Comparing Methods of Trend Assessment

Radek Malinský and Ivan Jelínek

Department of Computer Science and Engineering, Faculty of Electrical Engineering
Czech Technical University in Prague,
Karlovo náměstí 13, 121 35 Prague, Czech republic
{malinrad,jelinek}@fel.cvut.cz
http://webing.felk.cvut.cz

Abstract. This paper deals with a comparison of selected webometric methods for the evaluation of Internet trends. Each of the selected methods uses a different methodology to the trend assessment: frequency, polarity, source quality. It can be assumed that a combination of individual methods can provide much more accurate results with respect to the desired area of interest. This will lead to improve the quality of search engines on the principle of webometrics and thereby the reduction of irrelevant web search results. The introductory part of the paper explains a concept and basic functional background for all selected webometric methods.

Keywords: Webometrics, Web Mention Analysis, Sentiment Analysis, Social Network Analysis, Trend Assessment.

1 Introduction

With the growing popularity of social networking and blogging, there have been arising a large number of comments on various topics from many different types of users on the web. Some of these comments may be totally unimportant to the other Internet users. On the contrary, other comments might be very important, and do not only for an ordinary user, but also for some commercial companies that want to know a public opinion on price, quality and other factors of their products.

However, web content diversity, variety of technologies and website structure differences, all of these make the web a network of heterogeneous data, where things are difficult to find for common internet users.

Web search engines are the easiest way to find specific information in such diversified network for ordinary users. The search engines are based on complex algorithms that allow search structured but also unstructured data sets and return the most relevant results in a correlation to user-entered query. Webometric methods are often used as supportive assessment methods for search engines algorithms; Web Mention Web Analysis, Sentiment Analysis and Social Network Analysis are among the most widely used webometric methods.

Each of the selected methods uses a different methodology to the trend assessment: frequency, polarity, source quality. It can be assumed that a combination

S. Casteleyn, G. Rossi, and M. Winckler (Eds.): ICWE 2014, LNCS 8541, pp. 554–557, 2014.

of individual methods can provide much more accurate results with respect to the desired area of interest. This will lead to improve the quality of search engines on the principle of webometrics and thereby the reduction of irrelevant web search results.

2 Selected Methods of Trend Assessment

Web Mention Analysis, Sentiment Analysis and Social Network Analysis are among frequently used methods for searching and evaluating of web pages [7]. The selected methods were primarily chosen for their diversity and applicability in various areas of web and social engineering; blogs and social networks are an ideal data source for social science research because it contains vast amount of information from many different users.

Web Mention Analysis [7] is used for the evaluation of the "web impact" of documents or ideas by counting how often they are mentioned online. This idea essentially originates due to a study of an academic research. The researches wanted to know the place and the context which their works occurred in. This approach is applied in a commercial search service Google Scholar [1], which covers not only academic works but also journal articles, patents, etc. Another example of the use of Web Mention Analysis is an identification of how often and in which countries is some product (e.g. camera, book) mentioned online [2].

Sentiment Analysis or Opinion Mining [3], [4] enables us to automatically detect opinions from structured but also unstructured data. The research in this field originated from the demand of commercial companies, who wanted to know public opinion on price, quality and other features of their products. The classification of sentences or whole documents is very often based on the identification of sentiments of individual words or phrases [5]. Several approaches for the purpose have been explored and basically divided into three categories [8]: full-text machine learning, lexicon-based methods and linguistic analysis.

Social Network Analysis is the mapping and measuring of relationships on the web [2], [6]. A centrality is a part of the social network analysis, which is very often used in the web link analysis [6]. The centrality is a single node feature, which explains the node position in a network. It is for example used to determine the most active collaborator in the collaboration scientific networks [2]. There are several methods to measure the centrality of the nodes in a network [6].

3 Methodology of Study

The methods selected for the evaluation of the trends were compared over the data from film industry. Blog posts published in 2012-2013 served as the data source for this research. The five best rated movies which premiered in 2012 were chosen as trends for the assessment. All movies were selected according to IMDb[1]

[1] IMDb (Internet Movie Database) - an online database of information related to films, `http://www.imdb.com`

ranking, which is based on the site visitors rating. The rating is performed by selecting a numeric value from 1 to 10; with 10 being the best.

Because it is very difficult to find a correlation among the methods, the output of each assessment is represented as a list of movies rated from best (1) to worst (5). Table 1 shows the evaluation of trends for each compared method. The names of movies were for all comparing methods used as exact phrase searched with quotes on both sides, e. g. "The Avengers".

