

Usually, IE uses *tags* to identify the presence of a relevant information. For example, the word "*years*" in a text about some person may indicate the age of this person. But pattern matching and syntactic analysis may also be used for information extraction.

Grammars, regular expressions, parsers and finite automata are used for recognizing relevant information.

IE systems only work in specific domains and for extracting specific information. They are dependent on the domain, on the task (information to be extracted) and on the text collection used (CROFT, 1995).

4 DESCRIPTION OF THE FASTNEWS SYSTEM

FastNews is a Web-based software system where users register their profiles in order to receive by mail or in their cellular phones special kinds of information.

The extraction of information is automatically made by the system, analyzing Web sites (information sources).

The difference from traditional Web clipping systems is that FastNews only retrieves information relevant to the user's need, that is, it has an intelligent engine that extracts only information parts according to the interest of the user, as defined in his/her profile.

At the moment, the system offers information about news (written in Portuguese), currency conversion and weather forecasting.

FastNews uses Filtering techniques to select relevant news and Information Extraction techniques to find currency conversion values and weather forecasting information (local, day, temperature and rain probability). The system automatically captures information each 10 minutes in predefined Websites.

Profiles are used by the system to determine users' interest. Profiles are formed by one or more queries. Each query is associated to only one kind of information (news, currency conversion or weather forecasting).

In the case of news, user has to determine the keywords to be used by the system to find relevant news. This strategy works like Boolean queries in traditional Web search engines. It is possible to use operators like AND (when two or more words are put together in the same query) and NOT (adding the signal "–" before the word).

The OR operator may be generated by creating different queries. The AND operator leads the system to find texts where all the words are present, and the NOT operator limits the results to only texts where the associated word is nor present.

It is also possible to use expressions or compost words. In this case, the expression must be between quotes (for example, "*information filtering*"). Other possibility is to use radicals for finding word variations (for example, "*work**" is used to represent "*work*", "*working*", "*worker*", etc).

In each query (news, currency conversion or weather forecasting), the user has to mark the information sources where the current query should be performed over. Besides that, the user can state the week days when the query will be performed and the initial and final time. This service allows the user to receive information in only the specified days and time periods. This is especially useful for the currency conversion information (for example, to receive the final value of a day or to not receive the information in holidays or weekends). Time may be useful to allow receiving information only once in a day (for example, the weather forecasting).

Queries work while the user marks them as active. But it is possible to set temporarily a certain query as inactive or even change a query specification.

The information resulting from the selective search (or filtering) is sent to the count registered by the user (an e-mail address or a cellular phone number. Information are sent to cellular phones using Short Message Services (SMS). Counts must be informed by the user in his/her profile. It is allowed to the user to register different counts for receiving different kinds of information.

The information sources are predefined. Currently, there are about 15 different sources, each one dealing with a special kind of news (for example, general about the World and Brazil, sports, economy, computers, etc.).

To extract information from the sources, the system was developed with special templates, one for each source. These templates are designed to "understand" the structure and format of each Web source. The template identifies the region where the desired information appears (Information Filtering) and is able to extract exactly this information and nothing more (Information Extraction).

Templates were generated manually by human experts analyzing the structure and format of each Web source. Each template only works extracting information from the specified source. If the source changes its format or structure, the template may not work appropriately.

The system is under test in the Knowledge Discovery Portal, available at the URL

www.descobertadeconhecimento.com.br.

The history of information sent to the user is registered, allowing users to review messages or to visualize the original Web source.

Additionally, user can control the maximum number of messages to receive per day.

4.1 Predefined sources X New information sources

Traditionally, clipping or information filtering systems use predefined sources, that is, the search is made in sources from an existing list.

Unfortunately, the predefined list never covers all the desired sources, because some relevant information may be only present in other sources. The alternative would be to find new information sources.

In this way, PERKOWITZ ET AL. (1997) presents the ShopBot system, that visits Web sites of CD and Software vendors in order to extract information about products.

The novelty in this system is that the extraction rules are automatically discovered by the system based on analysis of preexisting sites. Using supervised learning strategies, the system receives from human experts a list of sites about some subject and identifies patterns common to all of them (for example, position, keywords, formats and proper names used for refereeing a specific information).

