IADIS International Journal on WWW/Internet
Vol. 14, No. 2, pp. 58-71
ISSN: 1645-7641

# THE APPLICATION OF MODERN WEBOMETRIC METHODS ON THE EVALUATION OF TRENDS IN THE SOCIAL NETWORK SPHERE

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*

**ABSTRACT**

In recent years, the Internet has been experiencing a huge boom in social networking, blogging and discussing on online forums. With the growing popularity of these communication channels, there have been arising a large number of comments on various topics from many different types of users. Such information source is not only useful for academic research, but also for commercial companies that would like to gain a direct user feedback on price, quality, and other factors of their products. However, obtaining comprehensive information from such a source is a challenging task nowadays. Our research is focused on the evaluation of the Internet trends, where the trend may be defined as anything from an event, product name, name of a person or any expression, which is mentioned online. The main emphasis of this study has been placed on the evaluation of trend in the social network sphere.

**KEYWORDS**

Webometrics, Social Network Analysis, Sentiment Analysis, Web 2.0

## 1. INTRODUCTION

In recent years, the Internet has been experiencing a huge boom in social networking, blogging and discussing on online forums. With the growing popularity of these communication channels, there have been arising a large number of comments on various topics from many different types of users. Such information source is not only useful for academic researchers, but also for commercial companies that would like to gain a direct user feedback on price, quality, and other factors of their products. However, obtaining comprehensive information from such a source is a challenging task nowadays.

Several models have been proposed for the social media analysis on the Web. However, many of these solutions are usually tailored to a specific purpose or data type, and there is still a lack of generality and unclear approach to handling the data. Moreover, a web content diversity, a variety of technologies along with the website structure differences, all of these make the Web a network of heterogeneous data, where things are difficult to find. It is, therefore, necessary to design a suitable metric that would reflect a semantic content of single pages in a better way.

Our research is focused on the evaluation of the Internet trends, where the trend may be defined as anything from an event, product name, name of a person or any expression, which is mentioned online. The main emphasis of this study has been placed on the evaluation of trend in the social network sphere. Degree Power and Power Threshold have been defined to propose a new evaluation methodology. Degree power represents a prestige of individual actors. Power threshold is the smallest value of the degree power that specifies the high-quality actors. This study deals with the influence of the threshold changes on the evaluation in the network of trends.

## 2. RELATED WORK

Our work is related to four main research topics: Sentiment Analysis; Social Network Analysis; Web Mention Analysis; and the evaluation of Internet trends.

### 2.1 Sentiment Analysis

Sentiment Analysis or Opinion Mining (Pang and Lee, 2008) enables us to automatically detect opinions from structured but also unstructured data. That involves several research areas such as natural language processing, computational linguistic and text mining. The main goal of sentiment analysis is to identify a positive/negative polarity of the text and recognise a subjective/objective impression of the text (Prabowo and Thelwall, 2009).

One of the problems of Sentiment Analysis is a sentiment classification (Liu, 2011; Ohana and Tierney, 2009), which classifies text, sentences or words as positive, negative, or neutral, and determines their strength. Sentiment analysis research involves three main approaches to determine sentiment classification, full-text machine learning, lexicon-based method and linguistic analysis, although many algorithms have elements of all.

Full-text machine learning algorithm receives a set of texts annotated for polarity by a human. According to the data input, the algorithm acquires knowledge for recognition of phrases, which identify the polarity of the text. The recognised phrases typically consist of one to three words. After that, the algorithm can be used for some non-annotated text and in relation to the learned features, the algorithm could predict a polarity of individual phrases of the non-annotated text (Pak et al., 2010).

Lexicon-based methods use a list of words, where each word is associated with its polarity and sometimes also its strength. These lists are along with a set of rules used to predict sentiment of analysed text (Malinský and Jelínek, 2014; Ohana and Tierney, 2009). Many existing lexicons can be utilised for this kind of lexicon-based analysis: SentiWordNet (Baccianella et al., 2010), Linguistic Inquiry and Word Count (LIWC) (Pennebaker, 2011), The Affective Norms for English Words (ANEW) (Lang et al., 2016). Lexicons are usually built manually for some specific corpus (Taboada et al., 2011) or semi-automatically starting with several annotated words and using heuristics to predict the sentiment of another word.

Linguistic analysis studies the grammatical and linguistic structures of analysed text, and in conjunction with the lexicon or the machine-learning methods, it tries to predict polarity of the structures. An interesting example of the use of linguistic analysis is part-of-speech (POS) tagging (Rush et al., 2012). The part-of-speech is a linguistic category of a word that is generally defined by its syntactic or morphological behaviour. The POS tagger assigns a part of speech to each word in a sentence, such as noun, verb, adjective, adverb, etc. and also recognised finite/infinitive and plural/singular form of the word. The same POS Tagger was used, e.g. for enriching textbooks produced from India, which are not written well and they often lack adequate coverage of important concepts (Agrawal et al., 2010).