Table 1. Comparing Methods of Trend Assessment

Movie	IMDb	SA	WMA	SNA	SNA+SA	Rank
Django Unchained	2	3	5	4	3	4 (12)
Life of Pi	4	5	4	5	5	5 (14)
The Avengers	3	2	3	2	4	2 (7)
The Dark Knight Rises	1	1	1	1	1	1 (3)
The Hobbit: An Unexpected Journey	5	4	2	3	2	3 (9)

For Sentiment Analysis (SA), the surroundings of each searched expression had been recognized and a list of sentences for the trend had been created. All sentences were processed using Lexicon-Based method with SentiWordNet as a lexicon of words. Web Mention Analysis (WMA) is based on counting how often searched words were mentioned in the corpus of blog posts. For Social Network Analysis (SNA), the Degree Centrality was used to determine the most read blog posts and thereby to identify the trend assessment. The combination of methods (SNA+SA) represents the evaluation of trends by using Sentiment Analysis for only the most important blogs that were selected in the previous step through Social Network Analysis.

Values in the column "Rank" were determined by the sum of (SA)+(WMA)+(SNA); result of the sum is given in parentheses. Trend assessment is represented by the first number; lower sum means better ranking. The combination of methods (SNA+SA) had not been included into the sum because its output does not reflect all the data, but only the selected most important blogs are evaluated. The final ranking in the "Rank" column in comparison with ranking in the "IMDb" column, it represents the rating difference between "common users" and "film fans from IMDb".

The result shows that movie *The Dark Knight Rises* was rated as the best by all methods; it is also correlated with ranking from IMDb. On the contrary, *Django Unchained*, very well ranked on IMDb, it did not gain too much popularity on blogs (Rank is 4). This may be caused primarily by the value of WMA, which shows that this movie was at least mentioned on the web in comparison with other movies. Another important influence of the overall assessment by WMA is observed on the *The Hobbit: An Unexpected Journey*. This movie was the worst rated of the selected set of movies from IMDb. This negative evaluation is also reflected by SA, which shows the high number of negative words. However, the low value of WMA proves that the movie was high interested.

4 Conclusion

We have selected three webometric methods, which are often used as supportive search engines assessment algorithms. Each of the selected methods was used to analyze five trends (movie titles) over a set of blog posts published in 2012-2013. The output of the analysis is by popularity ordered ranking of trends (movies).

The output of each method represents a different view on the evaluation of trends: Web Mention Analysis - emphasizes the frequency of blog posts that mention the trend; Sentiment Analysis - defines the output based on the positive / negative feedback from bloggers; Social Network Analysis – defines the output by quality of blogs that mention the trend. The combination of individual methods can provide much more accurate results with respect to the desired area of interest. In our case the ranking defined by the all three methods in comparison with ranking from IMDb represents the rating difference between "common users" and "film fans from IMDb".

The subject of future work is especially in the finding a correlation among the methods. This means to define criteria for quality assessment of found information, and "distance" among each trend. On this basis, rules for evaluation of semantic content in relation to user's queries can be designed.

Acknowledgments. This research has been supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS12/149/OHK3/2T/13.

References

1. Gehanno, J. F., Rollin, L., Darmoni, S.: Is the coverage of Google Scholar enough to be used alone for systematic reviews. BMC medical informatics and decision making. Vol 13, No. 1 (2013)
2. Han, S. K., Shin, D., Jung, J. Y., Park, J.: Exploring the relationship between keywords and feed elements in blog post search. World Wide Web. 12:381–398 (2009)
3. Jagtap, V. S., Pawar, K.: Analysis of different approaches to Sentence-Level Sentiment Classification. In International Journal of Scientific Engineering and Technology. Vol. 2, No. 3, 164-170 (2013)
4. Liu, B.: Sentiment analysis and opinion mining. In Synthesis Lectures on Human Language Technologies. Vol. 5, No. 1, 1-167 (2012)
5. Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M. T., Urena-López, L. A.: Ranked WordNet Graph for Sentiment Polarity Classification in Twitter. In Computer Speech & Language. Vol. 41, No. 11, 373-381 (2013)
6. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta (2010)
7. Thelwall, M.: Introduction to webometrics: Quantitative web research for the social sciences. San Rafael, CA: Morgan & Claypool (2009)
8. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment in twitter events. Journal of the American Society for Information Science and Technology. 62:406–418 (2011)