A Novel Web Metric for the Evaluation of Internet Trends

Radek Malinský and Ivan Jelínek

***Abstract*—**Web 2.0 (social networking, blogging and online forums) can serve as a data source for social science research because it contains vast amount of information from many different users. The volume of that information has been growing at a very high rate and becoming a network of heterogeneous data; this makes things difficult to find and is therefore not almost useful. We have proposed a novel theoretical model for gathering and processing data from Web 2.0, which would reflect semantic content of web pages in better way. This article deals with the analysis part of the model and its usage for content analysis of blogs. The introductory part of the article describes methodology for the gathering and processing data from blogs. The next part of the article is focused on the evaluation and content analysis of blogs, which write about specific trend.

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

I. Introduction

WEB 2.0 (social networking, blogging and online forums) creates vast amount of comments on various topics from many different users. Thanks to that, Web 2.0 can serve as a data source for social science research [1]. According to Internet World Statistics [2], approximately 60% of the population in Europe and over 70% of the populations in North America are currently internet users. The number of Internet users has been constantly growing and Web 2.0 has been becoming a popular tool for finding information [3], [4].

Nevertheless, the volume of information on the web has been growing at a very high rate and becoming a network of heterogeneous data, this makes thing difficult to find and is therefore not almost useful. It is necessary to design suitable metric for such volume of information, 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 [5]. 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.

II. Related Work

We proposed a novel theoretical model [6] for gathering and processing data from Web 2.0, which we have further improved for a more effective obtaining of relevant information from the web (Fig. 1). The model builds on webometrics [7], [8] and starts from the idea that almost any text can be machine-recognized. This idea is supported by current research in sentiment analysis [5], [9], [10], 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:

- *The Crawler* - an automated program, which follows every link on the site and creates a copy of all visited pages.
- *The WWW Analysis* - an algorithm for analyzing crawled pages and storing important information from them.
- The Index - a repository for analyzed web pages, which returns a list of the result pages in a correlation to user’s query.
- *The Query Analysis* - an algorithm for parsing all the words in the search query into a form that the index can understand.

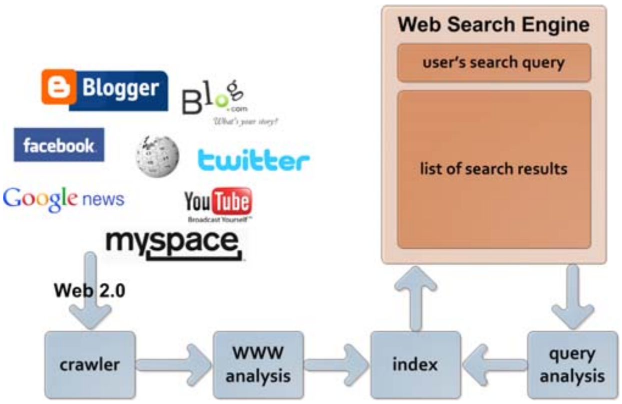

Fig. 1 Novel theoretical model for gathering and processing data from Web 2.0

Functionality of the model is similar to a typical web search engine. However, there are differences in analysis parts. The model provides complete content analysis of crawled pages. The analysis is performed in three phases:

1) *Webometric Web Mention Analysis* determines some basic information about the document such as the document type, geographic spread, and type of organization that is interested in the document, etc. and chooses keywords or entire phrases which represent it.

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).

2) *Sentiment Analysis* determines the polarity and impression of the text and evaluates the selected key words and phrases.

3) *Webometric Hyperlink Analysis* determines the impact of an analyzed document.

Original webometric techniques [7], [8] improved searching and provided trend detection. However, they are not able to distinguish a polarity of the text and its semantic meaning. Extension of the webometric techniques of sentiment analysis methods leads up to gaining insights into a public opinion with respect to some topic and to a better machine understanding of a text. Better understanding of the text on the web site could have a significant impact on the quality of site evaluation.

III. Objectives of Study

We assumed that the designed theoretical model [6] will reduce the irrelevant web search result and thereby facilitate user's access to the information on the web. The experimental system for gathering and processing data from blogs has been created and implemented to verify our theoretical assumptions. The study described in this paper is focused on the content analysis part of the system. Blogs are used as a data source for this research.

The scope of blog topics includes the range from the personal diaries through the official business news up to the political campaigns. The millions of people post information about events around them and they also share opinions on specific issues, e.g. political situation, travel information, technology review, gossip about celebrities, etc.

IV. Methodology of Study

This section describes the auxiliary tools for the research and the process of gathering and processing data from blogs.

A. Data Source and Auxiliary Tools

Hot Searcher

Hot Searcher[1] is the part of Google Trends, which reflects what people are searching for on the Internet. Google trends algorithm analyses web searches performed on Google search engine and provides the list of hot searches, which deviate the most from their historic traffic pattern. The list contains 10 fastest-rising search queries (in descending order) in the United States for each day.

Blog Pulse

BlogPulse[2] is an automated trend discovery system for blogs, which reflects what people are posting on the Internet. BlogPulse collects data from blogs, creates a full-text search index and provides a chronological summary of daily volume of blog post matching a trend. The service indexes over 170 million blogs and it increases approximately 90,000 blogs every day.