After, the system search through the Web looking for similar sites. When it finds a similar one, it uses its rules for extracting information.

At the moment, FastNews analyzes only predefined information sources (Web sites). An additional functionality is being developed to allow users to enter an URL that they want to monitor.

Using automatic machine learning techniques and some help from the user (supervised learning), the system will be able to automatically extract information from new sources (Web sites).

The system receives from the user an URL and automatically identifies the structure of the related site. After, the system presents this structure for the

user to select the area that he/she wants to monitor. The system stores a representation of this area and periodically monitors the same area in the Web. If a change occurs in this area, the system notifies the corresponding user, presenting the new text.

5 CONCLUSIONS

The great benefit of the proposed system is the automatic capture of relevant information in the Web, freeing the user from having to visit different sites, analyze some headlines and so read texts in order to find relevant information.

The difference from existing clipping systems is that the selection of information is based on the profile of the user. This profile may be created by the own user through the specification of queries.

At the moment, the FastNews system only captures information about news, currency conversion and weather forecasting.

To select news, the system uses Information Retrieval techniques based on Boolean operators, but other facilities are available. This functionality avoids users to receive irrelevant information, as in traditional Web clipping systems.

Information about currency conversion and weather forecasting is captured using Information Extraction techniques, based on the structure of the sites and on HTML tags.

An advantage from other systems is that the user can inform in what day and time he/she wants to receive the information.

A drawback of the system is that the capture of information in Web sites is made using specific templates, one for each kind of information and for each Web source.

Other restriction is that the system only extracts information from predefined sources. An ongoing work will allow users to specify new URLs to the system automatically monitor new information. An alternative being considered is the system automatically finds new information sources in the Web. The system will search through the Web, looking for Web sites similar to the existing ones (predefined in the system).

Other kinds of information are planned to be added to the system, as for example road traffic and stock exchange positions. To do that, it is necessary to create new and specific templates. A future study will evaluate the possibility of automatically extracting other kinds of information.

Currently, the system operates only over sites in Portuguese. A future work will consider other languages.

Other future work will study the use of dynamic profiles, that is, the system will learn about the user's interest, analyzing the news read by the user. Using text mining techniques, it will be possible to infer interesting subjects, not yet defined in the user's profile, and then select other kind of news to send to the user.

REFERENCES

BLOEDORN, Eric et al. (1996) Representational issues in machine learning of user profiles. In: HEARST, Marti; HIRSH, Haym (eds). AAAI Spring Symposium on Machine Learning in Information Access. **Proceedings...** Stanford, Março de 1996. Disponível por WWW em: http://www.ee.umd.edu/medlab/filter/

CHEN, Hsinchum (1994) The vocabulary problem in collaboration. **Computer**, 27(5), p.2-10, May.

COWIE, Jim; LEHNERT, Wendy (1996) Information extraction**. Communications of the ACM**, 39 (1), January.

CROFT, W. Bruce (1995) Machine learning and information retrieval. In: COLT '95 Conference. **Proceedings...** Lake Tahoe, July, 1995 (invited talk) http://www.ee.umd.edu/medlab/filter/

KORFHAGE, Robert R. (1997) **Information storage and retrieval**. EUA: John Wiley & Sons, 1997.

MOSTAFA, J. et al. (1997) A multilevel approach to intelligent information filtering: model, system and evaluation. **ACM Transactions on Information Systems**, v.15, n.4, Outubro de 1997.

OARD, Douglas W.; MARCHIONINI, Gary**.** (1996) **A conceptual framework for text filtering**. Technical Report, University of Maryland. Disponível por WWW em http://www.ee.umd.edu/medlab/filter/ (Maio de 1996).

PERKOWITZ, Mike et al. (1997).. Learning to understand information on the Internet: an example-bases approach. Journal of Intelligent Information Systems, 8, 1997, p.133-153.

SPARCK-JONES, Karen; WILLET, Peter (eds). (1997) **Readings in Information Retrieval**. San Francisco: Morgan Kaufmann, 1997.