## 2.2 Social Network Analysis

A social network is a structure made up of social entities (people or organisations), and their relationships. A social network analysis represents a set of techniques for the analysis of interactions and relationships between actors in a social network. An analysis output can be, just as in the Web link analysis, reported as a network diagram with nodes and links. The nodes, people or groups of people, are called *actors*; and the links, social interaction (such as friendship) between actors, are called *ties*. The ties are divided into directed and undirected. The directed ties are further divided into unidirectional and bidirectional. The bidirectional ties occur for example on Facebook (Viswanath et al., 2009), where two users have each other as a Friend. On the contrary, the unidirectional ties occur for instance on Twitter (Graham et al., 2013), where one user follows the second.

A lot of earlier researches have focused on the position of the individual node in the network and its interaction with adjacent entities. There has been introduced many network measures that help researchers to quantifies the importance of individual node and explain its position in a network. Some of these measures are focused on the node interaction with adjacent entities (Guns et al., 2011; Lee and Bonk, 2016), and others deal with a credibility of the node (Bross et al., 2012; Li et, al., 2012). Centrality and prestige are the most used measures in the social network analysis (Liu, 2011).

The idea of centrality comes from sociology, where Linton Freeman defined a set of methods called Centrality Measures based on degree, closeness, and betweenness counts (Freeman, 1977). Hanneman and Riddle (2005) described other centrality measures that extend Freeman's methods and besides they introduced software for the calculation of all specified metrics.

The centrality is a single node feature, which explains the node position in a network. In the context of a social network, a user which is followed by a group of many people and further communicates with another group of many is considered as more important than an individual with few followers. Freeman showed that certain positions in a network are more advantageous than others. The position and importance of nodes determine their impact on the network; central nodes have the greatest significance and affect most of the other nodes in the structure. All of the below mentioned Freeman’s methods are normalised to ranges between 0 and 1.

**Degree Centrality -** is the number of all connections of a node in a network. There are in-degree and out-degree centralities in the directed networks, which represent the number of incoming and outgoing connections of the node in the network. Normalised degree centrality $C_D(i)$ of the node $i$ is defined by the Equation 1, where $d(i)$ is a degree of the node $i$; $n$ is the total number of nodes

$$C_D(i) = \frac{d(i)}{n-1}$$

**Closeness Centrality -** is the sum of all shortest distances from a node to the other nodes in a network. Normalised closeness centrality $C_C(i)$ of a node $i$ is defined by the Equation 2, where $d(i,j)$ is the shortest distance between the nodes $i$ and $j$; $n$ is the total number of the nodes.

$$C_C(i) = \frac{d(i)}{\sum_{j=1}^{n} d(i,j)}$$

**Betweenness Centrality -** is the number of the shortest paths between two nodes, which pass through another measured node. Normalised betweenness centrality $C_B(i)$ of a node $i$ is defined by the equation 3, where $p_{jk}(i)$ is the number of the shortest paths between the nodes $j$ and $k$, which pass through the node $i$; $p_{jk}$ is the number of the shortest paths between node $j$ and $k$.

$$C_B(i) = \sum_{j<k} \frac{p_{jk}(i)}{p_{jk}}$$

## 2.3 Web Mention Analysis

Web Mention Analysis (Han et al., 2009; Thelwall, 2009) is used for the evaluation of the "web impact" of documents or ideas by counting how often they are mentioned online. The assessment is a combination of several types of methods:

- **Web Mentions -** determine the popularity estimation of ideas or documents using reported hit count estimates from commercial search engines. The hit count estimates are the numbers reported by search engines in their result pages as the estimated maximum number of matching pages.
- **Content Analysis -** represents a systematic separation into categories, such as the document type, national origins, industrial sector, etc. It is used to reduce the irrelevant search results that have nothing to do with the specific category.
- **Hyperlink Analysis -** is based on the extraction of information from URLs. That is very useful in the content analysis because URL extraction can provide information such as the geographic spread or the type of organisation that is interested in the document.

This idea essentially originates due to a study of academic research. The researchers wanted to know the place and the context which their works occurred in. The online search is faster and more practical than gathering a customer feedback via phone or email surveys. The approach enhanced academic communication and found new methods for the evaluation of scientific documents (Cronin et al., 1998).

Similar approaches are applied in the commercial search engine Google Scholar (Rethlefsen et al., 2009), which does not cover only academic works but also journal articles, institutional repositories, patents, etc. Web mentions are used to find information and also to sort the list of search results. The content analysis divides the articles into categories according to their areas of interest. Hyperlink analysis is primarily used along with the content analysis for building relationships between the articles; that allows us to search directly for citations between the articles.

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, etc.) mentioned online Han et al., 2009). That may partly provide sales figures and information about the geographical spread of purchases. However, as stated above, Web Mention Analysis is based on counting how often searched words were mentioned online. That does not reflect the polarity of text, and therefore, it is not able to distinguish insights into the public opinion about the specific product. Such insights could be found by using Sentiment Analysis (see Section 2.1, Sentiment Analysis), which allows as to detect opinions automatically from text.

