A Trend Discovery System for Dynamic Web Content Mining

A. Méndez-Torreblanca[1,2], M. Montes-y-Gómez[1] and A. López-López[1]

[1] Instituto Nacional de Astrofísica, Óptica y Electrónica-INAOE.
Luís Enrique Erro No 1, Tonantzintla, Puebla., 72840. México
amendez@cseg.inaoep.mx, {mmontesg, allopez}@inaoep.mx
[2] Instituto Tecnológico de Puebla-ITP, Av. Tecnológico No 420,
Puebla, Pue., 72000. México.

Abstract. The rapid expansion of the web is causing the constant growth of information, leading to several problems such as an increased difficulty of extracting potentially useful knowledge. *Web content mining* confronts this problem gathering explicit information from different web sites for its access and knowledge discovery. Its current methods focus on analyzing static web sites and cannot deal with constantly changing web sites, such as news sites. In this paper, we propose a method for mining online news sites. This method applies *dynamic* schemes for exploring these web sites and extracting news reports, and uses *domain independent statistical* analysis for trend analysis. The overall method is an application of web mining that goes beyond straightforward news analysis, trying to understand current society interests and to measure the social importance of ongoing events.

Keywords: Web content mining, dynamic crawler, trend discovery, statistical analysis.

1 Introduction

The web is a medium for accessing a great variety of information stored in different parts of the world. The rapid expansion of the web is causing the constant growth of this information, leading to several problems: an increased difficulty of finding relevant information, extracting potentially useful knowledge and learning about consumers or individual users [1]. Web mining is an emerging research area focused on resolving these problems.

Basically, *web mining* is concerned with "the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services" [2]. It is categorized in three areas of interest: web usage mining, web structure mining and web content mining [1]. Web usage mining finds access patterns from web sites. Web structure mining provides structural information about web documents and sites, and web content mining finds useful information from data in the web.

Web content mining considers different kinds of data such as: images, audio, video and texts (e.g. web documents and free texts). For web documents, the mining methods are mainly focused on information extraction and integration (i.e., gathering explicit information from different web sites for its access). These methods are usually based

on simple wrappers that collect structured information [3]. Recently, there are also attempts for using NLP techniques to allow the discovery of previously unknown and potentially useful information from the collected data [4].

In this paper, we describe a complete method for *mining news* from online news sites. This method navigates across these web sites, extracts news reports from them, and analyzes these reports in order to discover interesting news trends.

Basically, the proposed method is adapted for the peculiar characteristics of news. For instance, it applies *dynamic* schemes for the extraction of news reports, and *domain independent statistical* strategies for topic identification and trend analysis. As a whole, our method is an application of web mining that attempts to go beyond straightforward news analysis, trying to understand current society interests and to measure the social importance of ongoing events.

The rest of the paper is organized as follows. Section 2 describes previous work on web content mining and discusses the main limitations of current approaches. Section 3 presents our method for doing a dynamic discovery of news trends. Section 4 shows some experimental results that illustrate the system operation. Finally, section 5 concludes the discussion.

2. Related Work

Web mining is traditionally decomposed in four stages [1]:

- *Resource identification* is the process of retrieving the intended web documents. It is done by web search and metasearch engines, or by crawlers [5,6,7]. These approaches focus on a one-time analysis of web sites and cannot deal with constantly changing web sites, such as news sites where the information is constantly added or modified.
- *Preprocessing* consists of two tasks: selecting interesting data from the downloaded web documents, and transforming this data into a formal representation. Most methods use wrappers for extracting simple data (e.g. proper names, prices, phone numbers, e-mail addresses, etc.) from web documents, and construct tables as formal representations [7,8].
- *Generalization* is the automatic discovery of patterns across multiple web documents. Most methods use data mining techniques for discovering association rules, clusters and classification trees and rules. For instance, Singh et al. [6] proposed a method for detecting association rules that describe the content of a set of scientific online papers; Ghani et al. [7] suggested extracting companies data from the web and constructing classification trees for predicting the growth of the economic sectors; Crimmins and Smeaton [5] concentrated on clustering web sites by their content; Gelbukh et al. [9] proposes a method to categorize documents based on a weighted topic hierarchy; and Alexandrov et al. [10] describes a method for clustering and classifying interdisciplinary documents based on qualitative and quantitative properties.
- *Analysis* involves the validation and interpretation of the mined patterns. The user, sometimes supported by graphical interfaces [5], does this interpretation.

