Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Computer Science and Engineering

NOVEL WEB METRICS BASED ON SENTIMENT ANALYSIS

Doctoral Thesis

RADEK MALINSKÝ

PhD Programme: Electrical Engineering and Information Technology
Branch of Study: Information Science and Computer Engineering

February 2017

Supervisor:
doc. Ing. Ivan Jelínek, CSc.
Department of Computer Science and Engineering
Faculty of Electrical Engineering
Czech Technical University in Prague
Karlovo náměstí 13
121 35 Prague 2
Czech Republic

Abstract and Contributions

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.

In this thesis, the main emphasis has been placed on the evaluation of 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. A general model has been proposed to collect and analyse data from the Web. The analysis part of the model is based on webometric principles that are enhanced by the methods of sentiment and social network analysis. The extension of webometrics by the combination of these methods leads up to gaining insights into the public opinion with respect to some topic, and a better machine understanding of a text.

In particular, the main contributions of the dissertation thesis are as follows:

1. Proposal of the new general model for gathering and processing data from Web 2.0.
2. Definition of the methodology for the evaluation of Internet trends.
3. Adaptation of the newly designed methodology for the evaluation in social network sphere.
4. Proposal of the new sentiment sense disambiguation methods to improve sentiment classification for multiple-topic related words.
5. Architecture design of the new framework that provides an end-to-end approach to the analysis of selected Internet trends.

Keywords:
Webometrics; Sentiment Analysis; Social Network Analysis; Sense Disambiguation; Web 2.0

Acknowledgements

First of all, I want to thank my supervisor, doc. Ing. Ivan Jelínek, CSc., for his advice and support throughout my studies. He has been a constant source of encouragement and insight during my research and helped me with numerous problems and professional advancements.

I would like to thank all my colleagues from the research group Webing for all the inspiring discussions we have had. I also want to thank the Department of Computer Science and Engineering for a pleasant and flexible environment for my research.

This work would not have been possible without the financial support of the Czech Technical University in Prague, CTU in Prague, for granting me funding to travel and visit conferences and doctoral forums that have had a great impact on my research.

I am forever thankful to my parents who have always supported me and encouraged me to study for as long as I want. I want to thank my family members for their infinite patience and care, and friends for support during the study and proofreading of this work. And last, but definitively not least, I want to thank my beloved partner, for her love, constant support, and understanding for those countless hours that I have spent in front of the computer screen. Thank you for being so awesome.

This research has also been performed under:

- The project *The Evaluation of Node's Power in the Social Network Sphere using Modern Webometric Methods* and partially supported the Grant Agency of the Czech Technical University in Prague, Grant No. SGS16/092/OHK3/1T/13.
- The project *A Novel Web-Based Framework for the Evaluation of Internet Trends* and partially supported the Grant Agency of the Czech Technical University in Prague, Grant No. SGS SGS15/087/OHK3/1T/13.
- The project *A Novel Web Metrics Based on Sentiment Analysis* and partially supported the Grant Agency of the Czech Technical University in Prague, Grant No. SGS12/149/OHK3/2T/13.
- The project *Adaption of the webometric techniques for Web 2.0* and partially supported the Grant Agency of the Czech Technical University in Prague, Grant No. SGS10/202/OHK3/2T/13.
- The research program No. 6840770014 and partially supported the Ministry of Education, Youth, and Sport of the Czech Republic.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We are all now connected by the Internet, like neurons in a giant brain.
— Stephen Hawking

1.1 Motivation

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 on the Web. More and more people share their thoughts and opinions on products and events around them. This development phase of the Web, collectively Web 2.0, has reached such a stage, where it is quite usual for an ordinary user to communicate mostly online. Many web services have been gradually adapting to this trend, and for instance, the online shops allow the consumers to post comments on goods they have bought. Such comment on the goods can facilitate a decision making in purchase to another customer. Moreover, the seller gets a quick feedback on the goods, and the producer obtains an opinion on how to improve his products.

The current Internet stage has become a rich information source for a social science research. People have friends with the same interests on social networks, and they have been sharing their feelings on the daily events, whether about work, politics, gossips or personal life. The virtual friends can contribute and further comment on their thoughts. Many types of research are therefore transferred to the

Web to exploit the potential of this source. However, such social medium 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.

The main disadvantage of such information source is that a significant amount of information may not be relevant at all, and it might also be unrelated to the desired area of interest. Also, a 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 for common Internet users. It is, therefore, necessary to design a suitable metric for such volume of information that would reflect a semantic content of single pages in a better way.

1.2 Problem Statement

As stated, the Web 2.0 has been becoming an important information source since it contains many ideas on various topics from many different users. The obtaining comprehensive information from such a source is a challenging task nowadays, which includes the investigation of reciprocal relationships, analysis of the website content and recognition of its meaning. Webometrics is a scientific discipline that studies the quantitative aspects of information sources and their use. Original webometric techniques are focused on hyperlinks, and they exploit their interconnection to measure the World Wide Web. However, the methods have been reaching their limits, and they do not fully reflect the needs of the current Web.

Complex solutions are usually realised for a data analysis in the Web 2.0. However, these solutions are typically tailored to a specific purpose or data type, and there is still a lack of generality and unclear approach to handling the data. There is currently no widely acceptable solution for a data analysis in such heterogeneous environment. Our research emphasis has been placed on the extension of Webometrics by the methods of Sentiment and Social Network Analysis. The main focus is the evaluation of 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.

Social Network Analysis provides a broad range of resources to analyse the relations in a social network. Sentiment Analysis allows us to detect opinions from structured and also unstructured data. The combination of individual methods can provide much more accurate results with respect to a desired area of interest. The extension of Webometrics by a combination of these methods leads up to gaining

insights into the public opinion with respect to some topic and a better machine understanding of a text. Better machine understanding of the content on the Web might have a significant impact on the quality of a website evaluation.

1.3 Contributions of the Thesis

The thesis aims to find a suitable solution for data analysis in the Web 2.0 environment. The results could be used in a complex analysis system that would provide insights into the public opinion on a specific search topic. The main contributions of the thesis are the following:

1. Proposal of the new general model for gathering and processing data from Web 2.0. A functionality of the model is similar to a typical web search engine; however, the model focuses on the evaluation of Internet trends.

2. Definition of the methodology for the evaluation of Internet trends. Webometric principles are used as the cornerstone that is further enhanced by the idea of Sentiment Analysis.

3. Adaptation of the newly designed methodology for the evaluation in social network sphere by using the Social Network Analysis.

4. Proposal of the new sentiment sense disambiguation methods to improve sentiment classification for multiple-topic related words.

5. Architecture design of the new framework that provides an end-to-end approach to the analysis of selected Internet trends.

1.4 Structure of the Thesis

The thesis is organised into seven chapters as follows:

1. *Introduction*: Describes the motivation behind our efforts together with the goals. There is also a list of contributions of this dissertation thesis.

2. *The World Wide Web*: Presents the dynamic structure and the basic functionality of the World Wide Web. The methodology of collection and analysis of information from the Web are described therein.

3. *Webometric Background and State of the Art*: Introduces the necessary theoretical background and surveys the current state-of-the-art. Some methods from Sentiment Analysis and Social Network Analysis that are used in this research are presented and explained in this chapter.

4. *A Novel Web Metric for the Evaluation of Internet Trends*: Overview of our approach and the introduction of the new methods for the analysis of Internet trends. The chapter presents the methodology of the evaluation of Internet trends along with a comparison to other methodologies, and it introduces new approaches for the sense disambiguation.

5. *Framework for the Analysis of Internet Trends*: Describes the architecture of the framework for the analysis of selected Internet trends. The framework associates crawler, analysis algorithms and fully configurable user interface to define which data should be analysed and how.

6. *Experiments and Main Results*: Demonstrate the experiments carried out to evaluate the theoretical assumptions.

7. *Conclusions*: Summarises the results of our research, suggests possible topics for further research and concludes the thesis.

Chapter 2

The World Wide Web

The World Wide Web (WWW, or simply the Web) is an information space in which the items of interest, referred to as resources, are identified by global identifiers called Uniform Resource Identifiers URI[1] (Jacobs and Walsh, 2004). The World Wide Web represents the part of the Internet that can be accessed through a web browser. It is a "live" network of interconnected websites and hyperlinks between them, where the website can be added or removed at any time from that network.

The World Wide Web was invented by English computer scientist Tim Berners-Lee (Berners-Lee, 1989) and it has been exerting under the baton of the World Wide Web Consortium (W3C). HTML and HTTP are among the cornerstone technologies developed by the W3C. HTML (HyperText Markup Language) is a markup language that is along with Cascading Style Sheets and JavaScript used for structuring and presenting of a content of web pages. HTML language, in the latest version (Hickson et al., 2014), syntactically and semantically describes a content structure that is interpreted via web browsers to a human readable form. HTTP (Hypertext Transfer Protocol) is an Internet protocol for exchange of hypertext documents in HTML format. HTTP protocol is along with the family of TCP/IP protocols used for transmission of documents between server and client in a computer network.

[1] URI (Uniform Resource Identifier) is a string of characters with a defined structure that identifies an exact information resource, especially in a network, typically the World Wide Web. A URI has the following structure: `scheme:[//[user:password@]host[:port]][/]path[?query][#fragment]`, which can be defined for a web page as `http://www.domain.com/path/file.html`, where *file.html* is located in a directory *path* under the top level domain *domain.com*.

2.1 Network Analysis

Königsberg, known as Kaliningrad today, the city of a mathematician Leonhard Euler, is famous for a puzzle with seven bridges over the Pregel River. At the beginning of the 18^{th} century, there was a popular activity to cross over the all seven bridges, each only once, and it did not matter where a walk began and ended. It was an unsolved puzzle until Leonhard Euler outlined the problem into a graph where lands are represented as nodes and bridges as links connecting the nodes. A number of links to one node are defined as the degree of the node. Euler determined conditions for a successful walk over the bridges from the graph - a walk would cross each bridge only once when all nodes in the graph, except at most two, having an even degree. That means that such a walk was not possible until the new bridge would be built. Euler publicly presented his solution in 1735 and this period is considered today as the beginning of a new research field, graph theory (Crilly, 2007).

As stated, the World Wide Web represents a network of interconnected websites and hyperlinks between them. As well as Seven Bridges of Königsberg, the WWW network can be shown as a graph by replacing sites with nodes and hyperlinks with links connecting the nodes. Graph theory or network analysis has become very popular research sphere because it can more easily answer the questions that would be otherwise difficult to detect; especially for relatively small networks because it is straightforward to draw a picture of a network with points and lines and answer a specific question by examining this picture. For instance, it can be investigated which node in a network may have a decisive impact on communication among all other nodes in the network if it will be removed. However, such analysis is relatively impossible in a network with millions of nodes; therefore, the analysis has become more mathematical and statistical for this kind of networks and rather deals with questions such what percentage of nodes need to be removed to impact a communication in a network (Newman, 2010). The analysis is also often focused on the relation between sub-networks of the huge network (Ausserhofer and Maireder, 2013).

Over the years, there has been a development of an extensive set of computational and statistical tools for analysing and modelling networks (Brown et al., 2009; Scott, 2012; Xu and Chen, 2005), and there have also been introduced many network measures. Two of the key measure for network analysis are centrality and prestige. Centrality is a single node feature, which explains the node's position in a network and quantifies the importance of that node. Centrality was for example used to

determine the most active researcher in the scientific collaboration networks (Guns et al., 2011). Prestige is also a single node feature, and for instance, in social networks, the prestige reflects how the actor is trusted by others (Li et al., 2012). The main difference between these two measures is that centrality focuses on links pointed out of the node (out-links) while prestige focuses on links pointed to the node (in-links).

Newman (2010) performed the comparative study of networks from different branches of science, with emphasis on properties that are common to many of them and divided the networks into four categories:

Social Networks A social network is a structure made up of social entities (people or organisations), and their relationships. Sociologists have developed their own language for social networks: the nodes, people or groups of people, are called *actors*; and the links, social interaction (such as friendship) between actors, are called *ties*. Social network studies can include for instance criminal network analysis (Xu and Chen, 2005), influence of friends and peer groups in shaping physical activity behaviours (MacDonald-Wallis et al., 2012), the analysis in economic geography (Ter Wal and Boschma, 2009) and many others.

Information Networks Scientific communication, especially citations between academic papers, is a typical example of an information network. This network is acyclic because the paper can only cite other papers that have already been written. Ranking of World Universities is a successful example of the analysis in this network (Aguillo et al., 2008).

Biological Networks Biological systems like the brain, heart, eyes, etc. can be represented as a network and allow to track interactions between individual parts (Bashan et al., 2012). Other studies are for instance focused on the identifying new disease genes (Barabási et al., 2011).

Technological Networks Human-based networks are designed typically for a distribution of some commodity. Examples of technology networks include roads between cities, railways, rivers or the Internet as a network of physical connections between computers (Lyon, 2005, 2015).

2.2 Web Data Mining

The World Wide Web has been rapidly growing in the last decades, and it has become the largest publicly accessible data source in the world. Unfortunately, a web content diversity, a variety of technologies, and the website structure differences, all of these make the Web a network of heterogeneous data, where things are difficult to find for common Internet users. All these poor characteristics present a challenge for the data mining and discovery of information from the Web.

Web data mining is an analytical methodology for obtaining a potentially useful information from the Web. Web data mining involves techniques for information extraction from texts, images, videos, audio; and it is commonly defined as the process of discovering patterns from data sources.

Data mining is widely used in a commercial sphere where it replaced a traditional customer feedback collection via phone or email. Over the last few years, the mining has been using for a cyber-crime detection (Chen et al., 2004), or for sentiment analysis on social media channels (Petz et al., 2013). Data analysts commonly perform the mining in three steps:

Pre-Processing The raw crawled data is usually not immediately suitable for the mining. Such data are cleaned from noises, abnormalities or irrelevant parts during the pre-processing phase, and there is also ensured the inputs have a same structure for the next phase.

Data Mining The pre-processed data is then processed by the selected mining algorithm. It can be searched for relations between variables (e.g. customer purchasing habits), or some unknown structures behaviour can be discovered based on another similar structure.

Post-Processing The final step is to verify result, process an evaluation and apply visualisation techniques to make the decision about the crawled data.

Thanks to the mentioned diversity and heterogeneity of the Web, there have been introduced a significant number of web mining tasks and discovered many algorithms to analyse them. Based on the mining issue and the data input type, the web mining tasks can be divided into three different categories (Liu, 2011):

Web Structure Mining Web structure mining uses graph theory to analyse the structure of web pages and connections between them. The mining discovers a

useful knowledge from hyperlinks and tries to identify important web pages that are connected by them (Renu and Gaur, 2014). It can also discover communities of users on the website that includes a structure of discussion forum. Web pages are analysed as a tree-level graph of individual HTML or XML tags.

Web Content Mining Web content mining is used to extract and analyse the information from the Web. By using of selected techniques, it is possible automatically to identify the main topic of the web page is about and thanks to that cluster a bunch of similar sites (John et al., 2016). Another approach is to obtain user's comments from blogs and analyse the public opinion on the particular product or feature (Petz et al., 2013).

Web Usage Mining Web usage mining discovers web user behaviour from application logs, which record every click or a search query made by each user (Sterne, 2003). Analysing such data may help web administrators to optimise the website functionality, or e-shops can evaluate the effectiveness of a promotional campaign and better design the future market strategies.

2.3 Web Data Collection

Network analysis and web data mining provide the techniques to analyse websites and web pages. They are able to work with offline and even online data. However, the data needs to be extracted from the Web before the analysis. There are generally defined three ways to extract data from the Web (Holmberg, 2009): 1) manually visit every site and collect the data, 2) use a commercial search engine, 3) develop a personal web crawler. The first option seems to be the simplest, and it can be useful to analyse several sites in this way. However, it would be very time consuming when the task will require hundreds of sites to visit. A commercial search engine can be more efficient for such large number of sites. A researcher just enters specific search keywords, and the engine returns a result. No doubt, this is the fastest way to get the data. However, the search engines may not index all the pages, or they may apply limitations to a number of search queries that can be entered in. With a personal web crawler, the researcher has full control to define which sites will be visited and when, how far from the first page to discover new pages, or what kind of content to be crawled.

A web crawler is an automated unit that follows links on the website and stores a key content of all the visited pages. The crawler selects links to follow until some termination condition is not satisfied; for instance, there are no more links on the website. It can be designed to crawl and index documents such as Word, Excel, PDF, etc. and multimedia files (Turek et al., 2011), or vice versa to skip some web page chunks like banners or advertising parts. A properly designed crawler should take ethical considerations into account when deciding where to crawl. It should especially not frequently crawl on a specific site in order to prevent the overload of the site and servers where it is hosted (Sun et al., 2010).

Commercial search engines like Google, Bing, Yahoo! provide another option to access a huge number of sites. Search engines are partly similar to web crawlers; they visit a page by page by following hyperlinks and trying to go through the entire World Wide Web. Each visited page is subjected to the analysis of content and hyperlinks and based on that the page is evaluated, indexed, and stored in a database. The in-links that led the crawler and the out-links that guide to the next page have a major role in the evaluation (Langville and Meyer, 2006). Search engines can provide the application programming interface and allow researchers to enter searched keywords to find relevant websites and their content (Thelwall and Sud, 2012). Some engines are additionally designed to provide a history of individual pages (The Internet Archive, 2016).

Search engine begins to crawl the Web on one page and using the links it continues for another. However, if any crawled page refers to another and the another page has no connection to the first then the crawler will not discover that page. Conversely, a personal crawler mostly uses a defined corpus of pages to go through, so it does not need to discover some new ones.

Chapter 3

Webometric Background and State of the Art

Webometrics is a scientific discipline that studies the quantitative aspects of information sources and their use. In other words, webometrics tries to measure the World Wide Web, analyses technology usage and provides methods for a simple content analysis. As Figure 3.1 shows, webometrics is affected by many scientific disciplines:

- **Informetrics** This scientific discipline uses mathematical and statistical methods to describe and analyse information phenomena and the relationship between them. Informetrics mainly deals with a quantification of the information, measuring of information flows and with the evaluation of information processes. The result of the informetric research may serve for analysis of the quantitative growth of literature, for a measure of the efficiency of information systems, for evaluation of scientific communication, etc.

- **Bibliometrics** Bibliometrics uses informetric methods like quantitative analysis and data visualisation to determine the characteristics of bibliographic references, citations, authors, institutions, keywords, etc. Bibliometrics further uses citation analysis to clarify the quality of written documents. Bibliometrics as such can be used to evaluate the intensity of the use of librarian and information services.

