A STUDY ON PATTERN DISCOVERY SYSTEM FOR WEB CONTENT MINING

M.Vasavi

Abstract—

There is an explosive growth of information in the World Wide Web thus posing a challenge to Web users to extract essential knowledge from the Web. Web data Extraction is the process of extracting the information that users are interested in, from Semi-structured or unstructured web pages and saving the information as the XML document or relationship model. In this 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. 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.

Keywords- Web Data Extraction, digital archives, Cloud computing, text mining, newspapers.

I. INTRODUCTION

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 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". Web content mining is related but different from data mining and text mining. It is related to data mining because many data mining techniques can be applied in Web content mining. It is related to text mining because much of the web contents are texts. However, it is also quite different from data mining because Web data are mainly semi-structured and/or unstructured, while data mining deals primarily with structured data. Web content mining is also different from text mining because of the semi-structure nature of the Web, while text mining focuses on unstructured texts. 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. Furthermore, online news is a special type of public information which has exclusive characteristics. These characteristics contribute news engines tasks such as discovering, collecting and searching to be different with similar tasks in traditional web search engines. The existence of numerous reliable news sources (high trust) and fast news update are the two most important differences. News engines provide many services and contain various tasks, the quality of each task can affect the other tasks quality. The most important tasks are:

- Collecting News
- News Retrieval
- Categorizing Search Result
- Summarization
- Automatic pattern discovery

II. COLLECTING NEWS AND NEWS RETRIEVAL

News collection and retrieval is the process of retrieving the intended web documents. It is done by web search and meta-search engines, or by crawlers. 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.

- Downloads the page from the current url (initially this url corresponds to the main page of the news site).
- 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 extraction of the topics of a news report is done in the following steps

- 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.

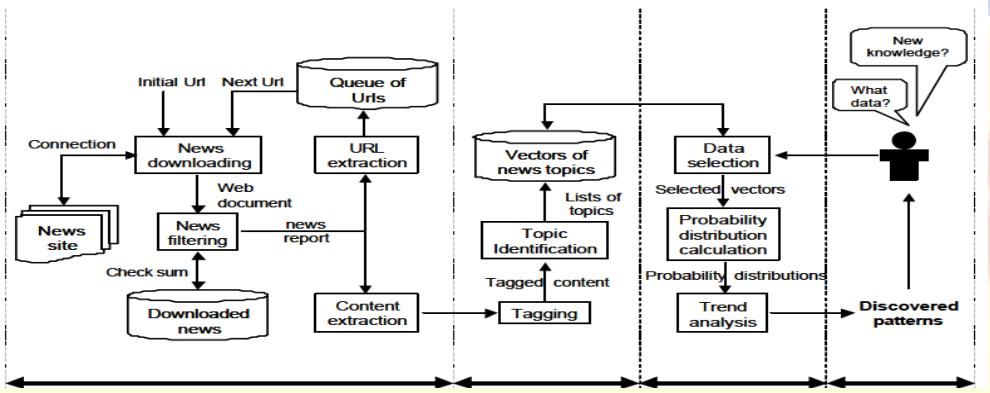

Fig:1 Architecture of the System

III. CATEGORIZING SEARCH RESULT

Classification and clustering are the main methods of search result categorization. Clustering on the results of search engines has obvious differences with traditional text clustering. One of these differences is the existence of links between web pages. Also, Due to the fact that search result

clustering is an online process, fast computation is needed. The final difference is that clustering is based on small snippets instead of whole document. According to good characteristics for clustering on the search results are defined as bellows:

- No necessity for all pages to be clustered
- Cluster Overlapping
- Incremental clustering

The process involved the extracting of the n most frequently occurring words in each article. If *S1* and *S2* are the sets which includes these extracted words of two news, the similarity between two news is defined as bellow:

$$Sim(n_1, n_2) = \frac{|s_1 \cap s_2|}{n}$$

$$Dissim(n_1, n_2) = 1 - Sim|(n_1, n_2)$$

Where $Dissim(n_1,n_2)$ is the distance criterion of two news. By virtue of these criteria, K-Nearest Neighbour is done on news. Single-link algorithm is also used in that paper, but the combination of both algorithms leads to the better results.

IV. AUTOMATIC PATTERN DISCOVERY

We discover patterns by comparing the probability distributions $P_{D_i} = \{p_k^i\}$ of the news topics for two given periods i = 1, 2 where $p_k^i = f_k^i / \sum_{j=1}^{n} f_j^i$ expresses the probability of occurrenceof the topic k in the period i, n indicates the number of topics cited in the wholeperiod i and 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. Since we are interested in the change regardless of the direction and a reference information source, we compare the distributions by the measure M_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.

$M_C = \frac{R_C}{R_M}$ Change coefficient, where:

$R_c = \sum_{k=1}^{n} d_k$ Change area

$R_M = \sum_{k=1}^{n} \max(p_k^1, p_k^2)$ Maximal area

$d_k = | p_k^1 - p_k^2 |$ Individual topic change

If the change coefficient between the two probability distributions tends to 1, then there exists 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_α be a "typical" value of d_k (see below) and d_β be a measure of the "width" of the distribution. Then a topic k for which $d_k > d_\alpha + (C \times d_\beta)$ is identified as a change factor. The tuning of the constant C determines the criterion used to identify an individual change as noticeable.