Different to the approaches described above, our method for web content mining considers the analysis of dynamic web sites (i.e., constantly changing web sites). Basically, it proposes several techniques specialized in news analysis.

The proposed method is a complete application of web content mining. It implements a *dynamic crawler* that considers changing news and improves the identification of up to date news. It also applies simple *nlp methods* to identify news topics richer than simple key words. Finally, it integrates some previous ideas about trend discovery in static manual-collected news collections [11] into the web mining process. These ideas use straightforward *statistical metrics* to discover general trends among news reports of different topics.

3. Dynamic news analysis

The proposed method, as any other web mining method, consists of four major stages: resource identification; preprocessing, generalization and analysis (see figure 1). These stages are described in the next four subsections.

3.1 Resource finding

This stage extracts all news reports from a given news site. It is implemented as a crawler that navigates across a web site and continuously extracts the news reports from it. This crawler operates as follows (see Fig. 1.a):

- Downloads the page from the current url (initially this url corresponds to the main page of the news site).
- Filters the downloaded web page, i.e. eliminates the page if it was previously downloaded or belongs to another web site. Also it verifies if the page is a news report and it is not outdated, becoming then a document of interest.
- Analyzes the identified news report. It eliminates irrelevant information such as tags, and stores the content of the news report for its processing. It also identifies and extracts urls for further exploration. These urls are stored in a queue
- Repeats the steps described above until the queue of urls is empty. This condition means that the web site was fully explored.

The whole resource finding process is activated periodically, e.g. daily or twice a day. After this process is completed, the documents obtained constitute a "snapshot" of current events and are subsequently preprocessed and stored. Later on, the user can select a period for analysis.

3.2 Preprocessing

The preprocessing stage transforms incoming news reports into a structured representation. This representation consists of its source information, date, and a formal representation of its contents. For the latter, we reduce the text to a list of keywords or topics. In our experiments, we used a method where the topics are related to noun strings [12].

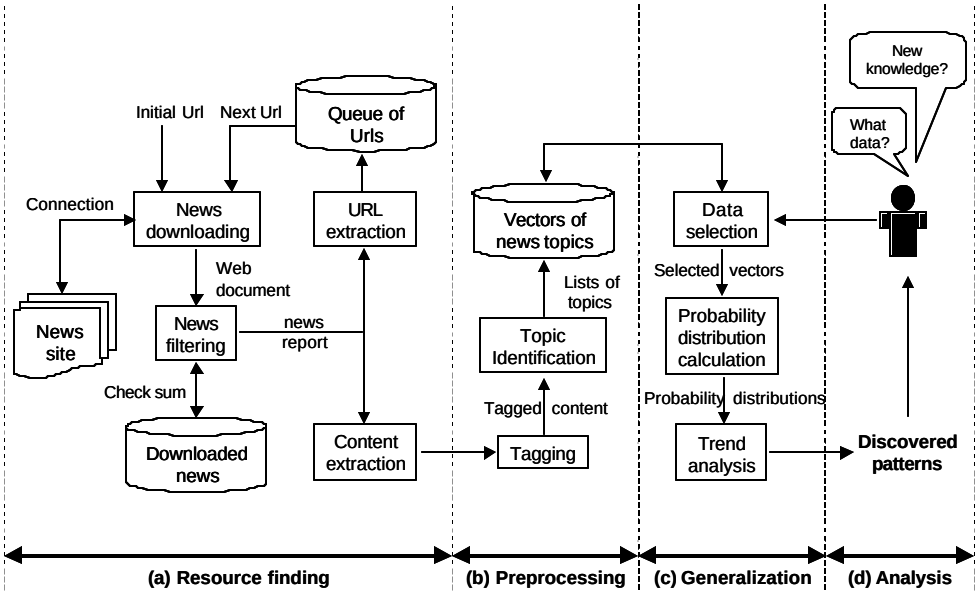

Fig. 1. The process for trend discovery in dynamic web sites

The extraction of the topics of a news report is done in the following steps (see Fig. 1.b):

- The sentences are marked with part-of-speech tags.
- Based on the POS tags, the nouns are identified and joined to form a unique item when appearing in a sequence.
- The most frequent items are selected and inserted into a list of topics.