- **Scientometrics** Scientometrics is an extension of bibliometrics that is focused on the evaluation of scientific research or individual researchers. The assessment is primarily based on the number and quality of citations of scientific work.

Using the scientometric approaches, it is also possible to characterise the historical evolution of scientific communication within a given research field.

- **Cybermetrics** Cybermetrics includes all previously mentioned disciplines and carries them into the Internet environment. Above all, the Cybermetrics deals with quantitative research of information sources, structures and technologies on the Internet. The subjects of its investigation are all electronic information flows, such as posting in discussion groups, e-mail communication, texting or other forms of communication.

However, webometrics is primarily based on informetric and bibliometric approaches (Thelwall, 2008; White and McCain, 1989). The information sources that are studied by webometrics are websites, web pages, parts of web pages, words in web pages, hyperlinks, and web search results (Thelwall, 2009). According to a narrow definition, webometrics encompasses five basic categories: web link structure analysis, web page content analysis, web usage analysis, web technology analysis and the evaluation of search engines using informetric methods.

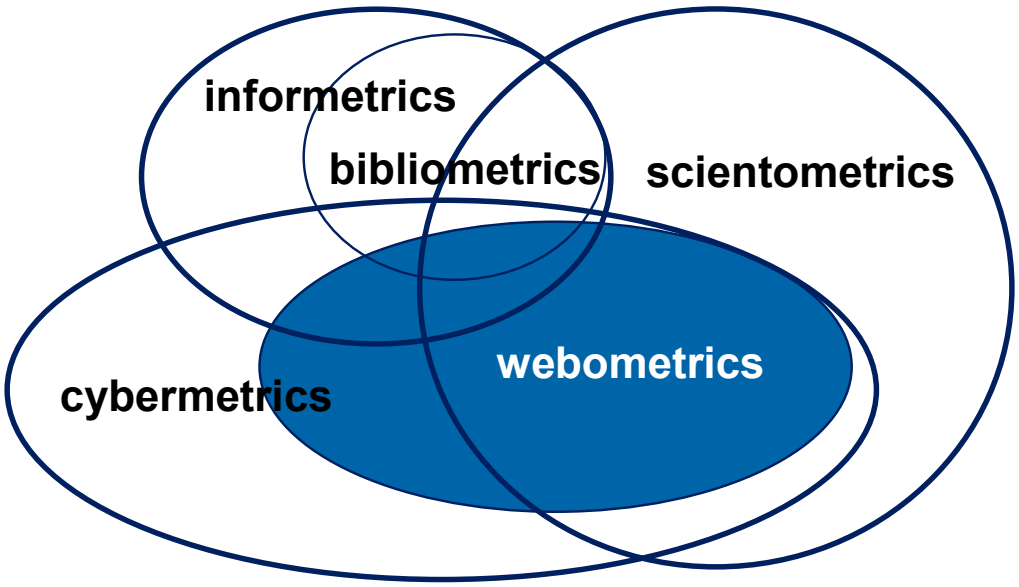


FIGURE 3.1: The interrelation of webometrics and bibliometrics, cybermetrics, informetrics, and scientometrics.

3.1 A Brief History of the Webometric Research

Thomas C. Almind and Peter Ingwersen were the first who defined the term "webometrics" and described it in their journal article (Almind and Ingwersen, 1997) as:

> "... research of all network-based communication using informetric or other quantitative measures.".

They used a quantitative data that was collected from the Web and compared the number of specific informetric analysis parameters of Danish and other Nordic countries on the Web. Web documents were processed in the same way as traditional printed documents during bibliometric analysis; however, bibliographic references were substituted by hyperlinks.

Nearly one year later, Ingwersen defined Web Impact Factor (Ingwersen, 1998), which is a parallel of Journal Impact Factor. The Journal Impact Factor is defined as the number of citations to a journal divided by the number of articles in that journal related to a given period (Garfield, 2005). The Web Impact Factor uses the same idea; it is defined for a website as the number of web pages receiving links from the other websites, divided by the number of web pages that are accessible to the crawler. The Web Impact Factor measures the impact of a Web area and, it is also used by commercial search engines to classify search results (Langville and Meyer, 2006; Thelwall, 2008).

In the following years, social science research techniques have been applied to webometrics (Kretschmer and Aguillo, 2004; Otte and Rousseau, 2002). One of the areas of the social science research is Social Network Analysis. The Social Network Analysis is a strategy for investigating of the structures that represent social relationships in terms of nodes and links. The evaluated output can be reported as a graph or as a network diagram with nodes representing individual actors (websites) within the network, and ties between the actors representing the relationships between them.

In 2004, there was the first conference devoted entirely to the Webometrics in the India. Hildrun Kretschmer and Isidro Aguillo introduced the use of Social Network Analysis in webometric research (Kretschmer and Aguillo, 2004). The Social Network Analysis had been applied to display the structure of a scientific collaboration network and thereby enable to detect the influence of one author within a scientific community.

In the same year, Lennart Björneborn and Peter Ingwersen provided a complete discussion of webometric terminology and techniques where was shown that the Web links are made for a different reason other than for bibliographic citations, and therefore, the links do not have the same semantics (Björneborn and Ingwersen, 2004). Björneborn also specified webometrics in his doctoral thesis (Björneborn, 2004) as:

> "The study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web drawing on bibliometric and informetric approaches.".

According to the narrow definition, webometrics encompasses the following categories: page content analysis, link structure analysis, usage analysis (including log files of searching and browsing) and web technology analysis (including search engine performance).

In 2009, Mike Thelwall introduced the new webometric definition, which is free from informetric and bibliometric approaches (Thelwall, 2009):

> "The study of web-based content with primarily quantitative methods for social science research goals using techniques that are not specific to one field of study."

The main emphasis is devoted to the use of applied methods for a social science research. For instance, the study of online political communication during election campaigns (Park, 2011) and seeking of differences and commonalities between parties (Foot and Schneider, 2006). Another example is focused on the spread of news created by an amateur journalist in blog space and the visualisation of news distribution (Takama et al., 2007). Graph style visualisation examines the relationships between news articles, blog entries and similar objects.

3.1.1 From History to the Present

In the beginning, the webometric research has been focused on the evaluation of web pages without any previous knowledge. Many of these quantitative studies were focused on hyperlinks. Google PageRank (Langville and Meyer, 2006) used by the Google search engine or Ranking of World Universities (Aguillo et al., 2008) are splendid examples of successful projects based on these studies. Another example

was focused on government websites and tried to investigate whether interlinking between the local administration bodies in Finland follows a strong geographic, or rather a geopolitical pattern (Holmberg and Thelwall, 2009).

Having regard to hyperlinks, the quality of pages was determined by bibliometric and informetric approaches (Langville and Meyer, 2006; Thelwall, 2008). However, these approaches are very simple for heterogeneous World Wide Web and do not reflect the semantic content of single pages. Therefore, many of current search engines excel in queries to a specific case, however, when a complex query is entered then the search engine returns a large number of irrelevant results.

Webometrics began to be widely used in the Web 2.0[1] area, for instance, to find relevant information on blogs (Bross et al., 2012; Han et al., 2009) or for a trend detection (Thelwall, 2007). In the beginning, it was a simple context analysis based on counting how often the searched word was mentioned online. However, current research is focused on a sophisticated analysis of sentences, which aims to determine the polarity of text and the attitude of a writer with respect to some topic (Potthast and Becker, 2010; Zhang et al., 2015), or to the analysis of a relationship between individual web pages and users who publish on them (Lee and Bonk, 2016).

Several types of techniques have been designed during the gradual development of webometrics: web link structure analysis, web page content analysis, web usage analysis, web technology analysis and the other new are still being made.

3.2 Hyperlink Analysis

Hyperlinks between web pages and websites have always been an interesting topic of scientific researchers. Hyperlinks primarily facilitate navigation on the Web, where they provide a way to move from one site to another. However, the reason why those sites have been linked, the way they are connected, and the type of link which connects them; these and many other similar questions bring the interest to the scientific community for deeper exploration.

The hyperlinks to web pages or websites may be formed for many reasons. For instance, an e-shop page with selling product may content a link to the blog where reviews, information and user experience of that particular product are described. Similarly, such review or blog post may again refer to another site which the author

[1]The term "Web 2.0" is a designation for the developmental stage of the Web, where static content is replaced by space for creating and sharing content.

used as a first-hand information source. Another example might be an online course, which refers to a website with detailed information that is not covered by the course. Above examples may indicate that links, out of the navigation links, are primarily formed to refer relevant, high-quality information pages, and therefore such useful pages should attract the most links. This phenomenon was used by Sergey Brin and Lawrence Page in the design of Google's PageRank algorithm, and also by Jon M. Kleinberg when he was forming HITS algorithm. Both of those algorithms are used by search engines to produce a list of websites that are ordered by relevance to the user search query. The search engines reflect many other factors when ranking the websites; however, the ranking algorithms that operate with hyperlinks are the cornerstones of modern web engines.

PageRank Algorithm

PageRank (Brin and Page, 2012; Page et al., 1999) used by the Google search engine is one the most reliable and efficient link analysis algorithm. The algorithm uses the structure of hyperlinks as a mutual "recommendation" of the connected pages; this is very similar to the evaluation of scientific work according to the number of citations. Contrary to the track the number of citations, the algorithm has been leading this principle beyond (Langville and Meyer, 2006): the page rank is not calculated just from the number of links that lead to the page, but it also reflects the evaluation of those pages that link to the rated page. The evaluation of individual pages is independent of the user-specified search query. PageRank thus falls into the category of static ranking algorithms because the evaluation can be performed off-line. Hyperlinks are divided into two categories for each page before the evaluation (using the terminology of Björneborn (2005)):

- **In-links** of page i represent the hyperlinks that point to page i from other pages.
- **Out-links** of page i represent the hyperlinks that point out to other pages from page i.

In-links that point to the page from other pages within the site are typically used as a navigation between all pages for a given site. Such links might artificially increase the rating of individual pages or loop the evaluation, and therefore they are not counted in the rating. The same approach is used for out-links that point to

pages on the same site. A special case of hyperlinks, which are also not counted in the PageRank calculation, are the self-links that point from one section to another within the same page. Another issue might be caused by the parallel links that point from several pages within the site to another page on another site. It would not be fair to calculate the page score using all that links. Therefore, there is usually used just one link between any pair of websites. In other words, the calculation is processed by using links between websites rather than links between pages.

For the calculating itself, the Web can be imagined as a directed graph $G = (P, H)$, where P is the set of nodes (all pages on the Web), and H is the set of directed links (hyperlinks) between the nodes. As mentioned, the rating is not only based on the number of links but primarily on the evaluation of the referring pages. The PageRank score 3.1 of page i (denoted by $PR(i)$) is therefore defined as the sum of ranks of all pages linked to i:

$$PR(i) = \frac{(1-d)}{|P|} + d \sum_{(j,i)\in H} \frac{PR(j)}{|O_j|} \tag{3.1}$$

where $PR(j)$ is the PageRank score of page j, which links to page i; $|O_j|$ is the number of all out-links on page j; $|P|$ is the total number of pages on the Web; $d \in [0, 1]$ is called the damping factor. The variable d itself denotes the probability that random visitor clicks on a link on the page he is currently viewing and thereby he will continue to another page (pass over the edge from one node to another in the graph terminology).

The damping factor was introduced to avoid problems with two pages that refer only to themselves and not to another page (rank sink), and problems with in-links to a page that has no out-links (dangling links). There are several studies about the correct value of the damping factor (Patel, 2014; Son et al., 2012; Wu et al., 2012); however, it was assumed that the damping factor would be set around 0.85.

HITS Algorithm

Hypertext Induced Topic Search (Kleinberg, 1999), known as HITS, is a link analysis algorithm that rates web pages. HITS is similar to PageRank, but with two differences. PageRank as a representative of static ranking algorithms is independent of a user-specified search query; the page ranking is performed offline before any user searching. However, HITS depends on the user's query, i.e. it works with a list of pages that are relevant to a user search query (pages returned by a search engine). The second

difference is that the PageRank assigns a single value (score) to each page that is included in the evaluation. HITS produces two rankings for every page, authority ranking which estimates the content of a page, and hub ranking, which estimates the quality of links to other pages.

The authority is a page with many in-links. HITS assumes that such a page is commonly popular because it may have good content on some specific topic, and therefore many people trust it and link to that page. The hub is a page with many out-links. Hub is often presented as the index of information, which points to many good authority pages. HITS uses a mutually reinforcing relationship between authority and hub pages. A good hub is a page that points to many good authorities, and a good authority is a page that is pointed to by many good hubs (Gupta et al., 2015).

As mentioned, HITS works with a list of pages that are relevant to the search query. Therefore, there is entered a query into a search engine before the actual calculation, and the search engine returns the unordered list of the most relevant sites (typically of 200). This list is then extended by pages that are linked from the list, and further by pages that pointing to pages from the list (typically 50 new pages for each page in the list). HITS algorithm to calculate the authority and hub scores is applied to all of these pages. The authority score $a(i)$ of page i and the hub score $h(i)$ of page i have mutually reinforcing relationship defined as:

$$a(i) = \sum_{(j,i)\in P} h(j) \tag{3.2}$$

$$h(i) = \sum_{(i,j)\in S} a(j) \tag{3.3}$$

Where P is the set of all predecessor's pages of page i and S is the set of all successor's pages of page i. At the beginning of calculation is set $a(i) = h(i) = 1$ for all pages and then authority and hub scores are updated using the power iteration method depends on 3.2 and 3.3. Both scores are also normalised after each iteration to ensure the sum of values for every score is equal to one. HITS then selects a few top ranked authority and hub pages and serves them to the user. HITS is part of Ask Jeeves/Ask.com search engine, which uses the algorithm to sort search results (Yang, 2006). Search engines do not commonly use the algorithm because of the need for real-time execution; however, it became the basis for many other types of research. A modified version of the algorithm was, for instance, used to

evaluate the professional skills of wine tasters (London and Csendes, 2013), or to solve the ambiguity in text-based image retrieval (Suganya, 2014), or to investigate the economic hubs and authorities of the world trade network (Deguchi et al., 2014).

3.3 Social Network Analysis

As mentioned in Section 3.1 (A Brief History of the Webometric Research), the social science research techniques have been applied to webometrics. One of the areas of the social science research is Social Network Analysis (SNA), which is closely related to web link analysis (Guns et al., 2011; Kretschmer and Aguillo, 2004; Scott, 2012). A social network is a structure made up of social entities (people or organisations), and their relationships. 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 research has focused on the position of the individual node in the network and its interaction with adjacent entities. There have been introduced many network measures that help researchers to quantify 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).

3.3.1 Centrality Measures

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, 1978). Hanneman and Riddle (2005) described other central-

ity 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 and quantifies the importance of that node. 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.

Degree Centrality

Degree centrality represents the number of all connections or connection weights of a single node in a network. There are defined in-degree and out-degree centralities in a directed network, which represent the number of incoming and outgoing connections. It is stated, that the node with a high number of connections is more central and has a greater ability to influence others in the graph structure. A node, which is linked a lot (high in-degree) is often coveted by other nodes, and it is known as popular or prestige (see Section 3.3.2, Prestige Measures). A node, which links to other nodes a lot (high out-degree) is classified as an influential node, which has a greater chance to influence the others. That kind of node is known as prominent.

The value of degree centrality is necessary to normalise between 0 (minimum degree) and 1 (maximum degree) to be able to compare the nodes of different graphs. The normalisation is performed by dividing of the degree of a node by the maximum possible number of links that the node may have, i.e. $(n-1)$, where n is the total number of nodes of a particular graph. The normalised degree centrality (Freeman, 1978) of an actor i (denoted by $C'_D(i)$) is therefore in an undirected graph defined as the node degree $d(i)$ divided by the maximum degree $(n-1)$:

$$C'_D(i) = \frac{d(i)}{n-1} \tag{3.4}$$

Graph centralisation expresses the degree of inequality in the whole graph. The centralization is defined as an average of all the degrees in a network compared against the average degree of a perfect star network of the same size (Hanneman and Riddle, 2005). A graph with the high centralisation value may consist of a majority of nodes with the low degree and just one node with the high degree that

usually being in the centre of the graph. Conversely, the low graph centralisation may consist of the nodes with an equal degree and position, so there is no single node with a high power.

Closeness Centrality

Closeness centrality is based on the sum of the shortest distances from a single node to the other nodes in a network. The sum of distances expresses how long it takes to disseminate information from a particular node to all other nodes, or how the node is close to all the others. A small sum value means that every part of the network could be reached through a relatively short chain of people. It follows that the node with the shortest distances to the others has the highest centrality; therefore, the closeness centrality is defined as the inverse of the sum of the shortest distances.

If the entire graph is not represented as one connected component, then the closeness centrality must be calculated separately for all its components. The result value of each component must be normalised, i.e. the sum of the shortest distances is divided by $(n-1)$, where n is the total number of nodes of the component. A comparison of closeness centrality of different graphs is performed in the same way. The normalised closeness centrality (Freeman, 1978) of an actor i (denoted by $C'_C(i)$) is therefore in an undirected graph defined as the inverse of the sum of the shortest distances $d(i,j)$ between the node i and the other nodes divided by the $(n-1)$:

$$C'_C(i) = \frac{n-1}{\sum_{j=1}^{n} d(i,j)} \tag{3.5}$$

There is not defined a direct correlation between degree and closeness centrality; therefore, the high degree does not mean high closeness (Opsahl et al., 2010). Degree centrality represents the connections a node has, while closeness centrality expresses a node position in the whole network.

Betweenness Centrality

Betweenness centrality expresses the ability of a node to connect (different) groups of nodes. The node with a high betweenness centrality has an important role in connecting different groups. For instance, an actor plays in two theatres, and thus he connects two different groups of actors from both theatres. If the actor is the only one who participates in both theatres, then the actor will have the greatest

betweenness centrality in the graph formed by the all other actors. Such node would have a significant influence on events in the entire graph, for example by blocking of messages from one group to another, or it can isolate some persons who have no choice to connect with a second group.