## 2.4 Evaluation of the Internet Trends

The trend may be defined as anything from an event, product name, name of a person or any expression, which is mentioned online. Several studies have been proposed to analyse and evaluate trends on the Internet. Some of those studies are focused on the web mention analysis (Han et al., 2009; Thelwall, 2009) where the trend's popularity is calculated based on the counting how often the trend is mentioned online. Another study deals with the number of the searches in Google search engine for a specific trend (Liu et al., 2015). Different research is based on the Sentiment Analysis and utilises SentiWordNet lexicon for the evaluation of the Internet trends (Malinský and Jelínek, 2014). This approach evaluates trends (Equation 5) according to the evaluation of sentences that mention the trend (Equation 4). The evaluation is based on the sum of words' positive and negative sentiments that are obtained from the SentiWordNet.

$$sentence_{eval} = \frac{\sum pos - \sum neg}{\#pos + \#neg} \quad ($$

$$trend_{eval} = \frac{\sum sentence_{eval}}{|sentence_{eval}|} \quad ($$

# 3. EVALUATION OF TRENDS IN SOCIAL NETWORKS

The methods described in Section 2 (Related Work), Sentiment Analysis, Social Network Analysis, and Web Mention Analysis are among the frequently used methods for searching and evaluating of web pages (Thelwall, 2009). Each method uses a different methodology to the trend assessment: frequency, polarity, source quality (Malinský and Jelínek, 2014). Each of these techniques is mostly used separately, but they could be utilised together and take advantage of all their properties. The combination of individual methods can provide much more accurate results with respect to the desired area of interest.

We have proposed a new methodology for the evaluation of trends in the social network sphere (Malinský and Jelínek, 2016). The methodology is a combination of Social Network Analysis and Sentiment Analysis techniques. Social Network Analysis is used to determine the most active actor who has written about a specific trend. Sentiment Analysis is used to determine the actors' evaluation of the trend. Figure 3.1 shows an example of the social network where it is possible to evaluate the trends.

Definition 3.1 (**directed graph**):
*A directed graph $G = (N, E)$ consists of a nonempty set of nodes $N$ and a set of directed edges $E$. Each edge $e \in E$ is specified by an ordered pair of nodes $u, v \in N$.*

Definition 3.2 (**trend**):
*A trend $t \in N$ is the node of a directed graph $G = (N, E)$ for which it holds that every edge $(a, t) \in E$ is pointing to the node.*

Definition 3.3 (**actor**):
*An actor $a \in N$ is the node of a directed graph $G = (N, E)$ for which it holds that every edge $(a, t) \in E$ is pointing out of the node.*

Definition 3.4 (**comment**):
*A comment $c \in E$ is the edge of a directed graph $G = (N, E)$ that is bounded by an ordered pair of nodes $a, t \in N$, where $a$ is an actor and $t$ is a trend.*

Definition 3.5 (**network of trends**):
*Let $T$ be a set of trends, $A$ be a set of actors, and $C$ be a set of comments. A network of trends $N = (T, A, C)$ is defined as a directed graph $G$, where $N = (T \cup A)$ and $E = \{C\}$.*

The entire evaluation process in the network of trends is defined in several steps. In the first step, an adjacency matrix is constructed from the sets of actors and trends. Table 1 reports an example of the matrix, where columns represent the trends, rows show the actors, and elements indicate whether the actor has written any comment about the trend. An actor-actor relationship might also be defined in the matrix, but then one of the actors would become a trend. A directed graph, a network of trends, can be created from the adjacency matrix to better illustrate the relationships (Figure 1).

Table 1. Adjacency matrix between trend and actor nodes. Columns represent the trends, rows show the actors, and elements indicate whether the actor has written any comment about the trend.

| **Node** | **T1** | **T2** | **T3** | **T4** | **T5** |
|---|---|---|---|---|---|
| **A1** | 1 | 1 | 0 | 0 | 1 |
| **A2** | 0 | 1 | 0 | 1 | 1 |
| **A3** | 0 | 0 | 1 | 1 | 1 |
| **A4** | 0 | 0 | 0 | 1 | 0 |
| **A5** | 0 | 0 | 1 | 1 | 0 |

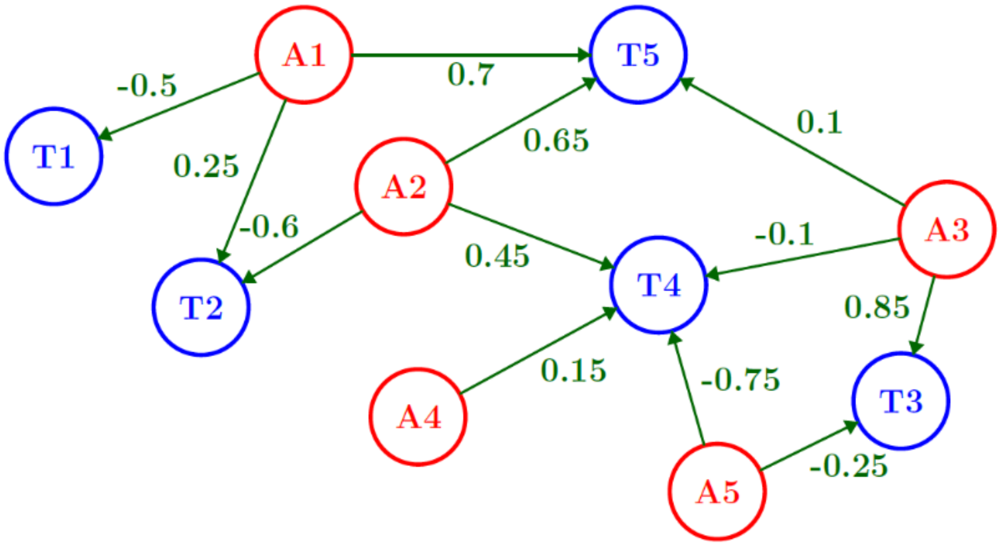


Figure 1. An example of the network of trends where it is possible to evaluate Internet trends. Red nodes represent the actors; blue nodes represent the trends. Edges indicate the actor has commented the trend. The number next to the edge represents the actor's evaluation of the comment