$d_\alpha = \frac{1}{n} \sum_{k=1}^{n} d_k$ Average change

$d_s = \sqrt{\frac{1}{n} \sum_{k}^{n} (d_k - d_\alpha)^2}$ Standard deviation of the change

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 $d_k < d_\alpha - (C \times d_\beta)$ and $p_k^i > p_\alpha^i$ for both periods i = 1, 2. Here, $p_\alpha^i = \sum_{j=1}^{n} p_j^i / n$ and C is a constant that establishes the criterion to identify an individual topic as sufficiently stable.

V. 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.

VI. RESULTS

The main goal of our system is to analyze current society interests and ongoing events by detecting news trends from online news sites. We analyse the insights of 2104 AP elections using twitter data. Many interesting facts about different political parties were revealed basing on the opinions of the people. This type of analysis gives reliable, efficient and effective results for different situations.

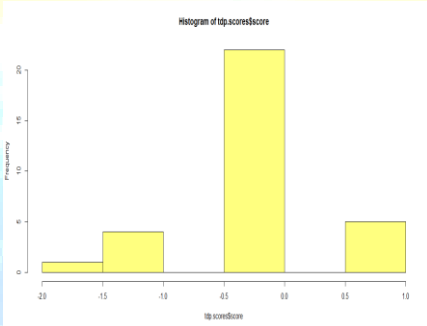

Fig:2 Histrogram for TDP favored twits

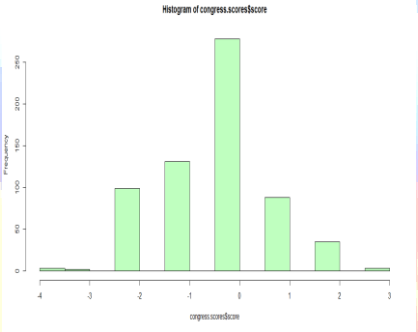

Fig:3 Histrogram for Congress favored twits

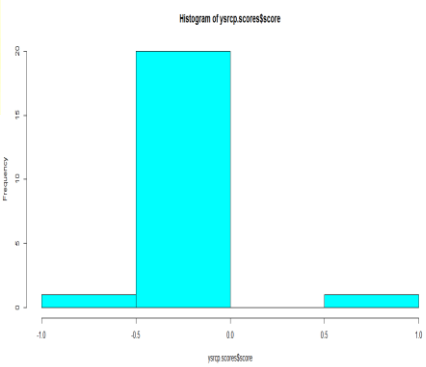

Fig 4: Histrogram for YSR Congress favored twits

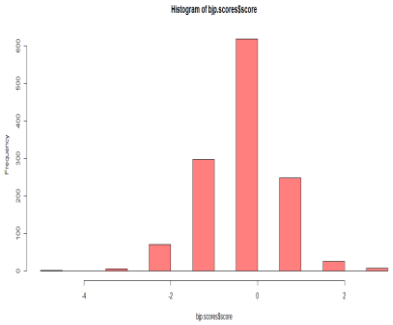

Fig 5: Histrogram for BJP favored twits

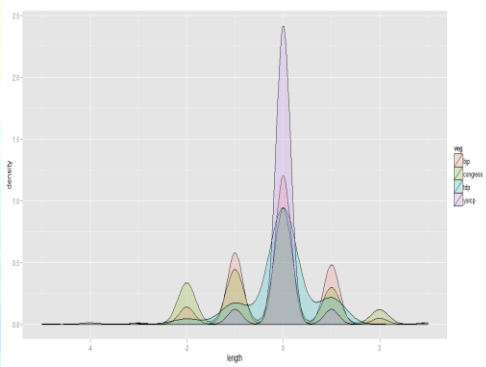

Fig 6: A graph which compares all the four political parties

VII. ANALYSIS

All the above graphs are created based on the data grabbed from twitter. Several facts are revealed basing on the data from a large number of opinions.

1. The analysis says that the party YSRCP has a neutral feed back
2. Both the parties BJP and TDP have a good positive score and most of the people support these two parties.
3. The party congress has a high negative score and most of the users oppose this party and users feel moody about this party.

In the state TDP has a good following from the people and inn central BJP has more positive feedback than any other political parties.

VIII. CONCLUSION

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. 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.

IX. REFERENCES

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

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

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

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

[5] D. R. Radev, S. Blair-Goldensohn, Z. Zhang, R. S. Raghavan,"Interactive, domain-independent identification and summarization of topically related news articles".

[6] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz, "Analysis of A Very Large Web Search Engine Query Log", SIGIR Forum, 33(1), 1999.

[7] G. M. D. Corso, A. Gulli, F. Romani, "Ranking a stream of news", In: WWW '05: Proceedings of the 14th international conference on World Wide Web, New York, NY, USA, ACM Press (2005) 97–106.

[8] M. Atallah, R. Gwadera, W. Szpankowski, "Detection of significant sets of episodes in event sequences". icdm 00 (2004) 3–10

[9] N. Maria, M. J. Silva, "*Theme-based retrieval of Web news*", Lecture Notes in Computer Science 1997

[10] N. A. Shah, E. M. ElBahesh, "*Topic-based clustering of news articles*", In: ACM-SE 42: Proceedings of the 42nd annual Southeast regional conference, New York, NY, USA, ACM Press (2004) 412–413