It is obvious that betweenness centrality is based on the number of paths between every two nodes in a graph that pass through a specific node. The betweenness centrality (Freeman, 1977) of a node i (denoted by $C_B(i)$) is therefore calculated as the sum of the number of shortest paths between nodes j and k that pass through i (denoted by $p_{jk}(i)$, $j \neq i$ and $k \neq i$) divided by the number of all paths between j and k (denoted by p_{jk}):

$$C_B(i) = \sum_{j<k} \frac{p_{jk}(i)}{p_{jk}} \tag{3.6}$$

As with the previous metrics, the value of betweenness centrality is necessary to normalise to be able to compare the nodes of different graphs. The normalisation is performed by dividing of the node's betweenness centrality by the maximum possible number of paths that the graph may have, i.e. $(n-1)(n-2)$ for a directed graph, and $\frac{(n-1)(n-2)}{2}$ for undirected graph, where n is the total number of nodes of a particular graph. The normalised betweenness centrality of an actor i (denoted by $C'_B(i)$) is therefore in an undirected graph defined as:

$$C'_B(i) = \frac{2\sum_{j<k} \frac{p_{jk}(i)}{p_{jk}}}{(n-1)(n-2)} \tag{3.7}$$

Another variant of betweenness centrality (Hanneman and Riddle, 2005) utilises all paths between two nodes, so not only the shortest paths. As a result, there may not be used just the shortest path for the communication between two nodes, however, if the shortest path is unavailable, then the second shortest is used, and so on.

3.3.2 Prestige Measures

Prestige was first examined in a sociological research in 1913 when Thomas H. C. Stevenson and statisticians in the British Civil Service performed the social ladder research (General Register Office, 1913). In the real world, the prestige is a basic indicator of a social status of an individual; and the same meaning is also applied

in Social Network sphere, where a prestigious actor is one who is more trusted by others.

The prestige is as well as centrality a single node feature. The main difference between these two measures is that centrality more focuses on links pointed out of the node (out-links), while prestige focuses on links pointed to the node (in-links). Hence, the prestige can be computed for directed relations and directed graph only (Wasserman and Faust, 1994).

Degree Prestige

Degree Prestige is the simplest measure of prestige which is derived from degree centrality. Based on the definition, an actor is prestigious if it receives many in-links. A greater number of in-links indicates a greater prestige.

The value of degree prestige is necessary to normalise between 0 (minimum degree) and 1 (maximum degree) to be able to compare the nodes of different graphs. The normalisation is performed by dividing of the degree of a node by the maximum possible number of in-links that the node may have, i.e. $(n-1)$, where n is the total number of nodes of a particular graph. The normalised degree prestige (Wasserman and Faust, 1994) of an actor i (denoted by $P'_D(i)$) is therefore in a directed graph defined as the node in-degree $d_I(i)$ divided by the maximum in-degree $(n-1)$:

$$P'_D(i) = \frac{d_I(i)}{n-1} \tag{3.8}$$

Proximity Prestige

Proximity prestige of an actor is based on the average distance of other actors that are in an influence domain of the measured actor. The influence domain of an actor is the set of other actors that can reach the actor. A larger influence domain and a smaller distance evoke a higher proximity prestige value. The average distance is calculated as the sum of shortest distances (shortest path distance from j to i is denoted by $d(j,i)$) of all actors in the influence domain I_i of actor i divided by the size of the influence domain (denoted by $|I_i|$):

$$\frac{\sum_{j \in I_i} d(j,i)}{|I_i|} \tag{3.9}$$

The proximity prestige (Wasserman and Faust, 1994) of an actor i (denoted by $P_P(i)$) in a directed graph is defined as the ratio of the proportion of actors who can reach i to the average distance these actors are from i:

$$P_P(i) = \frac{\frac{|I_i|}{n-1}}{\sum_{j \in I_i} \frac{d(j,i)}{|I_i|}} \tag{3.10}$$

As with the previous metrics, the value ranges between 0 and 1. In one extreme, the nominator, proportion of actors who can reach the actor i, is equal to 1 if every actor can reach the actor i. The denominator is equal to 1 if every actor is adjacent to i. Then the proximity prestige reaches the maximum, and it is equal to 1. The other extreme, the proximity prestige is equal to 0 when an actor has no influence domain.

Rank Prestige

Rank prestige is a measure that reflects the prominence of the individual actors who do the "voting" for the actor. That is a main difference from the above two prestige measures because they consider just in-degrees and distances. However, the rank prestige of an actor depends on the ranks of those who have voted for the actor, and the ranks of those who have voted rely on the ranks of the actors who voted them, and so on. To quantify this infinite regress, the rank prestige (Wasserman and Faust, 1994) of an actor i (denoted by $P_R(i)$) is defined as a linear combination of links that point to i:

$$P_R(i) = x_{1i}P_R(1) + x_{2i}P_R(2) + ... + x_{ni}P_R(n) \tag{3.11}$$

where x_{ji} takes the value 1 if j points to i or 0 otherwise. For instance, if actor number 2 is voted by actors with number 3 and 7, so that $x_{32} = x_{72} = 1$ and all the other $n-2$ actors do not vote, then the rank prestige for this actor is defined as $P_R(2) = P_R(3) + P_R(7)$. If actors 3 and 7 have a high rank, so the actor 2 will have it too. The proximity prestige increases if high ranking actors vote the actor.

3.4 Web Mention Analysis

Web Mention Analysis (Cronin et al., 1998; 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 3.6, Sentiment Analysis), which allows as to detect opinions automatically from text.

3.5 Social Media Analysis

Social media is a collection of online communication channels that allow people to create, share and exchange information among themselves or within a virtual community. Social media includes many different forms, for instance, blogs, microblogs, social networks, forums, and wikis. 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. Moreover, thanks to modern technologies, people may share that information almost immediately as soon as they are aware of them.

One of the first communication channels were blogs, which serve as users' diaries available on the Web. The scope of blog topics includes the range from personal diaries through the official business news up to the political campaigns. There have been many researches, which use blogs as a data source for their research. For example, the analysis of the spread of news created by amateur journalists in blog space and the visualisation of news distribution (Takama et al., 2007). There have also been efforts to find information on blogs (Han et al., 2009) or detect a trend among the most published topics (Thelwall, 2007).

BlogPulse was a blog search engine that monitored the daily activities on blogs and discovered trend information (Glance et al., 2004). The engine was primarily based on counting how often searched words were mentioned online (see Section 3.4, Web Mention Analysis). BlogPulse collected data from blogs and provided chronological summaries and insights into the public opinion on specific search topics. One of the outputs from BlogPulse is showed in Figure 3.2. In the picture is a chart which represents the public interest in the "Apple Worldwide Developers Conference" organised in 2010. The x-axis represents the published date, and the y-axis shows the percentage of all blog posts. The conference began on the 7th of June and finished on the 11th of June. As illustrated, almost no one was writing about the conference

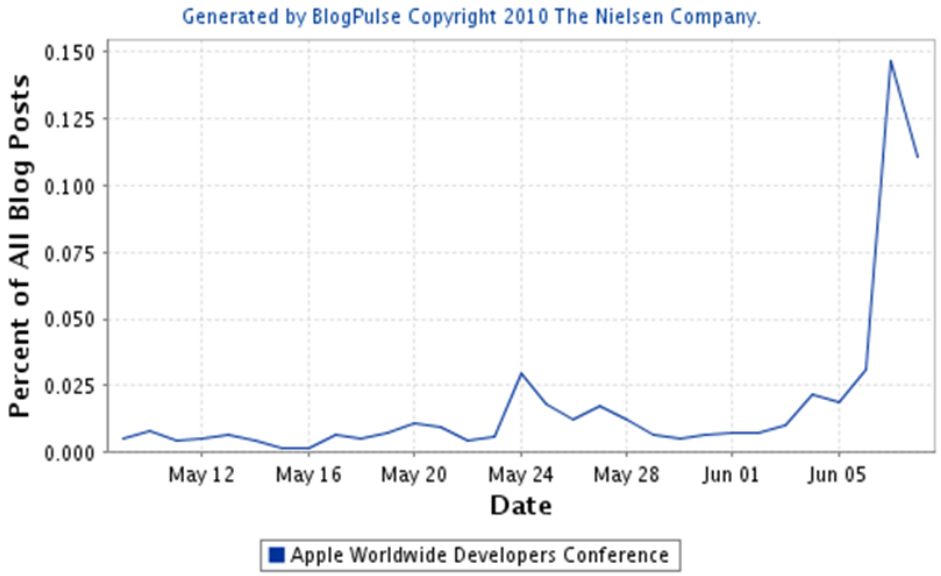


FIGURE 3.2: Blog trend chart of the Apple Worldwide Developers Conference organised in 2010.

before it starts. However, blogging had been rapidly rising during the conference. By the way, this presentation unveiled the new Apple iPhone. So many bloggers wrote about this new phone at that time.

Blogs are especially popular among people in the form of microblogs (Statista Inc., 2016a). A microblog is very similar to a blog, except the length of its content that generally does not exceed 200 characters, which may be the major reason for their popularity (Thelwall et al., 2011; Aichner and Jacob, 2015). Thanks to modern technologies, and especially due to the widespread usage of mobile devices, people may share whenever and whatever they want. That can be used as the advantage during disasters or crises when people may share information to improve the management of these emergencies (Terpstra et al., 2012).

Many web-based systems have been developed to automatically analyse messages (called "tweets") from Twitter[2] microblogging service. For instance, the TweetCred (Gupta et al., 2014), that automatically evaluates the credibility of posted tweets. The researchers have developed an algorithm that analyses a tweet content and author's characteristics and based on that determines the credibility rating of a

[2]http://twitter.com

tweet. Another example (Nagy and Stamberger, 2012) uses SentiWordNet to detect the basic sentiment of a tweet. Their approach excels primarily due to a self-created list of emoticons and slang words along with their sentiment evaluations that are not represented in the SentiWordNet lexicon (see Section 3.6.1.1, SentiWordNet Lexicon). One of the more complex systems, the Twitcident (Abel et al., 2012), performs real-time monitoring of Twitter messages. The system automatically detects an incident and starts tracking and filtering messages about that. Twitcident continuously profiles the incident based on the message stream and allows users to retrieve particular information and analyse the current situation on the Social Web.

3.6 Sentiment Analysis

Sentiment Analysis or Opinion Mining (Pang and Lee, 2008) enables us to detect opinions automatically from structured but also unstructured data. That involves several research areas such as natural language processing, computational linguistic and text mining. The studies in this field originated from the demand of commercial companies, who wanted to know the public opinion on price, quality and other features of their products.

Before the massive spread of the Internet, companies had been gaining customer feedback via phone, email surveys or interviews. It was very slow, expensive and annoying for some customers. With the Internet, companies may be able to gain feedback from comments in e-shops, blogs, customer reviews, social networks, etc. These methods of obtaining information are very fast and thanks to that the company that has just launched a major advertising campaign may gain quick public feedback on its impact.

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). As a subtask, the analysis is capable of determining the attitude of a writer on a specific topic. The attitude may express an affective state of the author when writing or intended emotional effect which author wishes to present to the reader.

3.6.1 Sentiment Classification

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. The main goal is to quickly identify the classification of the opinion on an object, which is described in the analysed text. The task is partly similar to the webometric mention analysis, which is focused on topic-related words (Han et al., 2009), e.g. sports, politics, industrial sectors, etc. However, the sentiment classification, in contrast to webometrics, is focused on opinion-related words (Pang and Lee, 2008), e.g. excellent, great, horrible, bad, etc.

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 algorithms usually start with simple rules (i.e., unsupervised machine learning) or with a collection of text that was previously annotated by a human for sentiment polarity and strength (i.e., supervised machine learning). According to the data input, the algorithm acquires knowledge of features, which identify a polarity of the text (Liu, 2012; Mejova and Srinivasan, 2011; Singh et al., 2013). The recognised features typically consist of one to three words (i.e., unigrams, bigrams, trigrams) which associate with the sentiment. 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 and Paroubek, 2010). A disadvantage of this approach is that the algorithm can extract non-sentiment features because they are frequently used in a text, and they can be wrongly used as sentiment features for non-annotated text. For instance, famous people's names may typically be associated with strong positive or negative meaning. Hence, the machine learning approach is rather used to identify patterns that are not directly sentiment related (Thelwall and Buckley, 2013).

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: Linguistic Inquiry and Word Count (LIWC) (Pennebaker, 2011; Tausczik and Pennebaker, 2010), Princeton WordNet (Fellbaum, 2005, 2010), SentiWordNet (Baccianella et al., 2010; Esuli and Sebastiani, 2006), 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 word and using heuristics to predict the sentiment of another word. For instance, two words divided by "and" may have the same polarity, and the words divided by "but" may have opposite polarity (Hatzivassiloglou and McKeown, 1997). The main problem of lexicon-based approach is that the incorrect sentiment, e.g. related to a different topic, can be assigned to the word. For instance, the word "nice" has a strong positive meaning. However, the meaning may also be neutral in the context of a city in south-eastern France.

Linguistic analysis studies grammatical and linguistic structures of analysed text, and it also tries to predict polarity of those structures in conjunction with the lexicon or machine-learning methods. An example of the use of linguistic analysis is part-of-speech (POS) tagging (Toutanova et al., 2003). The part-of-speech is a language category of a word that is defined by its syntactic or morphological behaviour. The part-of-speech tagger assigns part of speech to each word in a sentence, and it also recognises finite/infinitive and plural/singular form of the word. The part-of-speech tagger was for instance used to enrich textbooks produced from India, which are not written well and they often lack adequate coverage of important concepts (Agrawal et al., 2010).

3.6.1.1 SentiWordNet Lexicon

SentiWordNet (SWN) is one of the most widely used sentiment analysis lexica of English words (Baccianella et al., 2010). The SWN lexicon is an extension of Princeton WordNet dictionary (Fellbaum, 2010) and like that the words in SWN are grouped into sets of synonyms called synsets. SentiWordNet enriches each of these synsets by the sentiment, which represents an estimated degree of positivity, negativity, and neutrality. These three portions are described by a vector of scores $(score_{pos}, score_{neg}, score_{obj})$, where the sum of these scores is always equal to one. For example, the vector $(0.625, 0.000, 0.375)$ is assigned to the word "flush"; the sum of all scores of this vector is equal to one.

In Table 3.1, several senses of the word "flush" are presented. Each entry has assigned a part-of-speech (a – adjective, n – noun, r – adverb, v – verb) and a sense number where a lower number indicates a more frequent occurrence of the sense. Obviously, different word senses may have different polarities. Entries *flush#a#2*, *flush#n#1* and *flush#v#1* have a negative score only, whereas *flush#n#2* has

TABLE 3.1: Selected SentiWordNet entries for the word "flush". Each entry represents one of the word senses with a different sentiment. Pos, Neg and Obj represent the sentiment strength. A synset is a set of synonyms of the word.

Word#pos#sense	Pos	Neg	Obg	Synsets
flush#a#1	0	0	1	-
flush#a#2	0	0.250	0.750	wealthy, moneyed, loaded, affluent
flush#n#1	0	0.125	0.875	prime, peak, heyday, flower, efflorescence, blossom, bloom
flush#n#2	0.625	0	0.375	rosiness, blush, bloom
flush#r#1	0	0	1	-
flush#r#2	0	0	1	-
flush#v#1	0	0.125	0.875	redden, crimson, blush
flush#v#2	0	0	1	-

just a positive score. Other entries with objectivity equal to one have no sentiment importance and thus are not used too much in sentiment analysis. Synsets column contains a list of synonyms that have the same scores.

3.6.2 Sentiment Sense Disambiguation

As mentioned above, the main problem of the lexicon-based method is that the word may have multiple senses with a different sentiment polarity and strength, and it is hard to recognise which sense should be used in a specific context. There are several strategies for computing a prior polarity for all word senses (Gatti and Guerini, 2012) and the sense disambiguation. All strategies use part-of-speech tagger as a preliminary step. The determination of a word class may significantly reduce the number of word senses and thereby simplify disambiguation in the next step.

The simplest approach is based on a random selection of one of the word senses. For instance, sentiment "nice" has defined seven senses; six of them are adjective, and one is a noun. A part-of-speech tagger is used to recognise a word class in an analysed text, and the word may be marked as an adjective. The random sense disambiguation approach is then used to randomly select one of the six adjective senses and use its sentiment for the word. However, this approach is not much useful and reliable since the selected sense might not reflect a real word meaning. The more widespread approach uses only the first sense and skips other senses; this is equivalent for *word#pos#1* in SentiWordNet (Agrawal and Siddiqui, 2009; Malinský and Jelínek, 2012). This strategy utilises the fact that the first sense is the most frequent, however, the results may not always be very reliable.

Other approaches treat all word senses to calculate prior sentiment of the word. One approach (Equation 3.12) adds positive and negative parts of all senses and divides it by the number of those senses (Denecke, 2009; Sing et al., 2012; Thet et al., 2009). Another approach (Equation 3.13) differentiates the number of positive and negative words (Fahrni and Klenner, 2008; Neviarouskaya et al., 2009). Moreover, different approach prefers to calculate with some positive/negative senses instead of using scores (Neviarouskaya et al., 2011).

$$word_{pos} = \frac{\sum_{i=1}^{n} score_{pos}}{n} \quad \text{and} \quad word_{neg} = \frac{\sum_{i=1}^{n} score_{neg}}{n} \tag{3.12}$$

$$word_{pos} = \frac{\sum_{i=1}^{n} score_{pos}}{count_{pos}} \quad \text{and} \quad word_{neg} = \frac{\sum_{i=1}^{n} score_{neg}}{count_{neg}} \tag{3.13}$$

Some studies calculate the score as a sum of geometric series (Chaumartin, 2007) or harmonic series (Denecke, 2008). They use the assumption that more frequent senses are more important than those at the end of the series (Equation 3.14 and 3.15). The final score is the difference between positive and negative parts of the word.

$$word_{pos} = \frac{\sum_{i=1}^{n} \frac{1}{2^{i-1}} score_{pos}}{n} \quad \text{and} \quad word_{neg} = \frac{\sum_{i=1}^{n} \frac{1}{2^{i-1}} score_{neg}}{n} \tag{3.14}$$

$$word_{pos} = \frac{\sum_{i=1}^{n} \frac{1}{i} score_{pos}}{n} \quad \text{and} \quad word_{neg} = \frac{\sum_{i=1}^{n} \frac{1}{i} score_{neg}}{n} \tag{3.15}$$

An entirely different strategy is based on exploiting WordNet gloss for comparison with analysed text. The gloss is a textual description that briefly describes each word in the lexicon. Based on this approach, for instance, was developed an algorithm that chooses the sense of the word whose gloss contains the largest number of words presented in the text (Benedetti, 2013). However, the strategy may not be optimal for very short sentences or for sentences which words are not included in a gloss.