In the second step, a centrality measure (see Section 2.2, Social Network Analysis) is used to calculate a degree power for each of actors' nodes. The degree power for each node represents the result of the degree centrality calculation for the node's out-links. Degree power is very similar to the degree prestige (Liu, 2011). However, it calculates with the out-links in contrast to the prestige that counts with the in-links. Degree power defines a prestige of individual actors depending on the number of comments they have made; i.e. an actor is prestigious if he has commented many trends. A greater number of out-links indicates a greater power of an actor. The value of degree power is necessary to normalise between 0 (minimum degree) and 1 (maximum degree) to be able to compare the nodes of different networks of trends.

Definition 3.6 (**degree power**):
*Let $d_o(a)$ be a number of comments $c \in C$ the actor $a \in A$ has written about the all trends from $T$. Degree Power of the actor $a$ is then in the network of trends $N = (T, A, C)$ defined as $d_o(a)$ divided by the maximum possible comments $n = |C|$ the actor may have made.*

$$PW_D(\mathrm{a}) = \frac{d_o(a)}{n}; \quad \mathrm{a} \in \mathrm{A}, \mathrm{n} = |C|$$ (

In the third step, a power threshold is defined to ensure that the evaluation of trends depends only on high-quality sources, i.e. on comments from the actors that have a high degree power. The power threshold can be defined in the range from 0 to 1 to cover the whole scope of the degree power. The lower value of the power threshold implies the use of more sources from which the trend can be evaluated. On the other hand, it may also mean lower-quality sources. All sources are used when the threshold is equal to one.

Definition 3.7 (**power threshold**):
*The power threshold is the smallest value of the degree power $PW_D(a)$ that specifies the actor $a$ is a high-quality source.*

In the fourth step, Sentiment Analysis approach described in the previous section is used to evaluate all comments of the actors whose degree power is above the threshold. The calculated evaluation is then assigned to each directed edge between actor and trend. All edges are evaluated by the normalised polarity in the range from -1 to 1. The trend node positive/negative strength is then calculated as the sum of positive/negative evaluation of edges that are directed to the node (Equations 7 and 8). The overall trend evaluation is calculated as the difference between positive and negative strengths.

$$pos(t) = \frac{\sum_{a=1}^{n_a} d(a;t)}{|n_a|};\ \ n_a > 0 \wedge d > 0 \qquad (7)$$

$$neg(t) = \frac{\sum_{a=1}^{n_a} d(a;t)}{|n_a|};\ \ n_a < 0 \wedge d < 0 \qquad (8)$$

## 4. METHODOLOGY OF STUDY

The methods selected for the evaluation of trends have been examined over the data from the film industry. User reviews published in 2012-2013 on IMDb[1] serves as the source for this research. Five the best-rated and five the average-rated movies which premiered in the United States in 2012 have been chosen as the trends for the evaluation. The movies have been selected according to the IMDb Charts (IMDb, 2014) at the beginning of January 2014. The IMDb Charts contain the list of movies based on the rating of the website visitors.

All the selected movies are listed in Table 2, where the first five records represent the best-rated movies, and the last five are chosen from the average-rated movies. Because it is tough to find a correlation among the methods, the output of each evaluation is reported as a list of films rated from the best (1) to the worst (10). The evaluation of the individual methods is shown in brackets for each movie. The first column shows the movie rating obtained from the IMDb Chart, which is based on the rating of site visitors. The rating is performed by selecting a numerical value from 1 to 10; with ten being the best.

The Sentiment Analysis (SA) rating is determined according to the proposed evaluation (see Section 2.4, Evaluation of Internet Trends). All sentences have been processed using the Lexicon-Based method with SentiWordNet as the lexicon of words. For the Social Network Analysis (SNA), the degree power has been used to obtain a number of prestigious authors who have written the most reviews. The degree power value ranges from 0 to 1, and we wanted to cover all reviews from the authors who are above the average. Therefore, the value 0.5, which is the middle of the degree power range, has been selected as the evaluation threshold. Hence, the SNA value represents the number of authors who commented more than five movies. The Web Mention Analysis (WMA) is usually based on counting how often a searched word is mentioned online. However, this study deals with analyses of a closed corpus

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

data in which a counting of the words does not make much sense; therefore, the Web Mention Analysis represents the number of reviews that have been written about each movie.