B. Approach to the Data

Proposed system (Fig. 2) provides complete content analysis of crawled pages. The analysis is performed in four phases.

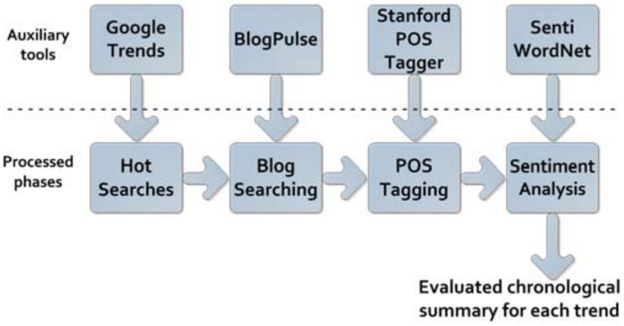

Fig. 2 Experimental system for gathering and processing data from blogs.

In the first phase, **Hot Searches**, hot searches are retrieved from Google Trends over a given period of time. Google Trends monitors searching in the United States, however, it does not matter because our research is focused on English-language sites only.

In the second phase, **Blog Searching**, for each of the detected multi-word trend, Boolean searches are generated to match relevant post. For example the expression "Ottawa earthquake" are restructured to "ottawa AND earthquake", however, names and specific expressions, e.g. "bill gates" have been stayed the same and are searched as the exact phrase. Each expression is searched in BlogPulse search engine within a specified interval e.g. 30 days before and 30 day after the trend has been detected (around the day of the trend deviation). The BlogPulse creates a chronological summary of daily volume of blogs for each trend.

In the third phase, **Part-of-Speech Tagging**, the surroundings of each searched expression from the chronological summary have been recognized and a list of sentences for the trend have created. Every sentence in the list is tagged using Stanford POS Tagger [11]. 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 fourth phase, the words are divided into the four part-of-speech categories: adjective, noun, adverb, verb (Table I). 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 [12].

[1] http://www.google.com/trends
[2] http://www.blogpulse.com/

TABLE I
THE CONVERSION TABLE BETWEEN PENN PART OF SPEECH TAGS AND SENTIWORDNET PART OF SPEECH TAGS

SentiWordNet POS	Penn POS abbr.
Adjective	JJ, JJR, JJS
Noun	NN, NNS, NNP, NNPS
Adverb	RB, RBR, RBS
Verb	VB, VBD, VBG, VBN, VBP, VBZ

The last phase; **Sentiment Analysis**, determines the polarity of tagged word and evaluates sentences for each day for each trend. For the evaluation there are used lexicon-based methods, which are based on SentiWordNet [13], [14]. SentiWordNet assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity. The evaluation of trends is being performed according to the following rules for each day:

1) SentiWordNet is used to identify positive/negative/ objective polarity of the words in a sentence. A polarity vector of scores is assigned to each word and the sum of these scores is always 1. For example the vector $\overline{X}$ = (1, 0, 0); (positivity, negativity, objectivity) is assigned to the word "excellent". The sum of all scores of this word is 1.
2) A word is positive if it has more positive score than negative score, and vice versa.
3) A sentence is positive if it has more positive words than negative words.
4) If the sentence has the same number of positive and negative words then the polarity of the sentence is determined by the sum of scores of individual words, and vice versa.
5) The sentence is positive if the sum of its words has more positive score than negative score, and vice versa.
6) Positive and negative evaluation of the trend is determined by the sum of positive sentences and by the sum of negative sentences for each day.

The rules can be written in a formal mathematical definition:

Def. 1 (alphabet):

Let $\sum$ be an alphabet, a non-empty finite set. Elements of $\sum$ are called characters.

Def. 2 (word):

A word over $\sum$ is any ordered n-tuple of characters from $\sum$.

Def. 3 (polarity):

Let W be a set of words, which can be identified by SentiWordNet. Let Z be a polarity of a word $w \in W$, a three-member set {positive, negative, objective}.

Now, let take the equation:

$$\forall word \in W \exists! \overline{X}; \quad \underset{i \in (pos,neg,obj)=Z}{score_i} \in \overline{X} \wedge \sum_i score_i = 1 \quad (1)$$

Then it holds that:

word is positive:

$$word_{pos} \Leftrightarrow score_{pos} > score_{neg} \qquad (2)$$

word is negative:

$$word_{neg} \Leftrightarrow score_{neg} > score_{pos} \qquad (3)$$

Def. 4 (sentence):

Let L be a language, a set of all words. A sentence over L is any ordered n-tuple of words from L.