Chapter 4

A Novel Web Metric for the Evaluation of Internet Trends

According to Statista Inc. (2016b), the social network penetration worldwide is ever-increasing, and social networking is one of the most popular activities on the Web. More and more people access social networks, blogs, discussions (collectively Web 2.0) and share their thoughts and opinions on various topics, products and events around them. Some of their comments might be totally unimportant to the other Internet users. However, many of them are very useful and do not just for ordinary users, but also for some commercial companies that would like to know the public opinion on price, quality and other factors of their products. Thanks to that, the Web 2.0 becomes a rich information source for social science research since it contains a large number of ideas on various topics from many different users. However, 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 for common Internet users. It is, therefore, necessary to design a suitable metric for such volume of information that would reflect a semantic content of single pages in a better way.

Webometrics is a scientific discipline that tries to measure the World Wide Web, and thus it can serve as a suitable solution for analysis in such heterogeneous environment. Original webometric techniques (see Chapter 3, Webometric Background and State of the Art) improve searching and provide a trend detection. However, they are not able to distinguish a polarity of a text and its semantic meaning. On the other hand, webometrics is purely a quantitative approach to the Web, which

can be enhanced by qualitative methods and thereby it allows us to expand the possibilities of a study problem.

Our research emphasis has been placed on the extension of Webometrics by the methods of Sentiment and the Social Network Analysis. The main focus is the analysis and evaluation of 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 extension of Webometrics by a combination of these methods leads up to gaining insights into the public opinion with respect to some topic, and a better machine understanding of a text. Better machine understanding of the content on the Web might have a significant impact on the quality of a website evaluation.

4.1 A General Model for Gathering and Processing Data from Web 2.0

The basis of the work is the general model (Figure 4.1) for gathering and processing data from Web 2.0. The model represents the simplest way for an Internet user to obtain relevant information from the Web. A functionality of the model is similar to a typical web search engine; however, the model focuses on the evaluation of Internet trends. The model builds on Webometrics (Aguillo et al., 2010; Thelwall, 2009) and starts from the idea that almost any text can be machine-recognised. This idea is supported by the current research in Sentiment Analysis (Potthast and Becker, 2010; Prabowo and Thelwall, 2009; Thelwall and Buckley, 2013), which aims at the sophisticated analysis of sentences using mathematical and statistical methods and linguistic analysis of a text. The model consists of several essential parts:

- **Crawler** - the automated unit that follows links on the Web and creates a copy of all the visited pages.
- **Content Analysis** - the algorithm unit which analyses the crawled pages and stores their key content in the database.
- **Index** - the repository for analysed web pages, which returns a list of the result pages in correlation to user's query.
- **Query Unit** - the unit for processing the user query into a format that the index can understand.

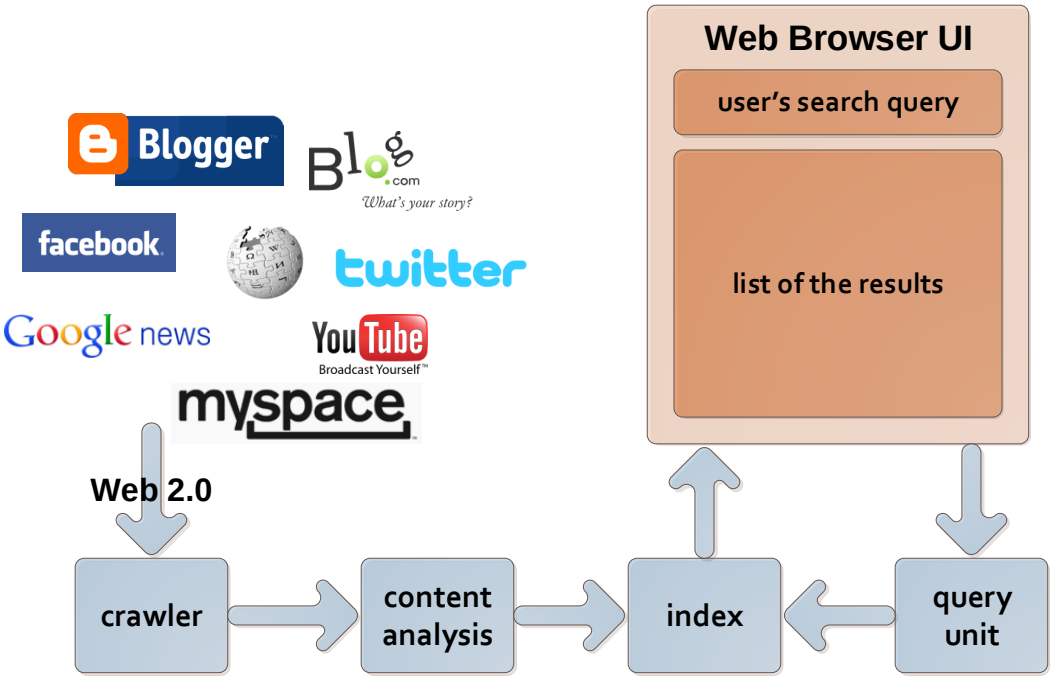


FIGURE 4.1: A general model for gathering and processing data from Web 2.0.

It can be assumed that the extension of the model by specific methodologies that are presented in later sections will improve the trend evaluation and thereby facilitate users' access to the information on the Web. The evaluation system for gathering and processing data from blogs has been created and implemented to verify our theoretical assumptions. The system is primarily focused on the content analysis part of the model. The final version that implements all of the introduced methodologies is described in detail in Chapter 5 (Framework for the Analysis of Internet Trends).

4.2 Evaluation System for Gathering and Processing Data

The evaluation system (Figure 4.2) is an implementation of the general model, which is used to verify our assumptions about the applicability of the Sentiment Analysis to the evaluation of Web content (Malinský and Jelínek, 2010, 2011b). The system accepts two types of data input. The first input includes the list of trends that will be evaluated, and the second represents the content of websites that will be analysed. The output of the system is a daily chronological evaluation for each trend.

It would be great to have very popular trends that are often mentioned online to allow the system to provide a relevant output. Google Trends, the service which reflects what people are searching for on the Internet, can serve as a good source to select popular Internet trends. Google Trends algorithm analyses web searches that are performed on Google search engine and provides the list of hot searches, which deviate the most from their historical traffic pattern. The service provides the list of ten fastest-rising search queries for user defined country and year.

The second input of the system is a list of blogs whose contents serve as the source for analytical algorithms to analyse the selected trends. The scope of blog topics includes the diverse range of the personal diaries through the official business news up to the political campaigns. Over millions of people post information about events around them and they also share opinions on specific topics, e.g. political situation, travel information, technology review or gossips about celebrities.

As mentioned above, the evaluation system is based on the proposed general model. The Hot Searches and the Blog Search units represent the Crawler. In the first mentioned unit, the ten the most searched expressions are retrieved from Google Trends over a given period. These expressions are used as a trend input for Blog Searching. The Blog Search unit modifies each of the received multi-word trend to a Boolean search expression to match a relevant blog post. For example, the expression "ottawa earthquake" is restructured to "ottawa AND earthquake". However, names and specific expressions, e.g. "bill gates", remain the same, and they are searched in blogs as an exact phrase. The Blog Search unit goes through the list of provided blogs and tries to find each prepared expression in a given time interval around the trend deviation. The trend deviation represents the day when the trend has been searched the most. This approach, to search in the interval around the trend deviation provides a chronological summary of daily blog posts for each trend.

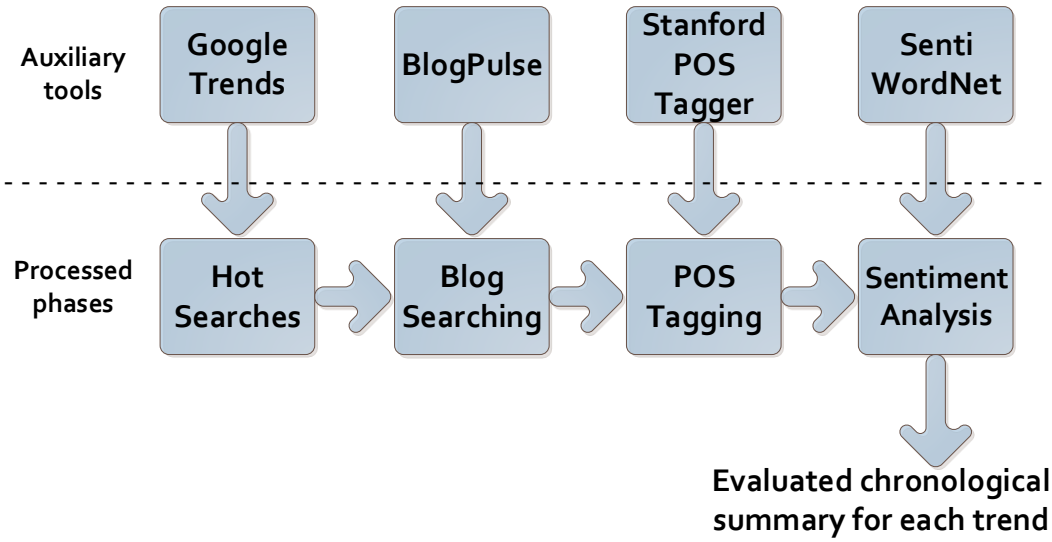


FIGURE 4.2: Evaluation system for gathering and processing data from blogs.

The POS Tagging and the Sentiment Analysis units represent the Content Analysis part of the general model. The Part-of-Speech Tagging unit parses a list of sentences for each blog post and further assigns a part-of-speech tag to each word in a sentence. The plural words are converted into singular for a more accurate word recognition in the Sentiment Analysis unit. The words are further divided into the four part-of-speech categories for a better search in the SentiWordNet lexicon: adjective, noun, adverb, verb. The Sentiment Analysis unit operates with the lexicon of words, where each word is associated with its polarity and strength. The unit determines a polarity and strength for each tagged word and further evaluates all sentences for each day based on the word designation. The evaluation of each trend is then performed for each day according to the following rules:

1. A word is positive if it has more positive score than negative score, and vice versa.

2. A sentence is positive if it contains more positive words than negative words.

3. If a sentence has the same number of positive and negative words, then the polarity of the sentence is determined by the sum of scores of the individual words.

4. The sentence is positive if the sum of positive scores is greater than the sum of negative scores, and vice versa.

5. A positive evaluation of a trend is determined for a specific day by the sum of positive sentences that have been written about the trend. Whilst, a negative evaluation is determined by the sum of negative sentences.

The rules can be written in a formal mathematical definition:

Definition 4.2.1 (alphabet):
Let Σ be an alphabet, a non-empty finite set. Elements of Σ are called characters.

Definition 4.2.2 (word):
A word is any n-tuple of characters from Σ.

Definition 4.2.3 (sentence):
Let L be a language, a set of all words. A sentence S is any n-tuple of words from L.

Definition 4.2.4 (polarity, strength):
A polarity is the property of word, sentence, or trend, which expresses a feeling toward something. Let $I = [-1, 1]$ be an interval of the real line. The interval I is partitioned into three subintervals, $I = I_1 \cup I_2 \cup I_3 = [-1, 0) \cup \{0\} \cup (0, 1]$. Values from the subinterval I_1 refer to a negative polarity, I_2 refer to an objective polarity, and I_3 refer to a positive polarity. Strength is a real number from the interval I that defines a degree of the polarity. The value -1 is the strongest negative strength, while 1 is the strongest positive strength.

Definition 4.2.5 (sentiment):
Let $W \subset L$ be a set of all words that are identified by SentiWordNet. A sentiment is the property of a word $w \in W$, which represents a degree of positive, negative, and objective polarity. These three portions are described by a vector of scores $(score_{pos}, score_{neg}, score_{obj})$, where each score $\in [0, 1]$, and the sum of all scores is equal to one:

$$(score_{pos} + score_{neg} + score_{obj}) = 1 \tag{4.1}$$

Definition 4.2.6 (word polarity):
A word $w \in W$ is positive if its sentiment has more positive score than negative score, and vice versa:

$$positive\ word:\ word_{pos} \Leftrightarrow score_{pos} > score_{neg} \tag{4.2}$$

$$negative\ word:\ word_{neg} \Leftrightarrow score_{pos} < score_{neg} \tag{4.3}$$

Definition 4.2.7 (word strength):
A word strength $word_{strength}$ defines a degree of the word polarity.

$$word_{strength} = score_{pos} - score_{neg} \tag{4.4}$$

Definition 4.2.8 (sentence polarity):
A sentence is positive if it contains more positive words than negative words. If a sentence has the same number of positive and negative words, then the polarity of the sentence is determined by the sum of scores of the individual words. The sentence is then positive if the sum of positive scores is greater than the sum of negative scores, and vice versa:

$$\begin{aligned} \text{positive sentence: } sentence_{pos} \Leftrightarrow & \sum word_{pos} > \sum word_{neg} \\ \vee & \sum word_{pos} = \sum word_{neg} \\ \wedge & \sum score_{pos} > \sum score_{neg} \end{aligned} \tag{4.5}$$

$$\begin{aligned} \text{negative sentence: } sentence_{neg} \Leftrightarrow & \sum word_{neg} > \sum word_{pos} \\ \vee & \sum word_{neg} = \sum word_{pos} \\ \wedge & \sum score_{neg} > \sum score_{pos} \end{aligned} \tag{4.6}$$

Definition 4.2.9 (sentence strength):
A sentence strength $sentence_{strength}$ defines a degree of the sentence polarity.

$$sentence_{strength} = \frac{\sum score_{pos} - \sum score_{neg}}{|word_{pos}| + |word_{neg}|} \tag{4.7}$$

Definition 4.2.10 (trend):
A trend is any word or subsequence of words from sentence S.

Definition 4.2.11 (trend polarity):
Let Υ *be an ordered set of sentences. A trend is positive if* Υ *contains more positive sentences then negative sentences, and vice versa:*

$$trend\ is\ positive:\ trend_{pos} \Leftrightarrow \sum sentence_{pos} > \sum sentence_{neg} \tag{4.8}$$

$$trend\ is\ negative:\ trend_{neg} \Leftrightarrow \sum sentence_{neg} > \sum sentence_{pos} \tag{4.9}$$

Definition 4.2.12 (trend strength):
A trend strength $trend_{strength}$ *defines a degree of the trend polarity.*

$$trend_{strength} = \frac{\sum sentence_{strength}}{|sentence_{strength}|} \tag{4.10}$$

Section 6.1 (Chronological Evaluation of Internet Trends) describes the experiment that has been used to verify the theoretical assumptions. A corpus of the blog posts has been used for the evaluation of the most searched expressions in the Google search engine. The results of the experimental system represent a chronological view of the trend evaluation according to the public opinion.

4.3 Comparing Methods of Trend Assessment

Original webometrics includes the techniques, which could be used for the evaluation of Internet trends (see Chapter 3, Webometric Background and State of the Art). One example is Web Mention Analysis (see Section 3.4, Web Mention Analysis), which evaluates the trends by counting how often they are mentioned online. Conversely, Sentiment Analysis does not evaluate the number of words, but the strength, i.e. it reflects the polarity of a text and helps to recognise its semantic meaning.

Another example is Social Network Analysis which represents a set of techniques for the analysis of interactions and relationships between actors in a social network. An analysis output can be reported as a network diagram with nodes and links (see Section 3.3, Social Network Analysis). The nodes, people or groups of people, are called *actors*; and the links, social interaction (such as friendship) between actors, are called *ties*. However, the node may also be imagined as a trend, and the tie between nodes may represent the trend evaluation.

The mentioned methods use a different methodology to the trend assessment; Web Mention Analysis uses frequency, Sentiment Analysis uses polarity, and Social

Network Analysis uses source quality. Each of these techniques is mostly used separately, but they could be utilised together and take advantage of all their properties. The methodology proposed in the previous section has been compared with the selected methods that can be used for evaluation on the Web (see Section 6.2, Comparing Methods of Trend Assessment). A corpus of the movie reviews has been served for this study. There have been found that the combination of individual methods can provide much more accurate results with respect to the desired area of interest.

4.4 Trend Evaluation in the Social Network Sphere

The new metric has been defined for the evaluation of trends in social networks (Malinský and Jelínek, 2016a,b). The metric 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 4.3 shows an example of the social network where it is possible to evaluate Internet trends.

Definition 4.4.1 (directed graph):
A directed graph $G = (N, E)$ consists of a non-empty 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 4.4.2 (trend):
Let $t \in N$ be a trend, and $a \in N$ be an actor. 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 t.

Definition 4.4.3 (actor):
Let $a \in N$ be an actor, and $t \in N$ be a trend. An actor $a \in N$ is the node of a directed graph $G = (N, E)$ for which it holds that every edge $(a, t) \in N$ is pointing out of the node a.

Definition 4.4.4 (comment):
A comment $c \in E$ is the edge of a directed graph $G = (N, E)$ that points from actor $a \in N$ to trend $t \in N$.

TABLE 4.1: 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.

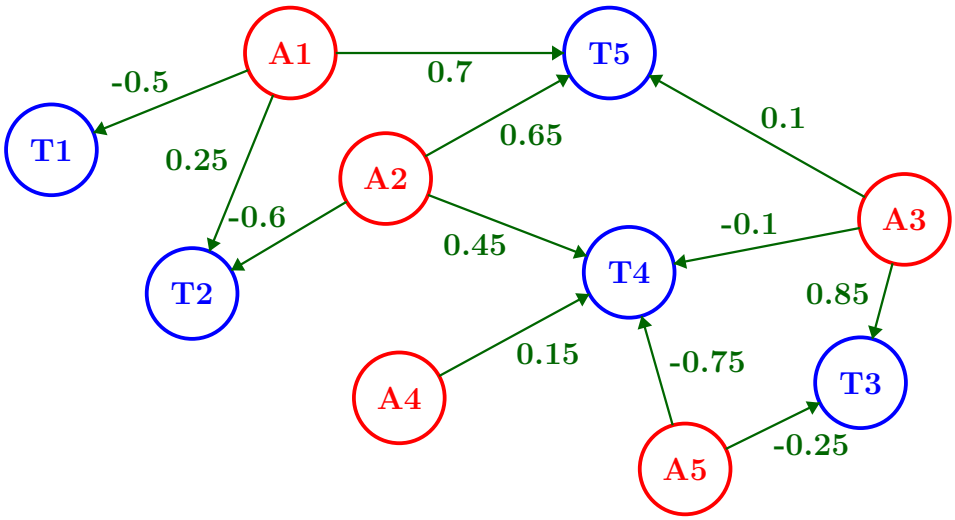
Node	T1	T2	T3	T4	T5
A1	1	1	0	0	1
A2	0	1	0	1	1
A3	0	0	1	1	1
A4	0	0	0	1	0
A5	0	0	1	1	0

Definition 4.4.5 (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 $\Theta = (T, A, C)$ 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 4.1 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 4.3).