The newly proposed methodology that is a combination of Social Network Analysis and Sentiment Analysis has been examined over the same data from the film industry. Social Network Analysis determines the most active actor who has written about a specific trend. Sentiment Analysis determines the actors' evaluation of the trend. The study described in the previous paragraph uses a newly proposed methodology to determine the most prestigious authors who have written reviews about the trend. In this newly proposed methodology, different variants of the power threshold are chosen to distinguish prestigious authors. Sentiment Analysis is utilised to determine the authors' evaluation for the specific trend.

As the first one, the adjacency matrix has been created between nodes of the movies and reviewers. Ten columns represent the films, 7,838 rows show the authors and 10,160 edges indicate whether the author has written a review about the trend. Degree power has been calculated to determine the author's prestige. Table 2 reports how many authors have assigned a given value of the degree power. The first row represents the scale of the reviewers' degree power. Values are not normalised, and thus they directly indicate how many reviews must be written for a given degree. The second row shows the number of authors who have assigned the degree, and the third row indicates how many reviews have been written by the authors. For instance, the first degree is assigned to 6,697 authors since each of them has created just the one review. On the other end of the scale, there are 13 reviewers and each of them has commented all of the evaluated movies. Thus, it can be stated that these ten reviewers are the most prestigious from all the rest.

Table 2. The distribution of the authors according to their Degree Power and the number of reviews they have written

| **Degree Power** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **# authors** | 6697 | 692 | 187 | 85 | 57 | 41 | 28 | 21 | 17 | 13 |
| **# reviews** | 6697 | 1384 | 561 | 340 | 285 | 246 | 196 | 168 | 153 | 130 |

## 5. RESULTS

As mentioned above, each of the selected techniques provides a different methodology to the trend assessment. Sentiment Analysis evaluates a textual content and provides the output based on the positive/negative feedback from the reviewers. Social Network Analysis determines the prestige of the authors and thus defines the quality of the source. Web Mention Analysis emphasises the frequency of making reviews and reports the overall number of reviews in a given period.

The result in Table 3 shows that the best rated IMDb movie, *The Dark Knight Rises*, is also the best rated by the SNA and WMA. The film has the highest WMA of all rated movies, i.e. there have been written a large number of reviews about the movie, which may evoke a high interest. The film also has the highest value of the SNA, so it is very interesting for prestigious authors. On the other hand, the film has an average SA evaluation. It is evident that the film is a big concern, but reviewers are not too happy.

The first by SA rated movie, *The Avengers*, is also very well evaluated on the IMDb. It can be concluded from the results that the film is loved by the general public. There have been written a lot of reviews about the movie (second highest WMA), and the reviews are very positive; polarity is the highest of all rated movies (SA). It is also obvious that there are many prestigious authors who are interested in the movie (third highest SNA). The movie is among the three most favourite movies of the prestigious authors, where the number of authors exceeds a hundred.

On the contrary, *The Amazing Spider-Man*, which is selected from the average-rated movies is evaluated very positively by the all three methodologies. Spider-man is the best IMDb evaluated movie among the average-rated movies. However, the movie has surpassed even many movies from the best-rated movies. That is primarily caused by the amount and positivity of the written reviews, and also by the high interest of prestigious authors (second highest SNA).

From an overall perspective, Sentiment Analysis reports only the positive values for all movies, which means that reviews are mostly positive rather than negative. Web Mention Analysis has a significant impact on the distance of individual evaluations. There would be a very similar evaluation of all movies without the WMA, and it would be more difficult to compare them as the trends. Social Network Analysis has a great importance especially in combination with degree power, and the result may be very different depending on the defined threshold.

Table 3. The individual evaluation results for selected methods. IMDb - Internet Movie Database, SA – Sentiment Analysis, SNA – Social Network Analysis, WMA – Web Mention Analysis

| Movie | IMDb | | SA | | SNA | | WMA | |
|---|---|---|---|---|---|---|---|---|
| Django Unchained | 2 | (8.4) | 5 | (0.0589) | 6 | (88) | 5 | (952) |
| Life of Pi | 4 | (8.0) | 3 | (0.0619) | 7 | (87) | 7 | (665) |
| The Avengers | 3 | (8.1) | 1 | (0.0820) | 3 | (101) | 2 | (1488) |
| The Dark Knight Rises | 1 | (8.4) | 4 | (0.0613) | 1 | (114) | 1 | (2491) |
| The Hobbit: An Unexpected Journey | 5 | (8.0) | 6 | (0.0568) | 4 | (92) | 3 | (1243) |
| Battleship | 10 | (5.9) | 10 | (0.0346) | 10 | (67) | 6 | (674) |
| Dark Shadows | 8 | (6.3) | 8 | (0.0522) | 9 | (74) | 10 | (440) |
| Snow White and the Huntsman | 9 | (6.2) | 9 | (0.0500) | 5 | (90) | 8 | (650) |
| The Amazing Spider-Man | 6 | (7.1) | 2 | (0.0818) | 2 | (103) | 4 | (1092) |
| Total Recall | 7 | (6.3) | 7 | (0.0546) | 8 | (77) | 9 | (465) |

Tables 4 and 5 report the result of the evaluation of movies by the newly proposed methodology that is combinations of Social Network Analysis and Sentiment Analysis. The resulting values are split into the two parts for each column in both tables. The value on the left side is the result of sentiment analysis based on the proposed Node Power evaluation. The values are normalised in the range from -1 to 1. However, these extreme values are a special case when all the reviews consist of negative or positive sentiment only. Therefore, most of the results are rather in the range from -0.1 to 0.1 and the evaluation scale must be adjusted for a larger number of trends. The second resulting value on the right side represents the number of people who have written the movie reviews. The entire corpus of reviews has no more than one review for one movie from a single author. The value on the right side also represents the number of reviews that have been written about the film.

