International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:6, No:12, 2012

Sentiment Analysis: Popularity of Candidates for the President of the United States

Radek Malinský, and Ivan Jelínek

***Abstract*—**This article deals with the popularity of candidates for the president of the United States of America. The popularity is assessed according to public comments on the Web 2.0. Social networking, blogging and online forums (collectively Web 2.0) are for common Internet users the easiest way to share their personal opinions, thoughts, and ideas with the entire world. However, the web content diversity, variety of technologies and website structure differences, all of these make the Web 2.0 a network of heterogeneous data, where things are difficult to find for common users. The introductory part of the article describes methodology for gathering and processing data from Web 2.0. The next part of the article is focused on the evaluation and content analysis of obtained information, which write about presidential candidates.

***Keywords*—**Sentiment Analysis, Web 2.0, Webometrics.

I. Introduction

THIS article deals with the popularity of candidates for the president of the United States of America. The president of the United States is elected by the Electoral College every four years. The Electoral College consists of representatives from each state of the United States and by the District of Columbia and it is elected directly by the people of the United States. The presidential election is the most important and observed event all over the world. This demonstrates that not only mass media are interested in, but especially large number of election-related news has been published on the Internet.

According to Internet World Statistics [1], almost 80% of the population of the United States is currently internet users and more than 50% of them visit social networking sites like Facebook, Twitter, etc. Social networking, blogging and online forums (collectively Web 2.0) are for common Internet users the easiest way to share their personal opinions, thoughts, and ideas with the entire world. Thanks to the openness of the Web 2.0 and its expansion throughout the Internet, the number of Web 2.0 users and the number of published articles has been constantly growing at a very high rate. Web 2.0 can serve as a data source for social science research because it contains vast amount of information from many different users.

On the other hand, the web content diversity, variety of technologies and website structure differences, all of these make the Web 2.0 a network of heterogeneous data, where things are difficult to find for common users. The Semantic Web, introduced by Tim Berners-Lee [2] already in 2001, it could be the way out of this situation. On the Semantic Web, there is information structured and organized according to standardized rules, which makes it easier to find. However, the implementation of the Semantic Web is too complicated and with few exceptions it is not almost applicable over the Internet.

It is necessary to design suitable methods, which would reflect the semantic content of pages in the better way and allow us quick and easy searches.

R. Malinský is with the Department of Computer Science and Engineering, Faculty of Electrical Engineering, Czech Technical University in Prague, Karlovo náměstí 13, 121 35 Prague, Czech Republic, (e-mail: malinrad@fel.cvut.cz).

I. Jelínek is with the Department of Computer Science and Engineering, Faculty of Electrical Engineering, Czech Technical University in Prague, Karlovo náměstí 13, 121 35 Prague, Czech Republic, (e-mail: jelinek@fel.cvut.cz).

II. Related Work

Webometrics is a scientific discipline that, among others, presents an approach to information on the Web. Webometrics is based on informetric and bibliometric methods [3]; however, the information sources that are studied by webometrics are web documents. Current search engines are based on the webometric approaches [3], [4], especially on hyperlink analysis. Google PageRank is a good example of this approach.

PageRank [4] used by the Google internet search engine is the most reliable and effective link analysis algorithm. The algorithm evaluates the relevance of search result pages by the number of in-links and by the quality of source of these in-links. The quality of in-links has a much higher significance than the number of in-links. This strategy especially excels in queries to a specific case, but when you enter a complex query, it returns a large number of irrelevant links and does not reflect the semantic content of single pages. Therefore, this strategy is used primarily to determine the impact of a given web page.

One of the options for more accurate comprehension of semantic information is to use a sophisticated analysis of sentences using mathematical and statistical methods and linguistic analysis of a text (called Sentiment Analysis [5], [6], [7]). Sentiment Analysis or Opinion Mining enables us to automatically detect opinions from structured but also unstructured data. The main goal of the sentiment analysis is to identify positive/negative polarity of a text and recognize an attitude of a writer to specific topic [9], [10].

III. Objectives of Study

The web content variety, diversity of technologies and website structure differences, all of these make the Web 2.0 a

network of heterogeneous data, where things are difficult to find for common users. It is necessary to design suitable metric for such data, which would reflect the semantic content of pages in the better way. One of the options for more accurate comprehension of semantic information is to use a sophisticated analysis of sentences called Sentiment Analysis. The knowledge gained will be useful for algorithm design to facilitate user access to the information on the web, and also to obtain the public opinion on specific issues.

The proposed model will be used to manage the information about the candidates the President of the United States. Gathered data will be analyzed and the analysis output will be displayed on a website with a popularity graph. The graph will show a popularity of both candidates over the time and offers a comparison between the candidates for each day. The graph will be automatically updated every 12 hours.