In the second step, a centrality measure (see Section 3.3.1, Centrality Measures) 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 (see Section 3.3.2, Prestige Measures). However, it calculates with the out-links in contrast to the prestige that counts with 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 it 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.

FIGURE 4.3: 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 on edge represents the actor's evaluation of the comment.



Definition 4.4.6 (degree power):
Let $d_o(a)$ be a number of comments $c \in C$ the actor $a \in A$ has written about all trends from T. Degree Power $PW_D(a)$ of the actor a is in the network of trends $\Theta = (T, A, C)$ defined as $d_o(a)$ divided by the maximum possible comments $|C|$ the actor may have made.

$$PW_D(a) = \frac{d_o(a)}{|C|}; \quad a \in A \tag{4.11}$$

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 4.4.7 (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, the Sentiment Analysis approach described in the previous section is used to evaluate all comments of the actors whose degree power is above the threshold. All comments (the edges between actors and trend) are evaluated by the normalised polarity in the range of $[-1, 1]$. A negative evaluation of a comment c from actor a that has been written about a trend t is denoted by $c^-(a,t)$, while a positive evaluation is denoted by $c^+(a,t)$. A set of all negative comments that have been written about a trend t is denoted by C_t^-, while a set of positive comments is denoted by C_t^+.

A negative strength $strength_{neg}(t)$ of a trend t is calculated as the sum of all negative evaluations $c^-(a,t) \in C_t^-$ divided by the number of negative comments that have been written about t (Equation 4.12), and vice versa for a positive strength (Equation 4.13). The overall trend evaluation is calculated as the difference between positive and negative strengths 4.14.

$$strength_{neg}(t) = \frac{\sum c^-(a,t)}{|C_t^-|}; \quad c^-(a,t) \in C_t^- \tag{4.12}$$

$$strength_{pos}(t) = \frac{\sum c^+(a,t)}{|C_t^+|}; \quad c^+(a,t) \in C_t^+ \tag{4.13}$$

$$strength_{overall}(t) = strength_{pos}(t) - strength_{neg} \tag{4.14}$$

A case study that uses this approach is described in Section 6.3 (Movie Evaluation in the Network of Trends). The movie reviews have been used for the evaluation.

4.5 Sentiment Sense Disambiguation

An important aspect of the public opinion is sentiment, which expresses whether people feel positive or negative towards some product or event. Studies of the sentiment classification have been arising, and they try to automatically detect a writer's opinion on some topic. The topic of an analysed text specifies which sentiment domain is used for analysis. For instance, the word "wine" is defined at least by two domains, colour, and beverage, and it can be difficult to recognise which sentiment domain is used in a text. Studies to many different domains have been presented, for example, to use a restaurant data set to determine personalised location recommendation (Yang et al., 2013), to determine hotel rating based on guest reviews (López Barbosa et al., 2015; Wang et al., 2011) or to rate movie

reviews (Dhande and Patnaik, 2014). Other studies draw on these domain-specific algorithms and try to use them for data analysis of similar or a different domain (Vilares et al., 2015). The significant problem of sentiment studies is that the analysis algorithm can be highly topic dependent (Rastogi et al., 2014), and it can be difficult to predict correct sentiment for a text under a different domain.

Two new methods have been proposed to improve sentiment classification for multiple-topic related words (Malinský and Jelínek, 2017). The first method, Domain Elimination, enhances lexicon-based analysis by combining sentiment and domain lexicons. The second method, Cosine Sense-Similarity, exploits lexicon gloss definition to calculate cosine similarity for a more accurate sentiment sense recognition. The experiments in Section 6.4 (Sentiment Sense Disambiguation) demonstrate that the both methods are recommended for sentiment analysis tasks.

4.5.1 Domain Elimination

As mentioned in Section 3.6.1.1 (SentiWordNet Lexicon), one word may have several senses, and each sense may express different sentiment strength; and even a polarity might be different for two senses of one word. For instance, the word "flush" listed in Table 4.2 has a negative SentiWordNet score for entry *flush#a#2* and conversely *flush#n#2* has a positive score. Each sense has assigned a domain that specifies the meaning of the word, and it allows us to determine which sense should be used in the context.

The list of domains allocated to English words is extracted from the MultiWordNet dictionary. MultiWordNet (Pianta et al., 2002) is an extension of Princeton WordNet (Fellbaum, 2010), which assigns a semantic field to each synset, i.e., a domain that most closely represents the meaning of a word. Both lexicons, SentiWordNet and MultiWordNet, arose from Princeton WordNet, so they have defined same or very similar gloss for all entries. This feature made it possible to unify both dictionaries into one and add a domain property to each entry of SentiWordNet lexicon (the last column in Table 4.2). MultiWordNet does not cover all of the WordNet words, nor all the SentiWordNet words. Therefore some of the words have no domain. Conversely, some of the words have multiple meanings. They have assigned multiple domains (e.g. entry *flush#n#2* from Table 4.2 has assigned two domains, health, and psychology).

In every sentence, which has assigned a domain from the same set of MultiWordNet domains, can be significantly eliminated the number of meanings for a given word.

TABLE 4.2: Selected SentiWordNet entries for the word "flush" enhanced by the MultiWordNet domains. Each entry represents one of the word senses with a different sentiment. Pos, Neg and Obj represent the sentiment strength. A synset is a set of synonyms of the word. A domain specifies the meaning of the word.

Word#pos#sense	Pos	Neg	Obg	Synsets	Domain
flush#a#1	0	0	1	-	quality
flush#a#2	0	0.250	0.750	wealthy, moneyed, loaded, affluent	economy
flush#n#1	0	0.125	0.875	prime, peak, heyday, flower, efflorescence, blossom, bloom	industry
flush#n#2	0.625	0	0.375	rosiness, blush, bloom	health, physiology
flush#r#1	0	0	1	-	factotum
flush#r#2	0	0	1	-	factotum
flush#v#1	0	0.125	0.875	redden, crimson, blush	psychological_features
flush#v#2	0	0	1	-	factotum

A domain of an analysed text can be determined in several ways, manually, semi-automatically and automatically. Research text domain might be determined from article keywords. A domain of social network comments, e.g. network for hotels or restaurants, might be adjusted according to the specific shape of the network. A domain identification is more complicated for general social networks and blogs, and therefore a deeper text analysis needs to be applied to them.

4.5.2 Cosine Sense-Similarity

Cosine similarity represents a mathematical measure of similarity of two vectors, $\vec{a}$ and $\vec{b}$, which is obtained by calculating the cosine of an angle of these vectors (Equation 4.15). Two vectors with the same orientation have a cosine similarity equal to 1, two vectors at 90° have a similarity of 0, and two vectors with opposite orientation have -1. This property is used to determine a similarity of two documents, where vectors represent for instance frequency of individual words.

$$sim(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{\| \vec{a} \| \| \vec{b} \|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} b_i^2}} \tag{4.15}$$

We use cosine similarity to determine which sense of a word is used in the context of an analysed sentence. The sense is determined by comparing the gloss of each sense against the analysed sentence. However, sentences are not only compared by the frequency of words, but also by the sentiment score of those words which bear a sentiment. The analysed sentence and the gloss sentence of each sense are converted into vectors before computing the cosine similarity. The individual vector components are represented by a tuple of the unique term and its value. The value is a measure of the importance of the term in the sentence, and it bears information to determine the similarity.

Unique vector terms are determined from an examined sentence. Each sentence is partitioned into words that are searchable in SentiWordNet; i.e. if a word is not in the lexicon then it is lemmatised and checked again; the word bears no sentiment if it is not found on the second attempt and as such it is not used in the vector. If the word is found in the lexicon, then it is added to the vector along with its synsets, which bear the same sentiment. Thanks to the use of the synsets, the texts with a variety of synonyms can also be compared.

The value of each term indicates how often a term appears in a text and how important it is. Term frequency (Equation 4.16) is normalised by dividing by the total number of words (vector length) of the sentence to avoid overestimation of long sentences in which the term may occur more frequently than in shorter ones. The numerator o_{ij} expresses the number of occurrences of the word w_i in the sentence s_j. The denominator o_{kj} represents the sum of the number of occurrences of all words in the sentence s_j. Inverse document frequency (Equation 4.17) represents the importance of each term; the more often a word is less important. The numerator $|S|$ expresses the number of all sentences. The denominator represents the number of sentences where the term t_i appears. The entire formula is evaluated as zero if the term is not found in any sentence to avoid a division by zero. The importance of a term is given by its sentiment score, which is calculated according to Equation 4.18. Term score ts_i is the difference of positive $score_{pos}$ and negative $score_{neg}$ components of the sentiment of the term.

The overall value of each term tv_i is then obtained by multiplying the term frequency with the inverse document frequency and with the term score (Equation 4.19). The value may range from -1 to 1. A term with the value close to 1 or -1 bears a strong sentiment, and it is frequently used. A term with the value close to 0 bears a weak sentiment, and it is not so often used.

$$tf_{ij} = \frac{o_{ij}}{\sum_k o_{kj}} \tag{4.16}$$

$$idf_{ij} = \log \frac{\mid S \mid}{\mid \{j : t_i \in s_j\} \mid} \tag{4.17}$$

$$ts_i = score_{pos} - socre_{neg} \tag{4.18}$$

$$tv_i = tf_{ij} \cdot idf_{ij} \cdot ts_i = \frac{o_{ij}}{\sum_k o_{kj}} \cdot \log \frac{\mid S \mid}{\mid \{j : t_i \in s_j\} \mid} \cdot (score_{pos} - socre_{neg}) \tag{4.19}$$

Chapter 5

Framework for the Analysis of Internet Trends

Based on the research, a novel framework that brings together well-known webometric techniques has been being developed for the analysis and evaluation of Internet trends. The proposed framework provides an end-to-end approach to the analysis of selected Internet trends. Visualiser, a graphical user interface, provides a complete system configuration along with a trend definition, analysis adjustment and visualisation of the analysed results to the user. The framework is vertically scalable to extend the analytical modules by the new one algorithms and auxiliary tools.

5.1 Framework Architecture

The framework architecture (Figure 5.1) is divided into three main interconnected layers, where each of them has its functionality and does not affect the others (Malinský and Jelínek, 2015a). The visualiser is an extension of the framework that provides a graphical interface for a complete system configuration along with a trend definition, analysis adjustment and visualisation of the analysed results to the user.

Crawler Layer

The units at this layer are designed to collect data from the Web and prepare them for subsequent analysis. A web crawler is an automated unit that follows links on the website and stores a key content of all the visited pages. The crawler intelligently selects links to follow based on a user defined keyword that occurs in the link, or a

depth of hyperlinks that defines how far the crawler can move away from the original link, or until there are no more links on the website. Each page is further analysed at the HTML tag level, and its content is divided into several categories such as main content, header, footer, metadata, navigation, advertisement, hyperlinks. All the processed content is continually being stored in a database during the HTML tag analysis.

Analysis Layer

The analysis layer is fully configurable from the visualiser. Trend Detection unit is used to find a text that is relevant to the user defined trend. The trend can be even defined as a multi-word expression with a Boolean term to increase the precision. For instance, the trend "Ottawa earthquake" is better to restructure to the expression "ottawa AND earthquake". However, names and specific expressions, e.g. "bill gates", should stay in the same form and then be searched as the exact phrase.

The trend analysis is performed by user interpretation in Visualiser. That can be done by assembling a graph where each node represents an analytical algorithm. The trend analysis is subsequently performed in the order in which the graph passes through. All the processed results are continuously being stored in a database.

Storage Layer

The layer is designed to quickly store and retrieve data from the database in order to be the trend evaluation displayed in real-time to the user. For this purpose, the implemented bespoke citation indexes are defined to optimise speed and performance for communication of the individual layers with the database. Among other things, crawled pages are stored at this layer; once the website is stored in a database, then its content is for further processing called a document. Also, parts of that websites divided into categories by Preprocessor and also the list of sentences for each trend created at Analysis Layer, all of these are stored using the storage layer components.

The framework communicates with a database on a higher level using object-relational mapping mediated by Java Persistence API (Keith and Schincariol, 2013). One of the main advantages of Java Persistence API is the ability to change the persistent tool without affecting the functionality of the application. Hibernate framework (Konda, 2014) is default persistence tool in the proposed framework. However, it can be replaced by any other tool thanks to the JPA specification.

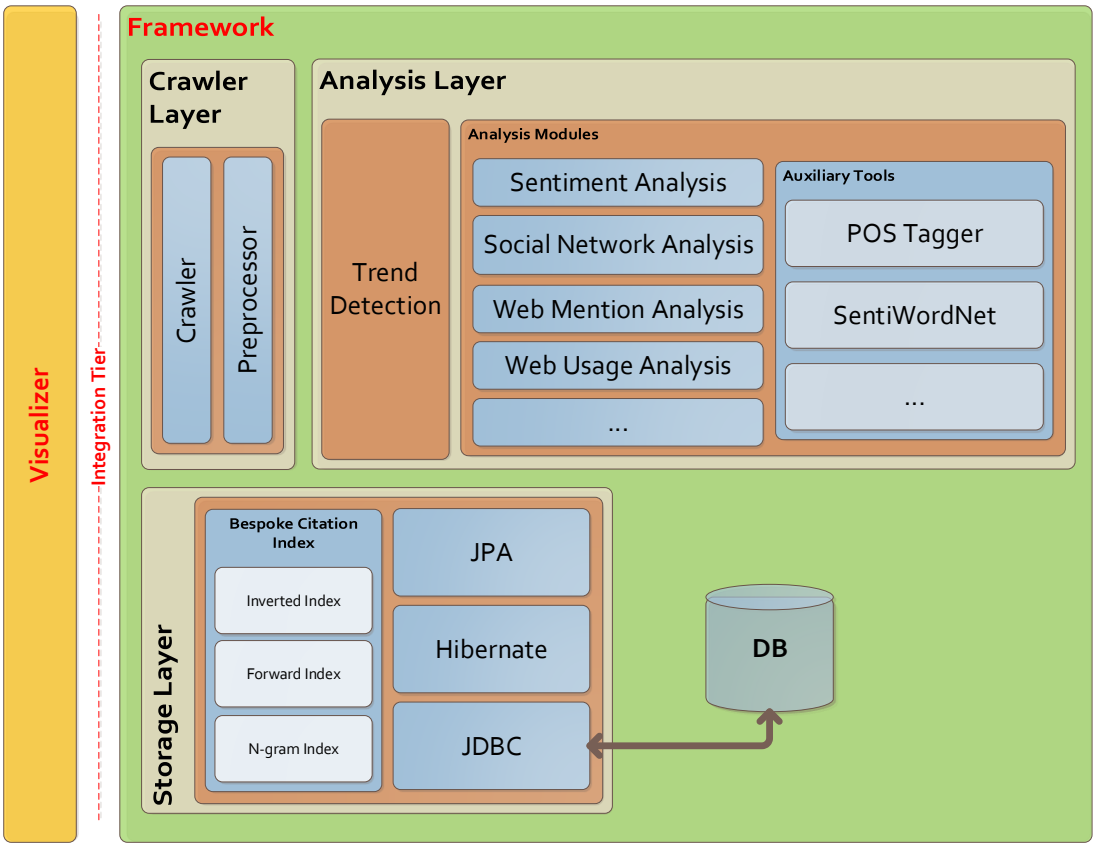


FIGURE 5.1: Architecture of the framework for the analysis of Internet trends.

Bespoke Citation Index is necessary to optimise speed and performance in finding a user selected trend relevant documents in the database. Inverted and Forward Index store a list of words for each document along with a list of references to each word and its position within the document. N-gram Index stores a list of the occurrences and frequency of all n-consecutive words, where n is typically 1, 2, or 3 (e.g.: "like-it-!").

5.2 Real-Time Trend Visualiser

Visualiser is the main part of the framework that provides a graphical user interface for a complete system configuration along with a trend definition, analysis adjustment and visualisation of the analysed results to the user (Malinský and Jelínek, 2015b).

The user can use the interface to track the progress in the analysis of all trends in the system, including those that have been created by other users. The configuration of the analysis of a trend requires defining the data sources and data collection frequency, as well as the choice of the methods for processing of collected data by the selected analytical modules.

Data Source Definition

The user defines URI and method of web data mining when creating a new data source. The data mining method is determined by the period that defines the frequency of crawling and by the level of depth of hyperlinks that defines how far the crawler can move away from the original URI while browsing hyperlinks. The entire content of each of the visited pages is stored as it was loaded in a database; i.e. plain text in HTML format.

Web Page Preprocessing

The content of each stored web page is analysed at the HTML tag level, and then it is divided into several categories: main content, header, footer, etc. Another category can be created, and a user can define rules for its recognition on a web page using an appropriate HTML tag or tag's identifiers.

Trend Analysis

Several analytical plans can be defined for each trend in the system. Every plan has defined a data source and categories, where the category describes which part of the source will be used for analysis. It is also possible to define a frequency of analysis; this is especially important if the content of the website itself is being changed. The last important part of the analysis plan is the selection of analytical modules, their configuration and assembling a graph where each node represents one module. Trend analysis is then performed by sequential pass of the assembled graph where the processed data is transferred from one node/module to the other node.

An example of the assembled Visualiser graph is illustrated in Figure 5.2. The whole graph is interactive, and it can be modified through several mouse clicks. After a double-clicking on a node in the graph, the user displays detailed configuration settings and options for the node. Furthermore, the user is able to add additional nodes and define edges between them.

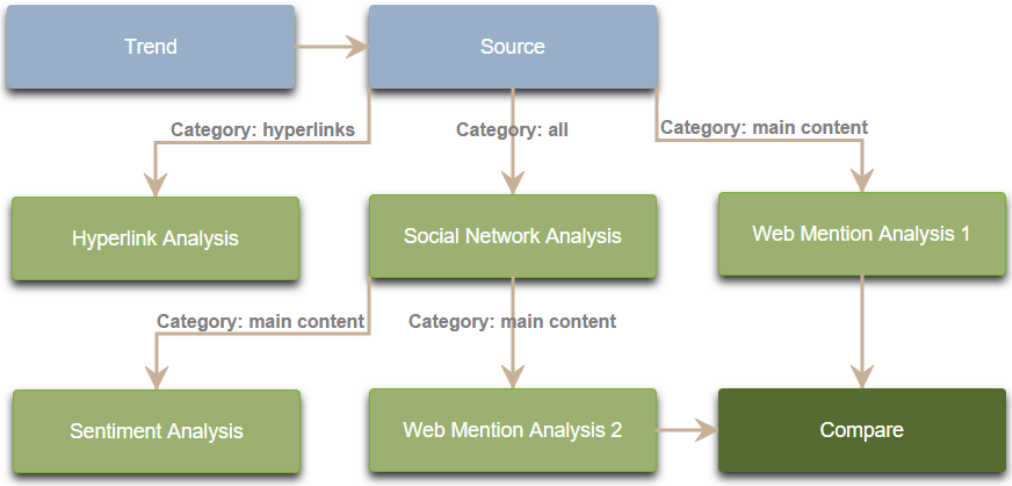


FIGURE 5.2: An example of the assembled graph in the Visualiser. Trend analysis is performed by passing through the graph.