The individual cells are tinged with a linear gradient of blue, white, and red colours. The shade of the colours represents the value in the cell. The blue colour indicates a higher value, the red colour a lower value, and the white represents the midpoint between the minimum and maximum values in the table.

Table 4. Evaluation results of the selected movies[2] according to the individual Power Threshold

| Movie | Threshold 1 | Threshold 2 | Threshold 3 | Threshold 4 | Threshold 5 | Threshold 6 | Threshold 7 | Threshold 8 | Threshold 9 | Threshold 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.0589 (952) | 0.0567 (327) | 0.0551 (214) | 0.0559 (161) | 0.0552 (124) | 0.0568 (88) | 0.0515 (60) | 0.0585 (44) | 0.0649 (27) | 0.0768 (13) |
| 2 | 0.0620 (665) | 0.0660 (273) | 0.0690 (177) | 0.0714 (138) | 0.0721 (109) | 0.0740 (87) | 0.0712 (65) | 0.0626 (43) | 0.0627 (26) | 0.0719 (13) |
| 3 | 0.0821 (1488) | 0.0828 (526) | 0.0884 (291) | 0.0873 (192) | 0.0874 (148) | 0.0983 (101) | 0.0996 (73) | 0.0931 (50) | 0.0995 (30) | 0.1017 (13) |
| 4 | 0.0614 (2491) | 0.0633 (677) | 0.0680 (339) | 0.0668 (221) | 0.0628 (155) | 0.0712 (114) | 0.0693 (78) | 0.0771 (50) | 0.0832 (30) | 0.0917 (13) |
| 5 | 0.0569 (1243) | 0.0587 (373) | 0.0574 (214) | 0.0562 (152) | 0.0561 (120) | 0.0574 (92) | 0.0555 (65) | 0.0484 (47) | 0.0500 (28) | 0.0612 (13) |
| 6 | 0.0346 (1092) | 0.0350 (430) | 0.0349 (250) | 0.0394 (183) | 0.0424 (140) | 0.0469 (103) | 0.0506 (71) | 0.0463 (46) | 0.0488 (27) | 0.0295 (13) |
| 7 | 0.0523 (465) | 0.0504 (199) | 0.0568 (142) | 0.0541 (116) | 0.0496 (89) | 0.0513 (77) | 0.0509 (60) | 0.0560 (42) | 0.0618 (28) | 0.0736 (13) |
| 8 | 0.0500 (440) | 0.0515 (193) | 0.0546 (149) | 0.0523 (118) | 0.0545 (98) | 0.0513 (74) | 0.0497 (57) | 0.0487 (43) | 0.0525 (30) | 0.0552 (13) |
| 9 | 0.0819 (650) | 0.0876 (238) | 0.0869 (162) | 0.0865 (131) | 0.0869 (109) | 0.0903 (90) | 0.0930 (67) | 0.1004 (47) | 0.1027 (29) | 0.1059 (13) |
| 10 | 0.0547 (674) | 0.0627 (227) | 0.0626 (141) | 0.0593 (106) | 0.0601 (86) | 0.0621 (67) | 0.0667 (51) | 0.0669 (39) | 0.0633 (28) | 0.0724 (13) |

Table 4 shows the results according to the specific threshold. The lower value of the threshold implies the use of a larger amount of reviews for the evaluation process; a higher number within the brackets in the left-hand columns. On the contrary, it may also mean lower-quality reviews. The higher number of the threshold means the evaluation to be processed using the high-quality reviews. However, that also means a smaller amount of reviews for the evaluation; a lower number within the brackets in the right-hand columns.

The values in brackets represent the number of authors whose degree power is equal or above the threshold in the corresponding column. Thus, for instance, the number in the first column "Threshold 1" is the sum of same numbers from the other columns; i.e. the evaluation on the left side reflects the reviews from all authors. The values in the first column also represent the state where no threshold has been applied since the review corpus does not contain any author who has made no comment.

Table 5. Evaluation results of the selected movies[2] based on the reviews by authors with a given Degree Power