3.3 Generalization

The goal of this stage is to discover interesting trends among news topics. It considers two tasks: the construction of the topic distributions, and the analysis of trends (see figure 1.c)[1]. The following subsections give details on the two tasks.

3.3.1 Construction of topic distributions

Based on the formal representation of news, a frequency f_k^i is assigned to each topic discussed in the period of interest (i.e. the time span indicated by the user). It is calculated as the number of the news reports in the period i that mention the topic k. Then, using these frequencies, a probability distribution $D_i = \{p_k^i\}$ of the news topics in the

[1] More detail on our trend discovery method can be found in [11].

period i is constructed, where $p_k^i = f_k^i \Big/ \sum_{j=1}^{n} f_j^i$ expresses the probability of occurrence of the topic k in the period i, and n indicates the number of topics cited in the whole period i.

3.3.2 Trend analysis

We discover trends by comparing the probability distributions $D_i = \{ p_k^i \}$ of the news topics for two given periods $i = 1, 2$. Since we are interested in the change regardless of the direction and a reference information source, we compare the distributions by the measure C_c expressed as the quotient of the change area and the maximal area. This measure reflects an overall trend and does not measures individual proportions of change of each individual factor.

$$
\begin{aligned}
C_c &= \frac{A_c}{A_m} && \text{change coefficient, where:} \\
A_c &= \sum_{k=1}^{n} d_k && \text{change area} \\
A_m &= \sum_{k=1}^{n} \max\left(p_k^1, p_k^2\right) && \text{maximal area} \\
d_k &= \left| p_k^1 - p_k^2 \right| && \text{individual topic change}
\end{aligned}
\qquad (1)
$$

If the change coefficient between the two probability distributions tends to 1, then there exist a considerable change between the news topics of the two periods. On the contrary, if the change coefficient tends to 0, then we can conclude that news of both periods are similar.

For the case of a change trend, it is important to identify the news topics with a major contribution to this trend. We call this topics change factors, and define them as those with a change noticeably greater than the typical change. Let d_m be a "typical" value of d_k (see below) and d_s be a measure of the "width" of the distribution. Then a topic k for which $d_k > d_m + (C \times d_s)$ is identified as a change factor. The tuning of the constant C determines the criterion used to identify an individual change as noticeable.

$$
\begin{aligned}
d_\mu &= \frac{1}{n} \sum_{k=1}^{n} d_k && \text{average change} \\
d_s &= \sqrt{\frac{1}{n} \sum_{k=1}^{n} \left(d_k - d_\mu\right)^2} && \text{standard deviation of the change}
\end{aligned}
\qquad (2)
$$

On the other hand, for a stability trend, we try to detect the most important and popular topics in the two periods of interest. We call this topics stability factors, and define them as the set of topics that remain almost stable and maintain significant level of importance in both periods. Thus, a topic k is a stability factor if $d_k < d_m - (C \times d_s)$

and $p_k^i > p_m^i$ for both periods i = 1, 2. Here, $p_m^i = \sum_{j=1}^{n} p_j^i \Big/ n$ and C is a constant that establishes the criterion to identify an individual topic as sufficiently stable.

3.4 Analysis

In this stage, the user interacts with the system as follows:

- The user selects the timeframe of interest and establishes the parameters that control the generalization process.
- Then, the user analyzes the patterns discovered by the system in the generalization stage.
- If the discovered patterns are not interesting for the user, he can repeat this process selecting other period and parameters until he is satisfied with the results.

4. Experimental results

The main goal of our system is to analyze current society interests and ongoing events by detecting news trends from online news sites. In this section, we describe the results on the analysis of *The News*, a Mexican news site in English language. This analysis consist of two main processes:

1. Extracting news reports from the news site and transforming them into a formal representation (this process is activated automatically every day or twice a day).
2. Discovering the main news trends for a given time span (this process is triggered by a user request).