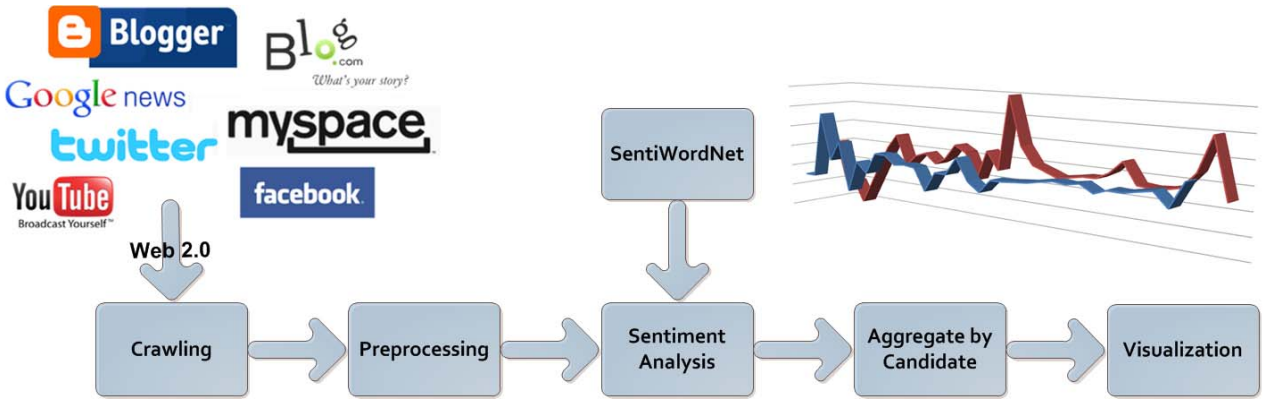

Fig. 1 System for the Visualization of Presidential Candidates' Popularity

IV. Methodology of Study

We have proposed a novel theoretical model [5] for gathering and processing data from Web 2.0 (Fig. 1). The model builds on webometrics [9], [10] and starts from the idea that almost any text can be machine-recognized. This idea is supported by current research in sentiment analysis [6], [7], [11], which aims at sophisticated analysis of sentences using mathematical and statistical methods and linguistic analysis of text. There are several essential parts in the model:

- *Crawling* – Crawler, an automated program, follows links on the Web 2.0 and stores a key content of all visited pages. The crawler intelligently selects links to sites to visit. Only links with high impact, related to the search topic, which are part of the main content of crawled pages, are selected. The content of crawled pages is analyzed and the important parts from them are stored for further processing.
- *Preprocessing* - In this phase, the all stored parts are scanned for the search keywords ("Barack Obama", "Obama", "Mitt Romney", "Romney") and the surroundings of each searched expression have been recognized and a list of sentences for the keyword have created. Every sentence in the list is tagged using Stanford POS Tagger [12]. The tagger assigns parts of speech to each word in a sentence, such as noun, verb, adjective, etc. and it predicts the part-of-speech even for an unknown word. For processing in the next phase, the words are divided into the four part-of-speech categories: adjective, noun, adverb, and verb. For more accurate recognition of words in the fourth phase, the plural words are converted to singular. The same POS Tagger was used, e.g. to enrich textbooks produced from India, which are not written well and they often lack adequate coverage of important concepts [13].
- *Sentiment Analysis* - determines the polarity of tagged word and evaluates sentences for each day for each presidential candidate. For the evaluation there are used lexicon-based methods, which are based on SentiWordNet [14], [15]. SentiWordNet assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity. The evaluation of searched expression is being performed according to our defined rules [8] for each day.
- *Aggregate by Candidate* - In this phase, all the evaluated sentences are split into two groups according candidates Obama and Romney. The individual sentences are then normalized by the number of positive and negative expressions for each day.
- *Visualization* - We have designed a website that shows the popularity graph of the candidates and the overall result of the analysis. List of analyzed articles and a breakdown most frequently positive / negative sentiments for each day are included in the analysis. The popularity graph shows a popularity of both candidates over the time and offers a comparison between the candidates for each day. Web 2.0 content is real-time processed and all the information on the website is automatically updated every 12 hours.

V. Data Analysis

One of the outputs from proposed system is showed in Fig. 2. In the picture is a graph, which represents evaluated chronological summary of both presidential candidates, Barack Obama and Mitt Romney in 15 days. The x-axis of the

graph represents the published date and y-axis shows the polarity of candidates. Positive values of the y-axis represent positive evaluation of the candidate. Negative values of the y-axis represent negative evaluation. As it seen, the polarities of both candidates are relatively similar all the time. However, Mitt Romney popularity began to change rapidly on the 4rd of October. One day before the deviation, there was the first presidential debate between candidates and the graph shows that Mitt Romney was more positively evaluated in the public eye. However, the evaluation of Mitt Romney was gradually coming back to previous state in the following days.