| Movie | Group 1 | Group 2 | Group 3 | Group 4 | Group 5 | Group 6 | Group 7 | Group 8 | Group 9 | Group 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.0601 (625) | 0.0597 (113) | 0.0527 (53) | 0.0584 (37) | 0.0515 (36) | 0.0681 (28) | 0.0321 (16) | 0.0484 (17) | 0.0540 (14) | 0.0768 (13) |
| 2 | 0.0592 (392) | 0.0605 (96) | 0.0606 (39) | 0.0686 (29) | 0.0646 (22) | 0.0823 (22) | 0.0881 (22) | 0.0623 (17) | 0.0535 (13) | 0.0719 (13) |
| 3 | 0.0817 (962) | 0.0758 (235) | 0.0905 (99) | 0.0872 (44) | 0.0640 (47) | 0.0948 (28) | 0.1137 (23) | 0.0835 (20) | 0.0978 (17) | 0.1017 (13) |
| 4 | 0.0606 (1814) | 0.0586 (338) | 0.0702 (118) | 0.0762 (66) | 0.0395 (41) | 0.0752 (36) | 0.0555 (28) | 0.0680 (20) | 0.0767 (17) | 0.0917 (13) |
| 5 | 0.0561 (870) | 0.0604 (159) | 0.0603 (62) | 0.0565 (32) | 0.0520 (28) | 0.0618 (27) | 0.0740 (18) | 0.0462 (19) | 0.0402 (15) | 0.0612 (13) |
| 6 | 0.0345 (662) | 0.0350 (180) | 0.0216 (67) | 0.0264 (43) | 0.0266 (37) | 0.0351 (32) | 0.0643 (25) | 0.0400 (19) | 0.0656 (14) | 0.0295 (13) |
| 7 | 0.0538 (266) | 0.0288 (57) | 0.0670 (26) | 0.0764 (27) | 0.0442 (12) | 0.0526 (17) | 0.0353 (18) | 0.0427 (14) | 0.0528 (15) | 0.0736 (13) |
| 8 | 0.0492 (247) | 0.0447 (44) | 0.0644 (31) | 0.0413 (20) | 0.0695 (24) | 0.0560 (17) | 0.0521 (14) | 0.0427 (13) | 0.0503 (17) | 0.0552 (13) |
| 9 | 0.0782 (412) | 0.0885 (76) | 0.0882 (31) | 0.0850 (22) | 0.0775 (19) | 0.0842 (23) | 0.0794 (20) | 0.0972 (18) | 0.0996 (16) | 0.1059 (13) |
| 10 | 0.0487 (447) | 0.0629 (86) | 0.0771 (35) | 0.0567 (20) | 0.0477 (19) | 0.0457 (16) | 0.0662 (12) | 0.0741 (11) | 0.0554 (15) | 0.0724 (13) |

[2] 1 - Django Unchained, 2 - Life of Pi, 3 - The Avengers, 4 - The Dark Knight Rises, 5 - The Hobbit: An Unexpected Journey, 6 - Battleship, 7 - Dark Shadows, 8 - Snow White and the Huntsman, 9 - The Amazing Spider-Man, 10 - Total Recall.

Table 5 shows the evaluation results of only those reviews that have been written by the authors whose degree power is equal the threshold in the corresponding column. For instance, the first column "Group 1" represents the evaluation of 625 reviewers whose degree power is equal to one; each of them has written only the one review.

The results, for instance, show that the movie 6 - *Battleship* reaches the lowest of all the values in both tables. The Threshold 7 is the only exception, where the film is slightly better than 8 - *Snow White and the Huntsman*. However, the value in the corresponding columns Group 7, 8, 9, and 10 in the second table explain this exemption. There are 39 people in the Group 7 and 9 who like the movie more than people from the other groups. It is 39 people from 71 who like the movie more, therefore the exception in the Threshold 7. There is no exception for Threshold 9 even though the rating by Group 9 is almost double in comparison to the other groups. That is because the evaluation by the Group 10 is much lower, and people from that group meet the Threshold 9 together with Group 9.

Overall, the movie 6 is popular only for a very narrow group of people. There is a huge difference in the evaluation of 39 individuals in comparison to 1,053. The prestigious reviewers who meet the Threshold 10 evaluate the movie by the lowest value 0.0295. For comparison, a film with the second lowest value has received the rating 0.0552 from the same group, i.e. 43% difference. Compared to the best movie, the difference is even 73%.

There can be found another discrepancy in the evaluation by the Group 7. The movie 1 - *Django Unchained* has an excellent rating from almost all reviewers. However, the assessment of the Group 7 is nearly half that of the others. It is 16 people from 936 who do not like the move as the others. That shows once again that this is a very narrow group of individuals having a different requirement on a movie genre.

A different example can be seen in the evaluation of movies 3 - *The Avengers* and 9 - *The Amazing Spider-Man* which is rated as the best by the all the all reviewers. These movies are popular both for the general public and also for prestigious film reviewers. These results also correspond with the evaluation in Table 3, where these films are placed in the top three.

## 6. CONCLUSION

Three the most frequently used methods for searching and evaluating of web pages have been used to analyse trends from the film industry. The output of each method represents a different view on the evaluation of trends: Web Mention Analysis - emphasises 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 a 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.

The new methodology based on the Social Network Analysis and Sentiment Analysis has been proposed for the evaluation in the network of trends. The network of trend has been represented by the movie titles and the reviewers who have written any comment about the movie. Degree power has been determined for each reviewer to recognise his prestige. Power threshold has been used to divide the reviewers into several groups according to their prestige. The lower value of the threshold implies the use of a larger amount of reviews for the evaluation. On the contrary, it may also mean lower-quality reviews. The higher number of the threshold means the evaluation to be processed using the high-quality reviews. However, that also means a smaller amount of reviews for the evaluation. The data from each group has been used to evaluate the movie titles.