In the first step, we achieve the following results:

The crawler downloaded, on average, 350 web pages each day. From these, only 130 pages described current news and were selected for further analysis; the rest were eliminated at the filtering stage. The figure 2 shows some statistics on this stage. Basi-

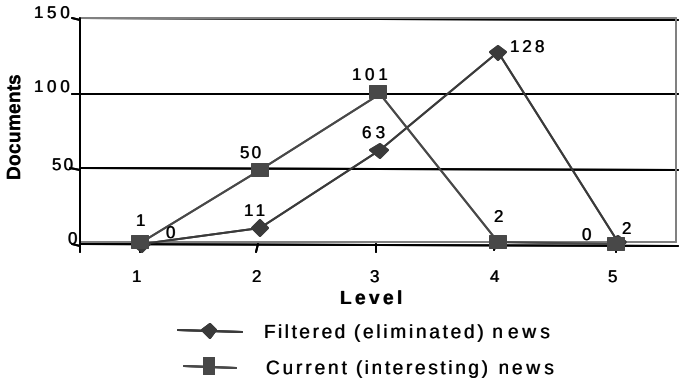

Fig. 2. Data about the exploration of "*The News*" site

The News - *Government* identifies areas of *strife* in *Oaxaca*.
The *government* has identified dozens of areas of conflict, including some designated as red alert zones, in the southern state of *Oaxaca*, where land disputes have raged for decades.
One long-standing *land dispute* between indigenous and peasant communities came to a head last Friday, with the *massacre* of 26 Indians in the *village* of Santiago Xochiltepec.
Xochitl *Galvez*, the director of the president's Indigenous Peoples' Development Office, said the federal *government* needs to conduct an exhaustive analysis of the region's problems, which also affect other states in southern Mexico, such as Chiapas and Guerrero……..

{Government, strife, Oaxaca, government, land-disputes, massacre, Galvez, Wednesday, Authorities, areas, PGR,…}

Fig. 3. Example of a formal representation of a news report

cally, it shows the number of web documents downloaded from each level of the given news site (in accordance with a breadth-first search). It is important to mention that the analysis of the first five levels of the given site (i.e. the complete resource finding process) took approximately 18 minutes.

The selected news pages were preprocessed and transformed to formal representation (i.e., a list of keywords or topics). We obtained lists of approximately 20 topics for each news page. The figure 3 shows a fragment of a news report and its corresponding formal representation. The obtained topic lists were stored for later trend discovery.

In the second step, we analyzed the news corresponding to the last week of May and the first of June, i.e., from May 27 to June 11, 2002. These news reports considered 663 topics as a whole. Their analysis returned the results showed in the figure 4.

Since C_c = 0.591, we concluded that there was a light global change trend between these periods (i.e., there were enough differences on the news topics between the last week of May and the first of June). The main change factors discovered are: *World-Cup*, *massacre* and *Oaxaca* as emerging news topics, and *Bush*, *France* and *Uribe* as the disappearing ones. The change on these topics is showed in figure 4.

These results indicates that at the end of May the attention in Mexico focused on international events such as the *visit of president Bush to France*, and the *presidential elections at Colombia*. On the contrary, at the beginning of June, the main events concentrated on internal matters, such as the *murder of natives at Oaxaca State*.

Additionally, these results confirmed the obvious global attraction generated by *Soccer World-Cup Championship* that started in June.

5. Conclusions and future work

In this paper, we present a trend discovery system for dynamic web content mining. This system extends the capabilities of traditional web content mining approaches in order to analyze constantly changing web sites containing information about multiple topics (such as online news sites). Some important features of this system are the following:

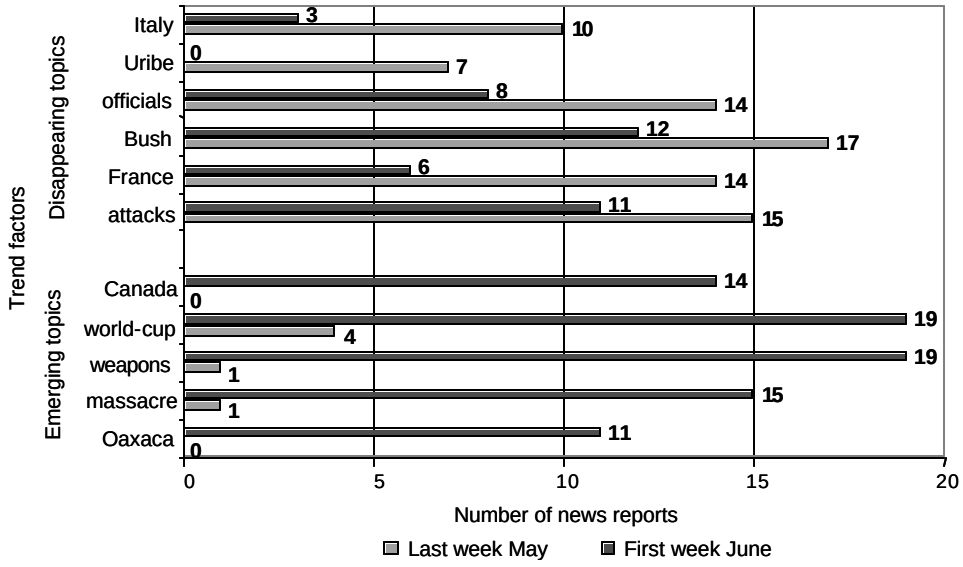

Fig. 4. Main change factors for the selected timeframe

- It uses a dynamic crawler for resource finding. This crawler permanently monitors a specific news site, and continually downloads the latest news reports.
- It applies simple nlp techniques (e.g. part-of-speech tagging) for preprocessing the news reports. These techniques allow the extraction of meaningful topics as news representations, but preserve the domain independency for analyzing news about multiple topics.
- It uses straight statistical measures in the discovery stage. These measures not only consider the detection of general trends, but also the identification of their factors (i.e. the topics contributing to these trends).
- It allows user interaction. For instance, the user selects the timeframe for analysis and defines some parameters that lead to the discovery of news trends.

As future work we plan to focus on the following tasks:

- Adapt the crawler for locating news in Spanish.
- Improve the document representations. Mainly, we plan to construct logic predicates or conceptual graphs as formal text representations.
- Extend the generalization methods in order to discover other kind of patterns such as: associations, clusters and deviations.
- Specialize the whole process to some domain such as economy or politics.
- Develop a graphical interface for supporting the user to interpret the discovered patterns.

Finally, it is important to point out that the discovery of this kind of news trends helps to interpret the society interests and uncover hidden information about the relationships between the events in social life.

References

[1] R. Kosala and H. Blockeel. Web mining research: a survey. SIG KDD Explorations, Vol. 2, pp. 1-15, July 2000.

[2] O. Etzioni. The World Wide Web: Quagmire or Gold Mine? Communications of the ACM, Vol.39, No.11, pp. 65-68. Nov. 1996.

[3] N. Kushmerick Ed. Adaptive Text Extraction and Mining (Working Notes). Seventeenth International Joint Conference on Artificial Intelligent (IJCAI-2001). Seattle, Washington, 2001.

[4] S. Soderland. Learning information extraction rules for semi-structured and free text. Machine Learning, Vol. 34, no 1-3, pp. 233-272, 1999.

[5] F. Crimmins, A. F. Smeaton, T.Dkaki and J. Mothe. TetraFusion: information discovery on the Internet. Journal of IEEExpert, pp 55-62, July 1999.

[6] L.Singh, B. Chen, R. Haight and K. Scheuermann. An Algorithm for Constrained Association Rule Mining in Semi-structured Data. Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 99), pp 148-158r, China, 1999.

[7] R. Ghani et al. Data mining on symbolic knowledge extracted from the web. Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD-2000), Workshop on Text Mining, pp 29-36, Boston, MA, August 2000.

[8] N. Kushmerick. Gleaning the web. IEEE Intelligent Systems. Vol. 14, No. 2, pp. 20-22, 1999.

[9] A. Gelbukh, G. Sidorov, A. Guzman-Arenas. Use of a weighted topic hierarchy for document classification. In Vaclav Matousek et al. (Eds.). Text, Speech and Dialogue. Lecture Notes in Artificial Intelligence, N 1692, Springer-Verlag, 1999.

[10] M. Alexandrov, A. Gelbukh, and P. Makagonov. Evaluation of Thematic Structure of Multidisciplinary Documents. Proc. NLIS-2000, 2nd International Workshop on Natural Language and Information Systems, IEEE Computer Society Press, 2000.

[11] M. Montes-y-Gómez, A. Gelbukh and A. López-López. Mining the News: Trends, Associations, and Deviations. Computación y Sistemas, Vol. 5 No. 1, pp. 14-24, Julio-Septiembre 2001.

[12] L. Gay and W.B. Croft. Interpreting Nominal Compounds for Information Retrieval. Information Processing and Management, pp 21-38. 1990.