The blue coloured nodes represent a selection. In this case, they determine which trend is analysed and which web pages from a data source are used for analysis. The output of each node may have a defined category, which constitutes a part of the website that is passed to the analysis. All the outputs of the node are transferred to further analysis if there is no category selected. For instance, the edge between node *Source* and *Hyperlink Analysis* is marked by category "hyperlinks"; i.e. only the hyperlinks gathered from analysed web pages are passed to the node Hyperlink Analysis. The user also has the option to specify which hyperlinks (from navigation, advertisement, main content, all, etc.) are used for the analysis; this option is available in the detailed settings of the node.

The light green coloured nodes represent the analytical modules. All the outputs of the analytical modules are automatically displayed to the user in the form of tables and charts. As illustrated, the output of the individual analytical module can be used as a data input for analysis in other modules. For example, the *Social Network Analysis* module is configured to use Degree Centrality to determine the most read web pages, and the entire output is then passed to *Sentiment Analysis* and *Web Mention Analysis 2* modules for further analysis.

The dark green coloured nodes are used to compare the outputs of two identical analytical modules. The output of the comparison is also displayed to the user in the form of table and chart showing both outputs.

The outputs of the individual modules and also complete results are gradually displayed to the user in chronological order in tables and charts. The update of the results to the user is almost instantaneous and independent of the user; data is displayed to the user as soon it is processed or analysed. That is caused because of the system architecture and the communication strategy with client web browser.

5.2.1 Visualiser Architecture

The visualiser architecture (Figure 5.3) is based on Java Server Faces (JSF) (Wadia et al., 2014) and supported by Primefaces (PrimeTek, 2015) component library. The components utilise JavaScript and AJAX to provide a rich user experience with support for real-time content updates. The real-time content updates are critical for automatic updates of results that are displayed to the user immediately after the data analysis.

The sequence of events resulting in a page update is as:

1. One portion of the data analysis is completed, and the results are stored in a database. The trigger, which is part of the business logic is activated, and the previously stored data are loaded from a database into the managed bean that manages a displayed web page with results.

2. JSF runtime re-renders the entire component tree stored on the visualiser server-side.

3. The component tree differences are calculated, and page update is packed into the XMLHttpRequest object via the AJAX Controller. The XMLHttpRequest object then calls the callback function on a client-side.

4. The XMLHttpRequest callback function updates a web page Document Object Model (DOM) and thereby automatically updates the web page with the new data.

5. The same process is invoked when the user changes any data in the result table and thereby the result chart is automatically updated.

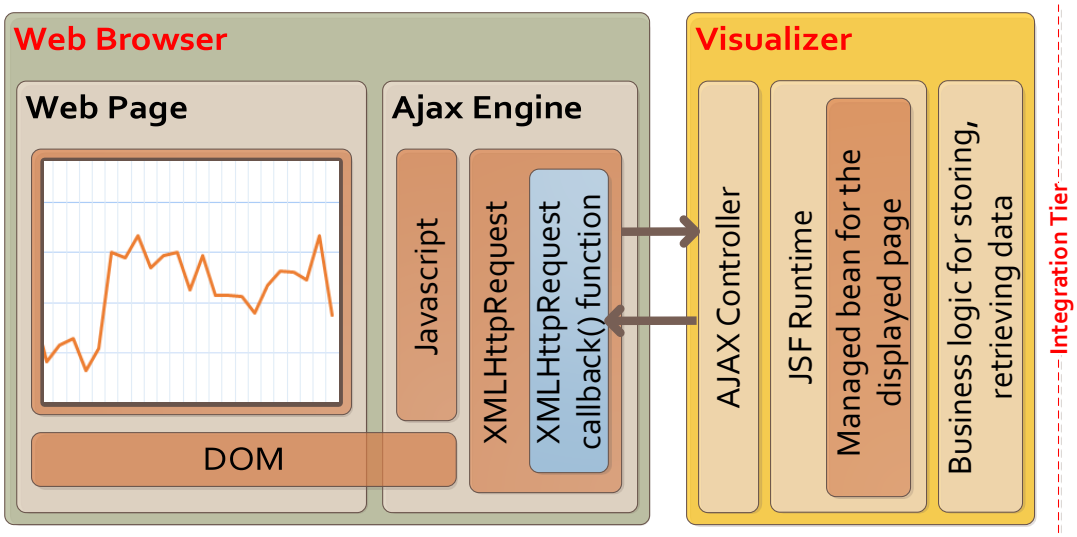


FIGURE 5.3: The visualiser architecture and communication with a client web browser.

5.3 Conclusion

The complex web application for the end-to-end evaluation of selected Internet trends has been proposed. The application consists of two main parts: framework and visualiser. The framework combines the tools for collection and processing data from the Web, the analytical tools that provide algorithms for analysis of collected data, and the data tools designed to quickly store and retrieve data from a database. The visualiser provides a graphical user interface for a complete system configuration along with a trend definition, analysis adjustment, and visualisation of analysed results to the user. The system architecture is vertically scalable which allows the addition of new custom analytical modules.

Chapter 6

Experiments and Main Results

6.1 Chronological Evaluation of Internet Trends

As mentioned in Section 4.1 (A General Model for Gathering and Processing Data from Web 2.0), a novel general model has been proposed for gathering and processing data from Web 2.0. Based on the model, a new methodology has been designed for gathering and processing data from blogs (see Section 4.2, Evaluation System for Gathering and Processing Data). The experimental system for gathering and processing data from blogs has been created and implemented to verify our theoretical assumptions.

The data input of the popular trends for the evaluation has been obtained from the Google Trends web service. The blogs intended for analysis have been obtained from the BlogPulse service. BlogPulse[1] was an automated trend discovery system for blogs, which reflected what people were posting on the Internet. BlogPulse collected data from blogs, created a full-text search index and provided a chronological summary of daily volume of blog post matching a trend. The service indexed over 160 million blogs, and it increased approximately 60,000 blogs every day when it was publicly available.

6.1.1 Methodology of Study

According to the designed methodology (see Section 4.2, Evaluation System for Gathering and Processing Data) the analysis has been performed in four phases:

[1] http://www.blogpulse.com

TABLE 6.1: Conversion table between Penn Part-of-Speech Tags and SentiWordNet Part-of-Speech classes. A complete list of tags and their description is given in Appendix A.

SentiWordNet Part-of-Speech class	Penn Part-of-Speech Tags
Adjective	JJ, JJR, JJS
Noun	NN, NNS, NNP, NNPS
Adverb	CC, RB, RBR, RBS
Verb	VB, VBD, VBG, VBN, VBP, VBZ

In the first phase, **Hot Searches**, Google Trends has provided 48 hot searches produced from March 16, 2011, to March 27, 2011; there were four the most deviate trends for every day. 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 have been generated to match relevant post. For example, the expression "ottawa earthquake" is better to restructure to "ottawa AND earthquake", however, names and specific expressions, e.g. "bruce pearl" stay in the same form and they are searched as the exact phrase. Each expression has been searched in the BlogPulse search engine on the interval between 30 days before and 30 days after the trend had been detected. The BlogPulse has created a chronological summary of daily volume of blogs for each trend; 88,711 blogs and 54,381 unique blogs in total.

In the third phase, **Part-of-Speech Tagging**, the surroundings of each searched expression from the chronological summary had been recognised, and a list of sentences for the trend has been created; 128,469 sentences in total. Every sentence in the list is tagged by using Stanford POS Tagger (Toutanova et al., 2003). The tagger assigns the Penn part of speech tag (Marcus et al., 1993) to each word in a sentence, and it even predicts the part-of-speech for an unknown word. For the processing in the fourth phase, the tagged words have been divided into the four part-of-speech categories: adjective, noun, adverb, verb (Table 6.1). The plural words have been converted into singular for a more accurate recognition of the words in the fourth phase.

The last phase; **Sentiment Analysis**, determined the polarity of the tagged word and evaluated sentences for each day for each trend; 1,505,869 sentimental words in total. The evaluation has been performed by the lexicon-based methods using SentiWordNet as a lexicon of words (Baccianella et al., 2010; Esuli and Sebastiani, 2006). SentiWordNet assigns to each synset of WordNet three sentiment scores:

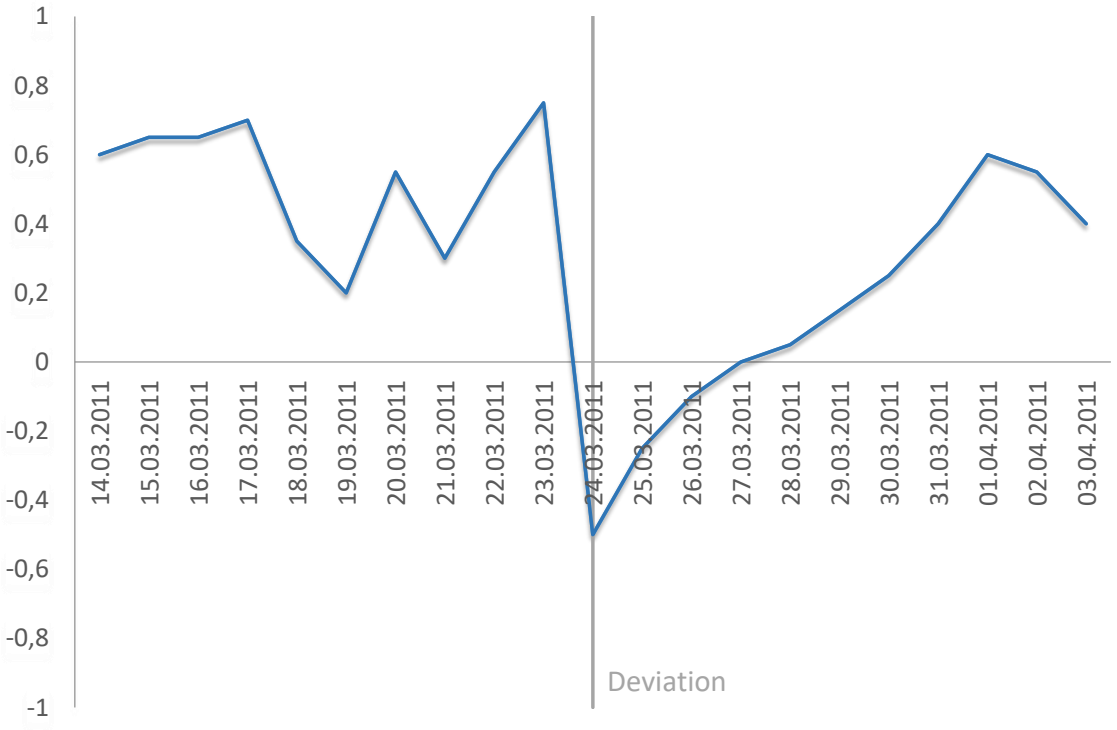


FIGURE 6.1: Evaluated chronological summary of the trend "myanmar" in 10 days around its deviation. The trend deviated on the 24th of March when there was a strong earthquake in Myanmar.

positivity, negativity, objectivity. The evaluation of trends has been performed according to the described rules (see Section 4.2, Evaluation System for Gathering and Processing Data).

6.1.2 Results

One of the outputs of the proposed experimental system is showed in Figure 6.1. In the figure is a chart, which represents evaluated chronological summary of the trend "myanmar" in 10 days around its deviation. Myanmar, also known as the Burma, officially the Republic of the Union of Myanmar, is a country in Southeast Asia. The x-axis of the chart represents the published date, and the y-axis shows the polarity of the trend. Positive values of the y-axis represent the positive evaluation of the trend. Negative values of the y-axis represent the 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 shown, people were writing relatively positive about the 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 assessment of the trend was gradually coming back to the positive values in the following days.

6.1.3 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 a specific trend.

6.2 Comparing Methods of Trend Assessment

This case study deals with a comparison of selected webometric methods for the evaluation of Internet trends (Malinský and Jelínek, 2014). Web Mention Analysis, Sentiment Analysis and Social Network Analysis are among 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. 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. The methods have been primarily chosen for their diversity and applicability in various areas of the Web and social engineering.

6.2.1 Methodology of Study

The methods selected for the evaluation of trends have been compared over the data from the film industry. User reviews published in 2012-2013 on IMDb[2] serves as the source for this research. Five the best-rated and five the average-rated movies

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

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 6.2, 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 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 4.2, Evaluation System for Gathering and Processing Data). All sentences have been processed using the Lexicon-Based method with SentiWordNet as the lexicon of words. 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 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. 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 last column "Rank" is designed to determine the impact of a particular methodology on the overall evaluation. The column reports the overall trend ranking based on the sum of previous evaluation (SA)+(WMA)+(SNA); the result of the sum is given in parentheses. Multiplication is used instead of the sum in case the result of the sum is same for more movies. The multiplication is not as default operation to ensure that the individual methodology results are equivalent, and an extreme evaluation of the one methodology will not strongly affect the others. The number before the brackets indicates the trend assessment where the lower value is the better rating. Comparison of the first and last columns may give an idea about the differences in the movie evaluation between a typical user and film fan. A typical

user usually simply evaluates a movie by selecting the numerical value of 1 to 10. A real film fan generally provides a text review in addition to the numerical evaluation.

6.2.2 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. Web Mention Analysis emphasises the frequency of making reviews and reports the overall number of reviews in a given period. Social Network Analysis determines the prestige of the authors and thus defines the quality of the source.

The result (Table 6.2) shows that the best rated IMDb movie *The Dark Knight Rises* is also the best rated by the Rank. 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 second best-rated Rank 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 many reviews about the movie (second highest WMA), and the reviews are very positive; polarity is the highest of all rated movies. 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 positioned on the third place of the Rank. The movie is the best in both IMDb and Rank evaluations for 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, the five best-rated movies on IMDb side is also among the top rated movies on the Rank side. There is just one exception; *The Amazing Spider-Man*, which has fallen among the high-rated movies. 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

TABLE 6.2: Comparing Methods of Trend Assessment.

Movie	IMDb	SA	WMA	SNA	Rank
Django Unchained	2 (8.4)	5 (0.0589)	5 (952)	6 (88)	5 (16)
Life of Pi	4 (8.0)	3 (0.0620)	7 (665)	7 (87)	6 (17)
The Avengers	3 (8.1)	1 (0.0821)	2 (1488)	3 (101)	2 (6)
The Dark Knight Rises	1 (8.4)	4 (0.0614)	1 (2491)	1 (114)	1 (6)
The Hobbit: An Unexpected Journey	5 (8.0)	6 (0.0569)	3 (1243)	4 (92)	4 (13)
Battleship	10 (5.9)	10 (0.0346)	6 (674)	10 (67)	9 (26)
Dark Shadows	8 (6.3)	8 (0.0523)	10 (440)	9 (74)	10 (27)
Snow White and the Huntsman	9 (6.2)	9 (0.0500)	8 (650)	5 (90)	7 (22)
The Amazing Spider-Man	6 (7.1)	2 (0.0819)	4 (1092)	2 (103)	3 (8)
Total Recall	7 (6.3)	7 (0.0547)	9 (465)	8 (77)	8 (24)

influence of on the first two top rated movies. Those movies would be moved to a lower position without the WMA, and *The Dark Knight Rises* would be the first. Web Mention Analysis has also impact on the distance of individual Rank evaluations. There would be a very similar evaluation for instance for *Life of Pi* and *The Hobbit: An Unexpected Journey*, and it would be more difficult to establish a position in the ranking. 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.

6.2.3 Conclusion

There have been selected three webometric methods, which are often used as supportive search engines assessment algorithms. Each of the chosen methods was used to analyse 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 - 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. In our case, the ranking defined by the all three methods in comparison with a 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. That means to define criteria for quality assessment of found information, and "distance" among each trend. On this basis, rules for evaluation of semantic content concerning user's queries can be designed.

6.3 Movie Evaluation in the Network of Trends

A new methodology has been proposed for the evaluation in the network of trends (Malinský and Jelínek, 2016a,b). The method is based on the Social Network Analysis and enhanced by the Sentiment Analysis (see Section 4.4, Trend Evaluation in the Social Network Sphere). 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 section uses a newly proposed methodology to determine the most prestigious authors who have written reviews about the trend. In this study, 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.

6.3.1 Methodology of Study

Like the first, 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 6.3 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.

6.3.2 Results

Tables 6.4 and 6.5 report the result of the evaluation of movies based on the corpus of reviews. 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. So 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 6.4 shows the results according to the individual threshold. The value of the threshold refers to a degree power in Table 6.3; i.e. for instance, the column "Threshold 5" refers to the degree power that is equal to five or higher. 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 upper 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

TABLE 6.3: 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

TABLE 6.4: Evaluation results of the selected movies[3] 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 6.5: Evaluation results of the selected movies[3] 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)

[3] 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.

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 6.5 shows the evaluation results of only those reviews, which 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; so 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.

An opposite example can be seen in the evaluation of movies 3 - *The Avengers* and 9 - *The Amazing Spider-Man* which are 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 6.2, where these films are placed in the top three.

6.3.3 Conclusion

The newly proposed approach for the evaluation in the network of trends has been examined over the data from the film industry. 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 upper 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.

6.4 Sentiment Sense Disambiguation

The two new methods have been proposed (Malinský and Jelínek, 2017) to improve sentiment classification for multiple-topic related words (see Section 4.5, Sentiment Sense Disambiguation). The first method, Domain Elimination, enhances lexicon-based analysis by combining sentiment and domain lexicons. The second method, Cosine Sense-Similarity, exploits lexicon gloss definition to calculate cosine similarity for a more accurate sentiment sense recognition.