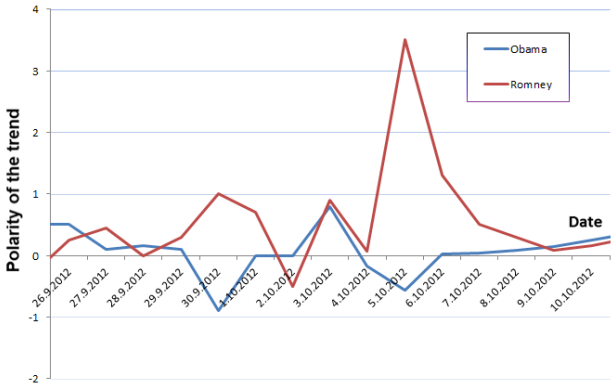

Fig. 2 Evaluated chronological summary of the popularity of presidential candidates Barack Obama and Mitt Romney

VI. Conclusion

We proposed and implemented system for gathering and processing data about candidates for the president of the United States from Web 2.0. Web 2.0 is a complex network of heterogeneous data, where things are difficult to find for common users. Our system provides an effective obtaining of relevant information from the web. The System builds on current research in sentiment analysis and allows us to detect opinions from structured but also unstructured data.

Gathered opinions are further analyzed using mathematical and statistical methods to recognize positive/negative polarity of each candidate for every day. The final results are shown as a graph, where it is possible to compare the popularity of both candidates for each day.

Acknowledgment

This research has been supported by MSMT under research program No. 6840770014. This research has been supported by the Grant Agency of the CTU in Prague, grant No. SGS12/149/OHK3/2T/13.

References

[1] Internet World Stats - Usage and Population Statistics, “Internet usage statistics.” [Online]. Available: http://www.internetworldstats.com/stats.htm
[2] T. Berners-Lee, J. Hendler, and O. Lassila, “The semantic web”, Scientific American, May 2001.
[3] Thelwall, M., “Bibliometrics to Webometrics”, Journal of Information Science, 34(4), 605, 2008.
[4] Langville, A.N., Meyer, C.D., “Google's PageRank and Beyond: The Science of Search Engine Rankings”, Princeton University Press, 2006, ISBN 978-0691122021
[5] R. Malinský and I. Jelínek, “Improvements of webometrics by using sentiment analysis for better accessibility of the web,” in Current Trends in Web Engineering, ser. Lecture Notes in Computer Science, F. Daniel and F. Facca, Eds. Springer Berlin / Heidelberg, vol. 6385, pp. 581–586. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-16985-4 59
[6] M. Potthast and S. Becker, “Opinion summarization of web comments,” in Advances in Information Retrieval, ser. Lecture Notes in Computer Science, C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. Rüger, and K. van Rijsbergen, Eds. Springer Berlin / Heidelberg, 2010, vol. 5993, pp. 668–669, 10.1007/978-3-642-12275-0 73. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-12275-0 73
[7] R. Prabowo and M. Thelwall, “Sentiment analysis: A combined approach,” Journal of Informetrics, vol. 3, no. 2, pp. 143–157, 2009.
[8] R. Malinský and I. Jelínek, “A Novel Web Metric for the Evaluation of Internet Trends”, Proceedings of World Academy of Science, Engineering and Technology. 2011, vol. 7, no. 81, p. 504-507. ISSN 2010-376X.
[9] I. Aguillo, J. Ortega, M. Fern´andez, and A. Utrilla, “Indicators for a webometric ranking of open access repositories,” Scientometrics, vol. 82, pp. 477–486, 2010, 10.1007/s11192-010-0183-y. [Online]. Available: http://dx.doi.org/10.1007/s11192-010-0183-y
[10] M. Thelwall, “Introduction to webometrics: Quantitative web research for the social sciences.” San Rafael, CA : Morgan & Claypool, 2009.
[11] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1–135, Jan. 2008.
[12] K. Toutanova, D. Klein, C. Manning, and Y. Singer, “Feature-rich part-of-speech tagging with a cyclic dependency network,” in HLT-NAACL, 2003, pp. 252–259.
[13] R. Agrawal, S. Gollapudi, K. Kenthapadi, N. Srivastava, and R. Velu, “Enriching textbooks through data mining,” in Proceedings of the First ACM Symposium on Computing for Development, ser. ACM DEV ’10. New York, NY, USA: ACM, 2010, pp. 19:1–19:9. [Online]. Available: http://doi.acm.org/10.1145/1926180.1926204
[14] S. Baccianella, A. Esuli, and F. Sebastiani, “Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining,” in Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC’10), N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds. Valletta, Malta: European Language Resources Association (ELRA), may 2010.
[15] A. Esuli and F. Sebastiani, “Sentiwordnet: A publicly available lexical resource for opinion mining,” in In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC’06), 2006, pp. 417–422.