Then for any sentence holds:

sentence is positive:

$$\begin{aligned} sentence_{pos} \Leftrightarrow & \sum word_{pos} > \sum word_{neg} \vee \\ & \vee \sum word_{pos} = \sum word_{neg} \wedge \\ & \wedge \sum score_{pos} > \sum score_{neg} \end{aligned} \qquad (4)$$

sentence is negative:

$$\begin{aligned} sentence_{neg} \Leftrightarrow & \sum word_{neg} > \sum word_{pos} \vee \\ & \vee \sum word_{neg} = \sum word_{pos} \wedge \\ & \wedge \sum score_{neg} > \sum score_{pos} \end{aligned} \qquad (5)$$

Def. 5 (trend):

A trend over L is an ordered n-tuple of words from L. The trend is fastest-rising search query per day.

Then for each trend holds:

trend is positive:

$$trend_{pos} \Leftrightarrow \sum sentence_{pos} > \sum sentence_{neg} \qquad (4)$$

trend is negative:

$$trend_{neg} \Leftrightarrow \sum sentence_{neg} > \sum sentence_{pos} \qquad (4)$$

V. DATA ANALYSIS

One of the outputs from proposed experimental system is showed in Fig. 3. In the picture is a graph, which represents evaluated chronological summary of the trend "myanmar" in 10 days around its deviation. Myanmar also known as Burma, officially the Republic of the Union of Myanmar is a country in Southeast Asia. The x-axis of the graph represents the published date and y-axis shows the polarity of the trend. Positive values of the y-axis represent positive evaluation of the trend. Negative values of the y-axis represent negative evaluation of the trend. The trend deviated on the 24th of March when there was a strong earthquake that killed more than 70 people in Myanmar. As it seen, there everyone was writing relatively positive about Myanmar before the deviation. However, the evaluation was rapidly changed on the day of trend deviation. So many bloggers had written negatively about the earthquakes at that time. The evaluation of the trend was gradually coming back to the positive values in the following days.

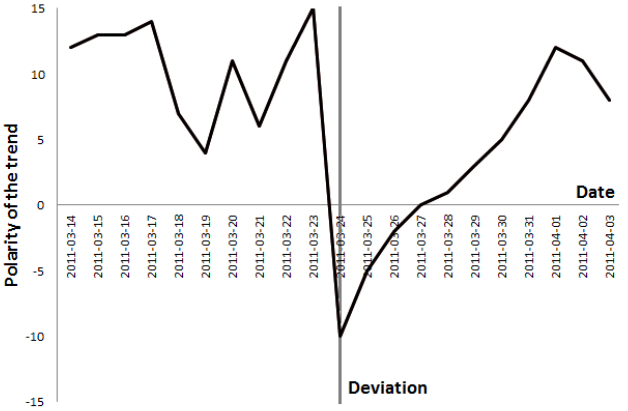

Fig. 3 Evaluated chronological summary of the trend “myanmar” in 10 days around its deviation. The trend deviated on the 24th of March when there was a strong earthquake in Myanmar.

VI. Conclusion

Thanks to the introduced evaluation of trends, it could be determined how is written about trends, which are searched on the Internet; it is positive or negative style. Furthermore, it can be found which blogs have been first writing about trend before its deviation and, it can be determined if it is possible to evaluate blogs according to the time since the trend was mentioned on them. There could also be found any correlation between sentiment polarity and the daily volume of blogs, which write about specific trend.

In the next phase of the research is necessary to verify our theoretical assumptions and compare measured results with others. For example, the trends could be represented by movie titles. Movie charts could be created according to the same methodology for gathering and processing data from Web 2.0. And finally, the film charts could be compared with another film evaluation, e.g. from The Internet Movie Database.

The final model will be used to develop an algorithm which improves the quality of search engines on the principle of webometrics, if our theoretical assumptions to be confirm. This will lead to a better machine understanding of user queries and thereby the reduction of irrelevant web search results.

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. SGS10/202/OHK3/2T/13.

References

[1] M. Thelwall, P. Wouters, and J. Fry, “Information-centered research for large-scale analyses of new information sources,” Journal of the American Society for Information Science and Technology, vol. 59, pp. 1523–1527, July 2008. [Online]. Available: http://dx.doi.org/10.1002/asi.v59:9

[2] Internet World Stats - Usage and Population Statistics, “Internet usage statistics.” [Online]. Available: http://www.internetworldstats.com/stats.htm

[3] S.-K. Han, D. Shin, J.-Y. Jung, and J. Park, “Exploring the relationship between keywords and feed elements in blog post search,” World Wide Web, vol. 12, pp. 381–398, 2009, 10.1007/s11280-009-0067-3. [Online]. Available: http://dx.doi.org/10.1007/s11280-009-0067-3

[4] M. Thelwall, “Blog searching: The first general-purpose source of retrospective public opinion in the social sciences?” 289, vol. 31, pp. 277+, 2007. [Online].Available: http://dx.doi.org/10.1108/14684520710764069

[5] 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.

[6] 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

[7] 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

[8] M. Thelwall, “Introduction to webometrics: Quantitative web research for the social sciences.” San Rafael, CA : Morgan & Claypool, 2009.

[9] 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

[10] R. Prabowo and M. Thelwall, “Sentiment analysis: A combined approach,” Journal of Informetrics, vol. 3, no. 2, pp. 143–157, 2009.

[11] 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.

[12] 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

[13] 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.

[14] 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.