The comparison of the results across the groups may help to identify trends in extreme, i.e. very popular and unpopular trends. Evaluation for this type of trends is usually identical for all groups. On the contrary, there can be found a trend that is popular only for a particular group. It can be helpful to identify the characteristics of different groups and determine their specific requirements.

In this study, the threshold has been used to ascertain the prestige of the author by the comments he wrote. However, the threshold value might also be utilised for the dividing into the groups according to the different parameters. That might help to identify more various groups of people and to determine their requirements better.

## ACKNOWLEDGEMENT

The research has been taking place under the aegis of the research group Webing (http://webing.felk.cvut.cz) and has been supported by the Grant Agency of the Czech Technical University in Prague, Grant No. SGS16/092/OHK3/1T/13.

## REFERENCES

Agrawal, R. et al., 2010. Enriching textbooks through data mining. *Proceedings of the First ACM Symposium on Computing for Development.* New York, USA, Vol. 19, pp. 1-9.

Baccianella, S. et al., 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. Proceedings of *LREC '10*, pp. 2200–2204.

Bross, J. et al., 2012. Identifying the Top-Dogs of the Blogosphere. In *Social Network Analysis and Mining*, Vol. 2, No. 1, pp. 53–67.

Cronin, B. et al., 1998. Invoked on the Web. In *Journal of the American Society for Information Science and Technology*, Vol. 49, No. 14, pp. 1319–1328.

Freeman, L. C., 1977. A Set of Measures of Centrality Based on Betweenness. In *Sociometry*, Vol. 40, No. 1, pp. 35–41.

Graham, T. et al., 2013. Between Broadcasting Messages and Interacting with Voters: The Use of Twitter During the 2010 UK General Election Campaign. In *Information, Communication & Society*, Vol. 16, No. 5, pp. 692–716.

Guns, R., Liu, Y., and Mahbuba, D., 2011. Q-Measures and Betweenness Centrality in a Collaboration Network: A Case Study of the Field of Informetrics. In *Scientometrics*, Vol. 87, No. 1, pp. 133–147.

Han, S. K. et al., 2009. Exploring the Relationship Between Keywords and Feed Elements in Blog Post Search. In *World Wide Web*, Vol. 12, No. 4, pp. 381–398.

Hanneman, R. A. and Riddle, M., 2005. Introduction to Social Network Methods. In *Network*, pp. 332. University of California Riverside.

IMDb, 2014. IMDb Charts. *IMDb*. Retrieved 8 January 2014, from http://www.imdb.com/chart.

Lee, J. and Bonk, C. J. (2016). Social Network Analysis of Peer Relationships and Online Interactions in a Blended Class Using Blogs. In *Internet and Higher Education*, Vol. 28, pp. 35–44.

Li, R. H. et al., 2012. A Framework of Algorithms: Computing the Bias and Prestige of Nodes in Trust Networks. In *PLoS ONE*, Vol. 7, No. 12.

Liu, B., 2011. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer-Verlag Berlin Heidelberg.

Liu, D. R. et al., 2015. Recommending blog articles based on popular event trend analysis. In *Information Sciences*, vol. 305, pp. 302-319.

Malinský, R. and Jelínek, I., 2014. Comparing Methods of Trend Assessment. In Casteleyn, S., Rossi, G., and Winckler, M., editors, *Web Engineering*, volume 8541, pp. 554–557. Springer International Publishing, Cham.

Malinský, R. and Jelínek, I., 2016. The Evaluation of Node's Power in the Social Network Sphere. *Proceedings of the IADIS International Conference WWW/INTERNET 2016*, vol. 15, pp. 136-142.

Ohana, B. and Tierney, B., 2009. Sentiment Classification of Reviews Using SentiWordNet. *School of Computing 9th IT & T Conference*, pp. 13.

Pak, A. and Paroubek, P., 2010. Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of the Seventh conference on International Language Resources and Evaluation*. Valletta, Malta.

Pang, B. and Lee, L., 2008. Opinion Mining and Sentiment Analysis. In *Foundations and Trends in Information Retrieval*, Vol. 2, No. 1–2, pp. 1–135.

Pennebaker, J. W., 2011. *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury Press.

Prabowo, R. and Thelwall, M., 2009. Sentiment Analysis: A Combined Approach. In *Journal of Informetrics*, Vol. 3, No. 2, pp. 143–157.

Rethlefsen, M. L. et al., 2009. Google Scholar. In *Internet Cool Tools for Physicians*, pp. 37–40. Springer Berlin Heidelberg.

Rush, A. M. et al., 2012. Improved parsing and POS tagging using inter-sentence consistency constraints. *Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1434-1444.

Taboada, M. et al., 2011. Lexicon-Based Methods for Sentiment Analysis. In *Computational Linguistics*, Vol., 37, No. 2, pp. 267–307.

Thelwall, M., 2009. *Introduction to Webometrics: Quantitative Web Research for the Social Sciences*, Volume 1.

Viswanath, B. et al., 2009. On the Evolution of User Interaction in Facebook. *Proceedings of the 2nd ACM Workshop on Online Social Networks*, WOSN ’09, pp. 37–42, New York, NY, USA. ACM.