6.4.1 Methodology of study

The proposed methods have been compared using the data from traveller sphere. TripAdvisor[4], a travel website company providing reviews of travel-related content has served as a data source for this research. Data was initially gathered for latent aspect rating analysis (Wang et al., 2010, 2011), and the authors have publicly provided data[5] for another research. TripAdvisor data set contains text reviews and a numerical rating for over 12,000 venues. Every venue has assigned an average of 500 comments along with overall venue evaluation. The evaluation ranges from 1 to 5 stars, where one is the worst and five the best. There have been selected 1000 reviews for each numerical evaluation. All reviews had been analysed using the described strategies for the word sense disambiguation, and the results were compared to the assessment based on the proposed methods. At first, all reviews were split into sentences, and then each sentence has been evaluated in several steps:

In the first step, **Part-Of-Speech Tagging**, the Stanford Log-Linear Part-Of-Speech Tagger (Toutanova et al., 2003) is used to assign part of speech to each word in an analysed sentence, and it even predicts the part of speech for an unknown word. From the tagged words are for subsequent processing selected those which fall within the four categories according to Table 6.1: adjective, noun, adverb, and verb. That is necessary because the SentiWordNet recognises only these four types, while Stanford Part-Of-Speech Tagger works with many part-of-speech classes that are identified by Penn Treebank tag set. Plural words are converted to a singular for more accurate recognition.

In the second step, **SentiWordNet Lookup**, the word along with its part of speech is looked up in SentiWordNet lexicon as word#pos. There are three cases for this search:

1. No sense is found for the word – the word is in the form that does not bear any sentiment. Get the canonical form of a word (i.e., lemma) and look up again. If a sense is not found again, then the word is not used for any further evaluation. This exclusion may happen for example for preposition, definite and indefinite articles, but also for the name of country, city or names of persons.

2. Only one sense is found for the word – the word has defined just one sense in SentiWordNet, and thus the other sense disambiguation steps are skipped.

[4]http://www.tripadvisor.com
[5]http://times.cs.uiuc.edu/ wang296/Data/

The word is not lemmatised before the first lookup to resolve the right sense. For instance, "addicted" has lemma "addict" and both words are listed in SentiWordNet. The use of lemmatization before lookup will always skip original word meaning.

3. Multiple senses are found for the word – there is not enough information to recognise correct word meaning. The word sense will be recognised in the next step.

In the third step, **Domain Elimination**, domains of all senses acquired at the previous step are compared against the set of travel-related domains. The following MultiWordNet domains have been selected as travel-related: food, gastronomy, tourism, transport, aviation, vehicles, nautical, railway. If there are some senses with an unrelated domain, then these are excluded from further evaluation. Only the senses with the undefined or travel-related domain are used in the next step. If there is just one sense with the undefined or travel-related domain, then the sense bears the word sentiment, and thus the other sense disambiguation may be skipped.

In the fourth step, **Cosine Sense-Similarity**, cosine similarity algorithm is used to compare the sentence with glosses of multiple senses which were obtained at the previous step. Glosses of synsets of obtained senses are also included in the comparison. The sense with the most similar gloss description is used for further evaluation. If there are more senses with same similarity evaluation, then the word sentiment is calculated as the sum of harmonic series of all recognised senses.

In the fifth step, **Review Evaluation**, the sentence evaluation is being performed as a division of the difference between positive and negative scores to the number of positive and negative words. The overall result of each review is then calculated as the arithmetic average of the evaluation of all sentences for the review.

In the sixth step, **Accuracy Measure**, the comparison of results for each review with the assessment from those who wrote the review is based on the F-measure that considers the precision and the recall to compute the score. The precision (Equation 6.1) is defined as the number of successfully evaluated reviews *ser* for a given rating divided by the number of all reviews that have been evaluated for a given rating, i.e., the number of successfully evaluated reviews *ser* plus the number of incorrectly evaluated reviews *ier*. For example, in the case where 657 reviews have been assessed for the rating 4, but only 181 of them are correct, the precision is defined as $181/657 = 0.275$. The recall (Equation 6.2) is defined as the number of

successfully evaluated reviews for a given rating *ser* divided by the total number of reviews that have been reserved for a given rating, i.e., the number of successfully evaluated reviews *ser* plus the number of unsuccessfully evaluated reviews *uer*. For example, in the case where 1000 reviews have been reserved for rating 4 and 181 of them are correctly evaluated for the same rating, the recall is defined as 181/1000 = 0.181. The F-measure (Equation 6.3) represents a harmonic mean of the precision and the recall; it is roughly average for close values, and it inclines to the lower for distant values.

$$precision = \frac{ser}{ser + ier} \tag{6.1}$$

$$recall = \frac{ser}{ser + uer} \tag{6.2}$$

$$F\text{-}measure = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{6.3}$$

6.4.2 Results

As mentioned above, there have been selected 1,000 reviews for each numerical evaluation; i.e. 5,000 reviews for the evaluation from 1 to 5. Selected reviews consist of variety length; the shortest review has nine words and the longest consists of 174 words. All reviews have been split into sentences; each review has on average seven sentences, and the whole corpus contains 34,632 sentences in total. Each sentence consists of 1 to 22 words; 567,225 words in total. One of the shortest words is "a", the longest word is "first-come-first-serve". There have been identified 287,379 words that bear any sentiment; 71,052 positive; 31,264 negative; 185,063 neutrals.

TABLE 6.6: Mapping schema between sentiment score distribution and derived rating that correspond to the TripAdvisor numerical rating.

Sentiment Score	Derived Rating
[-1; -0.05]	1
(-0.05; -0.01]	2
(-0.01; 0.01]	3
(0.01; 0.1]	4
(0.1; 1]	5

TABLE 6.7: Results of the two new methods applied to the corpus of TripAdvisor reviews and their relation to various modifications. The highest values for each rating are in bold (P – Precision, R – Recall, F – F-measure).

	no-dom + no-cos			**dom + no-cos**			**no-dom + cos**			**rel-dom + cos**			**rel-dom + tf-idf-cos**			**all-dom + tf-idf-cos**		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
1	**0.460**	0.333	0.386	0.458	0.347	0.395	0.425	0.471	0.447	0.405	**0.591**	**0.480**	0.391	0.573	0.465	0.405	0.493	0.445
2	0.299	0.244	0.269	**0.306**	0.232	0.264	0.283	0.236	0.258	0.266	0.211	0.236	0.289	**0.263**	**0.275**	0.273	0.245	0.258
3	0.240	0.199	0.217	0.256	**0.213**	0.232	0.252	0.186	0.214	0.254	0.155	0.193	0.263	0.197	0.225	**0.269**	0.205	**0.233**
4	0.224	0.197	0.210	0.235	**0.204**	0.218	0.252	0.173	0.205	**0.288**	0.169	0.213	0.275	0.181	0.218	0.282	0.196	**0.231**
5	0.381	0.666	0.484	0.377	**0.671**	0.483	0.391	0.639	0.485	0.409	0.634	**0.497**	**0.426**	0.518	0.468	0.422	0.602	0.496
	0.321	0.328	0.313	0.326	0.333	0.318	0.321	0.341	0.322	0.324	**0.352**	0.324	0.329	0.346	0.330	**0.330**	0.348	**0.333**

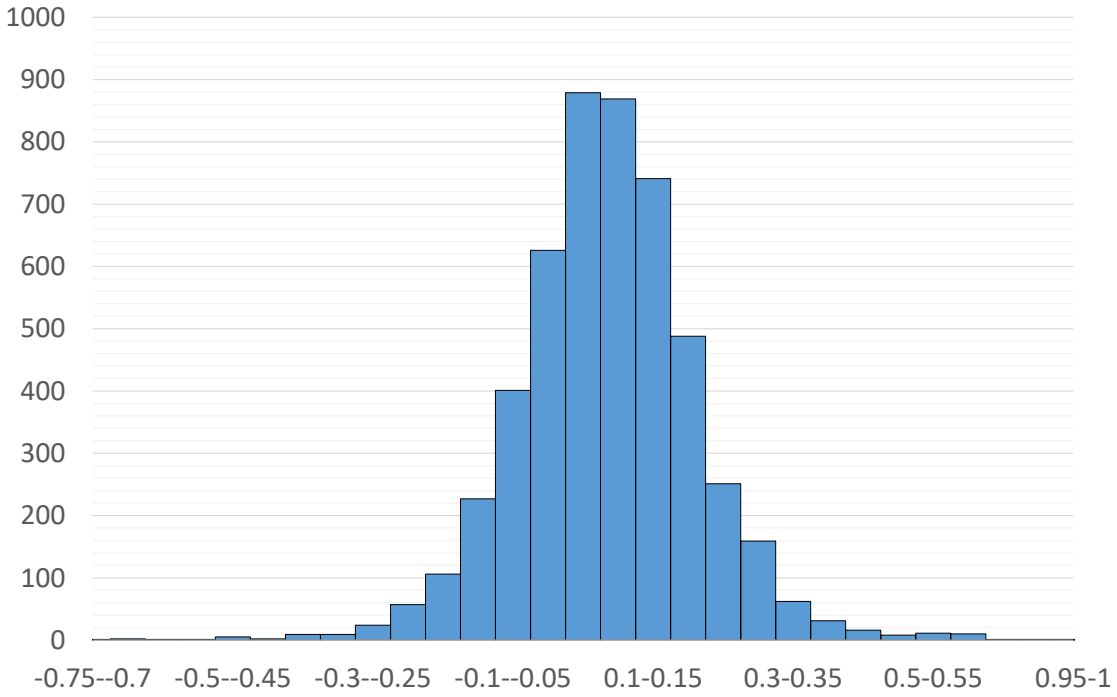


FIGURE 6.2: Sentiment score distribution for all reviews.

As shown in Figure 6.2, the distribution of sentiment scores is highly clustered around zero. This distribution implies that there are more neutral than positive or negative words in the review corpus. Furthermore, a slight increase towards positive scores is also observed, which implies people tend to leave more positive reviews. Considering such a distribution of sentiment scores, the rating has been derived for sentiment scores that correspond to the TripAdvisor numerical rating (Table 6.6); i.e. from 1 to 5, where one is the worst and five the best. The derived rating serves as a measure to compare the evaluation of the new methods to the original evaluation from people on TripAdvisor.

Tables 6.7 and 6.8 report the results of two new methods applied to the corpus of reviews. The first table shows how the review evaluation is changed depending on the addition of new methods. The second table compares the evaluation of the new methods with existing sense disambiguation strategies. In both cases, the number of correctly and incorrectly rated and unrated reviews are taken into account to obtain differences among the strategies. The resulting comparison is based on F-measure that considers the precision and the recall to compute the score.

TABLE 6.8: Comparison of the results of the new methods with existing sense disambiguation strategies. The highest values for each rating are in bold (P – Precision, R – Recall, F – F-measure).

	first sense			**all pos/neg**			**selected pos/neg**			**geom series**			**harm series**			**all-dom + tf-idf-cos**		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
1	0.460	0.333	0.386	**0.551**	0.250	0.344	0.537	0.245	0.337	0.448	0.211	0.287	0.387	0.275	0.322	0.405	**0.493**	**0.445**
2	0.299	0.244	0.269	0.324	0.249	0.282	**0.335**	0.269	0.296	0.296	**0.413**	**0.345**	0.307	0.328	0.317	0.273	0.245	0.258
3	0.240	0.199	0.217	0.246	0.243	0.244	0.250	0.272	0.260	0.246	0.374	0.297	0.268	**0.435**	**0.332**	**0.269**	0.205	0.233
4	0.224	0.197	0.210	0.237	0.256	0.246	0.232	0.249	0.240	**0.342**	0.317	0.329	0.341	**0.329**	**0.335**	0.282	0.196	0.231
5	0.381	**0.666**	0.484	0.380	0.650	0.480	0.384	0.605	0.470	0.416	0.311	0.356	0.396	0.253	0.309	**0.422**	0.602	**0.496**
	0.321	0.328	0.313	0.348	0.330	0.319	0.347	0.328	0.320	**0.349**	0.325	0.323	0.340	0.324	0.324	0.330	**0.348**	**0.333**

The first column reports the state where no new method has been applied, so only the first sense of each word has been used to calculate the score. The second and third columns show how the score changes using the individual strategy. A standalone use of the Domain Elimination method gives a small improvement compared to the first column; the overall accuracy increases from 0.313 to 0.318. This slight change probably occurs because just travel-related domains were chosen for the sense disambiguation. The scores might be more divergent with more variable data and the use of multiple domains. However, the method achieves excellent results in recall values, where it correctly evaluates many reviews for the ratings 3, 4 and 5.

A standalone Cosine Sense-Similarity method has much better results compared to the first sense method; the overall accuracy increases from 0.313 to 0.322. That is mostly due to the excellent recall for the ratings 1 and 5, where it correctly evaluates over 47% and 63% of reviews for these ratings. On the other hand, the method has a weaker recall for the ratings 2, 3, and 4, where it rates the smaller amount of comments than the first sense and the standalone Domain Elimination methods.

The column "rel-dom + cos" reports the results of the combination of both new methods; only travel-related senses were used for cosine evaluation if a single trend had not been identified by Domain Elimination method. It can be observed that the methods are considerably precise in the limit values for ratings 1 and 5. Therefore, it seems that this combination is effective for extreme values, i.e. very negative and positive sentiments. However, the mere combination of both methods is not much advantageous for the average reviews; the recall and the overall score for ratings 2, 3 and 4 are the worst of all the previous approaches in this case.

The last two methods extend the calculation by Term Frequency and Inverse Document Frequency. Both methods have due to the TF-IDF a high precision and accuracy compared to the previous "rel-dom + cos" in the average ratings. It seems that reviews with these ratings contain a bigger number of similar words and this combination of methods helped to favour the most important ones. Moreover, the last method operates with all senses regardless of their domain. However, Domain Elimination method still takes into the account the travel-related sense only. The results showed that the overall precision and score achieve the best results for this combination.

Table 6.8 reports the comparison among sense disambiguation strategies. The last column "all-dom + tf-idf-cos" represents newly introduced approach, which is based on Domain Elimination and Cosine Sense-Similarity. Other columns represent existing

strategies, which are described in Section 3.6.2 (Sentiment Sense Disambiguation): first sense, all positive/negative (Equation 3.12), selected positive/negative (Equation 3.13), geometric series (Equation 3.14), harmonic series (Equation 3.15).

First sense column is identical to the first column in Table 6.7 where no new method has been applied. As data indicates, despite the fact that the first sense is the most frequent, the other strategies provide much more accurate results. As the results of "all pos/neg" and "selected pos/neg" demonstrate, the both strategies are very similar, and they reach almost the same numbers in all the measured aspects. Both strategies also provide better results than the standalone Domain Elimination method, which can be most probably caused by the domain selection for the sense disambiguation. However, Domain Elimination may serve as a complementary method for these strategies to eliminate unrelated senses and improve the overall score. Both strategies lose the precision for the ratings 3, 4, and 5 in comparison to the "rel-dom + cos" and so it seems they are rather effective for the words that bear a positive sentiment.

The best results among existing strategies have been achieved by using the sum of the senses of geometric and harmonic series. These strategies are even more accurate than the individual use of a simplified version of one of the new methods. However, the combination of both new methods, "rel-dom cos", provides much more accurate results in both the precision and the recall.

Thanks to the sum of the series, the overall scores for all ratings almost identical, i.e. there is not much difference in the score, as in the case of other strategies, between the reviews rated as 1 and 5, and the average reviews rated as 2, 3, and 4. It entails the improvements to the average review evaluation, however, also the deterioration for the other ratings as compared to other methods.

The newly introduced approach in the last column brings significantly better results compared to the other strategies. In the comparison to other strategies, the best results have been achieved in the limit values for ratings 1 and 5. However, many reviews have also been evaluated with good precision in the average ratings. The overall accuracy has increased by 6% from 0.313 to 0.333.

6.4.3 Conclusion

Overall, both methods provide a significant increase in accuracy of the sense disambiguation, and thus they are recommended for sentiment analysis tasks. The domain elimination method achieves excellent results in the completeness of the identified

reviews that have been rated by people as average. The completeness and accuracy are slightly lower for extreme values, i.e. very negative and positive reviews. This defect is probably due to the fact that the analysed data are mainly focused on hotel reviews. The accuracy might be higher for more variable data where multiple domains could be used. The cosine sense-similarity method achieves much better results for extreme values, however, on the contrary, it is a slightly inefficient for average rated reviews. Similarly, the simplest combination of both new methods brings a very good result for extreme values but inferior in average values. Improved results for the average reviews are achieved by extension of the evaluation process by Term Frequency and Inverse Document Frequency; hence it seems that reviews with these ratings contain a bigger number of similar words, and this combination helps to favour the most important ones.

The combination of both new methods provides much more accurate results for all kinds of ratings. In comparison with the first sense strategy, the precision has increased by almost 3%, the completeness has grown by nearly 6%, and the overall accuracy has improved by 6%. In comparison with other strategies, new methods achieve excellent accuracy for very positive and very negative reviews. The accuracy along with the completeness are the best for these reviews among all the measured strategies. Somewhat worse results are given for the completeness of average reviews. However, the average reviews are evaluated worse than the very positive and very negative reviews by all the strategies. That could be caused by the chosen mapping scheme between sentiment score distribution and derived rating. However, individual reviews for ratings 1, 2, and 3 can be very similar. Someone can also rate a review as 2, and someone else may evaluate the same review as 4. The evaluation is, therefore, more complicated for these average reviews than for strongly positive or negative reviews. It is evident that it is much easier for people to rank a review as excellent (1) or awful (5) than to define any average rating (2, 3, and 4).

Despite the overall improvements, the standalone domain elimination method gives less than the expected results for reviews rated as 1 and 5. It would be interesting to try the method for more variable data and presumably this would show that it is more efficient when more domains can be identified. It would also be useful to identify which trends have the greatest impact on the evaluation of individual reviews, i.e. what factor is the most important for authors of reviews.

Chapter 7

Conclusions

The Internet is becoming the town square for the global village of tomorrow.
— Bill Gates

7.1 Summary

The Internet has been becoming an important information source since it contains many ideas on various topics from many different types of users. The obtaining comprehensive information from such a source is a challenging task nowadays, which includes the investigation of reciprocal relationships, an analysis of the website content and recognition of its meaning.

This thesis has aimed to find a suitable solution for data analysis in the Web 2.0 environment. The primary emphasis has been placed on the evaluation of 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 theoretical foundations along with their formal description have been determined for the evaluation of Internet trends. The proposed metrics follow up on the progress being already made in the Webometric research and further extends the research by the idea of the sentiment and social network analysis.

Several practical experiments have been designed and implemented to verify the theoretical assumptions about the evaluation of Internet trends. The experiments have been based on the determination of the polarity of an analysed text, and the definition of reciprocal relationships in social networks. All measures have been

performed using the various data input to obtain the results covering the most of the practical use cases. The results confirm all the theoretical assumptions, and thus the proposed web metrics are recommended for the evaluation of Internet trends.

7.2 Contributions of the Dissertation Thesis

The introductory part describes the motivation behind our efforts together with the goals of the thesis. The following chapters present the basics of the World Wide Web, describes its dynamic structure and introduces the methodologies for collecting and analysing data from it. The subsequent survey of the current state-of-the-art reports the necessary theoretical background and presents the important knowledge in the areas of Webometrics, Sentiment Analysis, and Social Network Analysis.

The core of the thesis is described in details in the following chapters. Chapter 4 defines the theoretical background and the methodology for analysis of the Internet trends. Chapter 5 introduces the architecture design of the new framework that provides an end-to-end approach to the analysis of selected Internet trends. Chapter 6 describes the experiments carried out to evaluate the theoretical assumptions. The contribution of the thesis is described in more detail below.

A novel general model has been proposed for gathering and processing data from Web 2.0 (see Section 4.1). The model builds on Webometrics and starts from the idea that almost any text can be machine-recognised. This idea is supported by the current research in Sentiment Analysis. Original Webometric techniques have been reaching their limits, and they do not fully reflect the needs of the current Web. Therefore, the Sentiment Analysis has been used along with original Webometrics to define the methodology for the evaluation of Internet trends. Section 6.1 describes the experiment that has been used to verify the theoretical assumptions. A corpus of the blog posts has been used for the evaluation of the most searched expressions in the Google search engine. The results of the experimental system represent a chronological view of the trend evaluation according to the public opinion.

The proposed methodology has been compared with the similar methods that can be used for the evaluation tasks on the Web (see Section 6.2). A corpus of the movie reviews has served for this study. Each of the selected methods provides a different methodology to the evaluation of Internet trends. The first of tested methods, Sentiment Analysis, evaluates a textual content and provides the output based on the positive/negative feedback from the reviewers. The second, Web Mention Analysis,

emphasises the frequency of making reviews and reports the overall number of reviews in a given period. The last tested method, Social Network Analysis, determines the prestige of the reviewer and thus defines the quality of the source. However, the combination of individual methods can provide much more accurate results with respect to the desired area of interest.

Social Network Analysis has been used to enhance the initial general model for the evaluation in the social network sphere (see Section 4.4). 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. Section 6.3 deals with the influence of the threshold changes on the evaluation in the network of trends.

All of the proposed Sentiment Analysis studies utilise the lexicon-based method to determine the sentiment of the trend. The SentiWordNet lexica of English words serve as a data input for this method. The main problem of the lexicon-based method is that the word may have multiple senses with a different sentiment polarity and strength, and it is hard to recognise which sense should be used in a specific context. There are several strategies for computing a prior polarity to determine the sense disambiguation (see Section 3.6.2). Most of these strategies utilise the fact that the first sense in the lexicon is the most frequent, however, the results may not always be very reliable. Section 4.5 describes two new methods to improve sentiment classification for multiple-topic related words. The first method, Domain Elimination, enhances lexicon-based analysis by combining sentiment and domain lexicons. The second method, Cosine Sense-Similarity, exploits lexicon gloss definition to calculate cosine similarity for a more accurate sentiment sense recognition. The experiments in Section 6.4 demonstrate that the both methods are recommended for sentiment analysis tasks.

Architecture design of the new framework for an end-to-end evaluation of the selected Internet trends is described in Chapter 5. The framework associates crawler, analysis modules and fully configurable user interface to define which data should be analysed and how. The interface allows the user to compose a graph where each node represents one analytical module. The evaluation of a trend is then performed by a sequential pass of the graph where the processed data is transferred from one node/module to the other. The outputs of the individual modules and also complete results are gradually displayed to the user in chronological order.

7.3 Future Work

The major approaches introduced in this theses are focused on the evaluation of Internet trends, where as the trend may be defined any expression which is mentioned online. However, in a wider perspective, the trend can be described in very general terms, and there might be many specific factors that affect its evaluation. It would also be useful to identify which trends have the greatest impact on the general trend. For instance, the name of a city can be defined as a general trend. Then, the analysis of underlying trends such as transport, sights, shopping, could significantly affect the evaluation of the main trend.

It would be interesting to use a language other than English for the evaluation of trends. This would greatly expand the possibilities for data analysis on any website. Several dictionaries for other languages have been appearing outside of academia. Their implementation in the lexicon-base sentiment classification method may also bring a significant improvement in the evaluation of trends.

The main problem of lexicon-based approach is that the incorrect sentiment can be assigned to the word. The implementation of our methodology could be further enhanced by linguistic and full-text machine learning algorithms. Each method has some disadvantages, however, it would be great to use them together to find their common benefits.

The proposed framework can be further extended with more analytical and auxiliary modules. The extension of a set of implemented algorithms will provide more precise results and enable more varied possibilities for the trend evaluation. Based on the continued use of the framework the recommendation can be formed about which algorithm use for a particular task, eventually how to work with data in a specific domain.

Bibliography

Abel, F., Hauff, C., Houben, G.-J., Stronkman, R., and Tao, K. (2012). Semantics + Filtering + Search = Twitcident Exploring Information in Social Web Stre. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, pages 285–294. ACM.

Agrawal, R., Gollapudi, S., Kenthapadi, K., Srivastava, N., and Velu, R. (2010). Enriching textbooks through data mining. In *Proceedings of the 1st ACM Symposium on Computing for Development*, ACM DEV '10, pages 19:1–19:9, New York, NY, USA. ACM.

Agrawal, S. and Siddiqui, T. j. (2009). Using Syntactic and Contextual Information for Sentiment Polarity Analysis. In *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human*, ICIS '09, pages 620–623, New York, NY, USA. ACM.

Aguillo, I. F., Ortega, J., Fernández, M., and Utrilla, A. (2010). Indicators for a Webometric Ranking of Open Access Repositories. *Scientometrics*, 82(3):477–486.

Aguillo, I. F., Ortega, J. L., and Fernández, M. (2008). Webometric Ranking of World Universities: Introduction, Methodology, and Future Developments. *Higher Education in Europe*, 33(2-3):233–244.

Aichner, T. and Jacob, F. (2015). Measuring the Degree of Corporate Social Media Use. *International Journal of Market Research*, 57(2):257–275.

Almind, T. C. and Ingwersen, P. (1997). Informetric Analyses on the World Wide Web: Methodological Approaches to "Webometrics". *Journal of Documentation*, 53(4):404–426.

Ausserhofer, J. and Maireder, A. (2013). National Politics on Twitter: Structures and Topics of a Networked Public Sphere. *Information, Communication & Society*, 16(3):291–314.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. LREC '10, pages 2200–2204.

Barabási, A. L., Gulbahce, N., and Loscalzo, J. (2011). Network Medicine: A Network-Based Approach to Human Disease. *Nat Rev Genet*, 12(1):56–68.

Bashan, A., Bartsch, R. P., Kantelhardt, J. W., Havlin, S., and Plamen, C. I. (2012). Network Physiology Reveals Relations Between Network Topology and Physiological Function. *Nature Communications*, 3(702):1–9.

Benedetti, F. (2013). Tutorial of Sentiment Analysis. In *Sentiment Analysis for the course of Big Data Analysis, Department of Computer Engineering of Modena and Reggio Emilia*. Retrieved 10 July 2016, from http://www.slideshare.net/faigg/tutotial-of-sentiment-analysis.

Berners-Lee, T. (1989). Information Management: A Proposal. *Word Journal Of The International Linguistic Association*, 2(5):1–10.

Björneborn, L. (2004). *Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach*. PhD thesis, Royal School of Library and Information Science, Copenhagen: Department of Information Studies, Denmark.

Björneborn, L. (2005). Identifying Small-World Connectors Across an Academic Web Space - A Webometric Study. In *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*, volume 1, pages 56–66. Karolinska University Press.

Björneborn, L. and Ingwersen, P. (2004). Toward a Basic Framework for Webometrics. *Journal of the American Society for Information Science and Technology*, 55(14):1216–1227.

Brin, S. and Page, L. (2012). Reprint of: The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks*, 56(18):3825–3833.

Bross, J., Richly, K., Kohnen, M., and Meinel, C. (2012). Identifying the Top-Dogs of the Blogosphere. *Social Network Analysis and Mining*, 2(1):53–67.

Brown, K. R., Otasek, D., Ali, M., McGuffin, M. J., Xie, W., Devani, B., Toch, I. L., and Jurisica, I. (2009). NAViGaTOR: Network Analysis, Visualization and Graphing Toronto. *Bioinformatics*, 25(24):3327–3329.

Chaumartin, F.-R. (2007). UPAR7: A Knowledge-based System for Headline Sentiment Tagging. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 422–425, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y., and Chau, M. (2004). Crime Data Mining: A General Framework and Some Examples. *Computer*, 37(4):50–56.

Crilly, T. (2007). *50 Mathematical Ideas You Really Need to Know*. 50 Ideas. Book Sales.

Cronin, B., Snyder, H. W., Rosenbaum, H., Martinson, A., and Callahan, E. (1998). Invoked on the Web. *Journal of the American Society for Information Science and Technology*, 49(14):1319–1328.

Deguchi, T., Takahashi, K., Takayasu, H., and Takayasu, M. (2014). Hubs and Authorities in the World Trade Network Using a Weighted HITS Algorithm. *PLoS ONE*, 9(7):1–16.

Denecke, K. (2008). Accessing Medical Experiences and Information. In *European Conference on Artificial Intelligence, Workshop on Mining Social Data*, volume 21.

Denecke, K. (2009). Are SentiWordNet Scores Suited for Multi-Domain Sentiment Classification? In *Proceedings of the 4th International Conference on Digital Information Management*, pages 32–37.

Dhande, L. L. and Patnaik, G. K. (2014). Analyzing sentiment of movie review data using Naive Bayes neural classifier. *International Journal of Emerging Trends & Technology in Computer Science*, 3(4):313–320.

Esuli, A. and Sebastiani, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation*, pages 417–422.

Fahrni, A. and Klenner, M. (2008). Old Wine or Warm Beer: Target-Specific Sentiment Analysis of Adjectives. *Computational Linguistics*, 2(3):60–63.

Fellbaum, C. (2005). WordNet and Wordnets. *Encyclopedia of Language & Linguistics*, 13:665–670.

Fellbaum, C. (2010). WordNet. In Poli, R., Healy, M., and Kameas, A., editors, *Theory and Applications of Ontology: Computer Applications*, pages 231–243.

Foot, K. and Schneider, S. (2006). *Web Campaigning*. Acting with technology. MIT Press.

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35–41.

Freeman, L. C. (1978). Centrality in Social Networks Conceptual Clarification. *Social Networks*, 1(3):215–239.

Garfield, E. (2005). The Agony and the Ecstasy: The History and Meaning of the Journal Impact Factor. *International Congress on Peer Review And Biomedical Publication*, pages 1–22.

Gatti, L. and Guerini, M. (2012). Assessing Sentiment Strength in Words Prior Polarities. In *Proceedings of COLING 2012: Posters*, pages 361–370, Mumbai, India. The COLING 2012 Organizing Committee.

General Register Office (1913). *Annual Report of the Registrar-General for England and Wales*. Number 74. Great Britain. Her Majesty's Stationery Office.

Glance, N. S., Hurst, M., and Tomokiyo, T. (2004). BlogPulse: Automated Trend Discovery for Weblogs. In *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, analysis and dynamics*, pages 1–8. ACM.

Graham, T., Broersma, M., Hazelhoff, K., and van 't Haar, G. (2013). Between Broadcasting Nessages and Interacting with Voters: The Use of Twitter During the 2010 UK General Election Campaign. *Information, Communication & Society*, 16(5):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. *Scientometrics*, 87(1):133–147.

Gupta, A., Kaur, H., and Batra, S. (2015). Topic sensitive web page ranking through graph database. In Shetty, R. N., Prasad, N., and Nalini, N., editors, *Emerging Research in Computing, Information, Communication and Applications*, volume 1, pages 519–527. Springer India, New Delhi.

Gupta, A., Kumaraguru, P., Castillo, C., and Meier, P. (2014). TweetCred: A Realtime Web-Based System for Assessing Credibility of Content on Twitter. In *Social Informatics: 6th International Conference, SocInfo 2014. Barcelona, Spain*, pages 228–243. Springer International Publishing.

Han, S. K., Shin, D., Jung, J. Y., and Park, J. (2009). Exploring the Relationship Between Keywords and Feed Elements in Blog Post Search. *World Wide Web*, 12(4):381–398.

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

Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the Semantic Orientation of Adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, pages 174–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hickson, I., Berjon, R., Faulkner, S., Leithead, T., Navara, E. D., O'Connor, E., and Pfeiffer, S. (2014). HTML5. In *W3C Recommendation 28 October 2014*. Retrieved 15 June 2016, from https://www.w3.org/TR/html5/.

Holmberg, K. (2009). *Webometric Network Analysis: Mapping Cooperation and Geopolitical Connections Between Local Government Administration on the Web*. PhD thesis, Åbo Akademis förlag-Åbo Akademi University Press, Biskopsgatan: Faculty of Economics and Social Sciences, Information Studies, Finland.

Holmberg, K. and Thelwall, M. (2009). Local Government Web Sites in Finland: A Geographic and Webometric Analysis. *Scientometrics*, 79(1):157–169.

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

Ingwersen, P. (1998). The Calculation of Web Impact Factors. *Journal of Documentation*, 54(2):236–243.

Jacobs, I. and Walsh, N. (2004). Architecture of the World Wide Web. In *W3C Recommendation*, volume 1. Retrieved 12 June 2016, from http://www.w3.org/TR/webarch/.

John, E. T., Skaria, B., and Shajan, P. (2016). An Overview of Web Content Mining Tools. *Bonfring International Journal of Data Mining*, 6(1):1.

Keith, M. and Schincariol, M. (2013). *Pro JPA 2*. Apress.

Kleinberg, J. M. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):668–677.

Konda, M. (2014). *Just Hibernate*. " O'Reilly Media, Inc.".

Kretschmer, H. and Aguillo, I. F. (2004). Visibility of Collaboration on the Web. *Scientometrics*, 61(3):405–426.

Lang, P. J., Mcteague, L. M., and Bradley, M. M. (2016). RDoC, DSM, and the Reflex Physiology of Fear: A Biodimensional Analysis of the Anxiety Disorders Spectrum. *Psychophysiology*, 53(3):336–347.

Langville, A. N. and Meyer, C. D. (2006). *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press.

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

Li, R. H., Yu, J. X., Huang, X., and Cheng, H. (2012). A Framework of Algorithms: Computing the Bias and Prestige of Nodes in Trust Networks. *PLoS ONE*, 7(12).

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

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

London, A. and Csendes, T. (2013). HITS Based Network Algorithm for Evaluating the Professional Skills of Wine Tasters. In *Proceedings of the 8th International Symposium on Applied Computational Intelligence and Informatics*, SACI '13, pages 197–200. IEEE.

López Barbosa, R. R., Sánchez-Alonso, S., and Sicilia-Urban, M. A. (2015). Evaluating hotels rating prediction based on sentiment analysis services. *Aslib Journal of Information Management*, 67(4):392—407.

Lyon, B. G. (2005). Opte as an Aesthetic Experience. *British Journal of Psychology*, pages 1–6.

Lyon, B. G. (2015). The Internet 2015. *The Opte Project*. Retrieved 19 April 2016, from http://www.opte.org.

MacDonald-Wallis, K., Jago, R., and Sterne, J. A. C. (2012). Social Network Analysis of Childhood and Youth Physical Activity: A Systematic Review. *American Journal of Preventive Medicine*, 43(6):636–642.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Mejova, Y. and Srinivasan, P. (2011). Exploring Feature Definition and Selection for Sentiment Classifiers. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, pages 546–549.

Nagy, A. and Stamberger, J. (2012). Crowd sentiment detection during disasters and crises. In *Proceedings of the 9th International ISCRAM Conference*, pages 1–9.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2009). SentiFul: Generating a Reliable Lexicon for Sentiment Analysis. In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2011). Affect Analysis Model: Novel Rule-Based Approach to Affect Sensing from Text. *Natural Language Engineering*, 17(01):95–135.

Newman, M. E. (2010). *Networks. An Introduction*. Oxford University Press.

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

Opsahl, T., Agneessens, F., and Skvoretz, J. (2010). Node Centrality in Weighted Networks: Generalizing Degree and Shortest Paths. *Social Networks*, 32(3):245–251.

Otte, E. and Rousseau, R. (2002). Social Network Analysis: A Powerful Strategy, Also For the Information Sciences. *Journal of Information Science*, 28(6):441–453.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. Number 66. Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Pak, A. and Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. LREC '10, pages 1320–1326.

Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Park, H. (2011). How Do Social Scientists Use Link Data from Search Engines to Understand Internet-Based Political and Electoral Communication? *Quality & Quantity*, pages 1–15.

Patel, P. (2014). Research of Page Ranking Algorithm on Search Engine Using Damping Factor. *International Journal of Advance Engineering and Research Development*, 1(1):1–6.

Pennebaker, J. W. (2011). *The Secret Life of Pronouns: What Our Words Say About Us*, volume 1890. Bloomsbury Press.

Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Stříteský, V., and Holzinger, A. (2013). Opinion Mining on the Web 2.0 – Characteristics of User Generated Content and Their Impacts. *Lecture Notes in Computer Science LNCS 7947*, pages 35–46.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). MultiWordNet: Developing an Aligned Multilingual Database. In *Proceedings of the 1st International WordNet Conference*, pages 293–302.

Potthast, M. and Becker, S. (2010). Opinion Summarization of Web Comments. In Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., and van Rijsbergen, K., editors, *Advances in Information Retrieval*, volume 5993 of *Lecture Notes in Computer Science*, pages 668–669. Springer Berlin Heidelberg.

Prabowo, R. and Thelwall, M. (2009). Sentiment Analysis: A Combined Approach. *Journal of Informetrics*, 3(2):143–157.

PrimeTek (2015). PrimeFaces. *PrimeTek Informatics*. Retrieved from http://primefaces.org.

Rastogi, S., Singhal, R., and Kumar, A. (2014). An Improved Sentiment Classification using Lexicon into SVM. *International Journal of Computer Applications*, 95(1):37–42.

Renu, T. V. and Gaur, D. (2014). Implementation of Web Structure Mining with Breadth First Search and Depth First Search. *International Journal of Applied Information Systems*, 7(2):19–21. Published by Foundation of Computer Science, New York, USA.

Rethlefsen, M. L., Rothman, D. L., and Mojon, D. S. (2009). Google Scholar. In *Internet Cool Tools for Physicians*, pages 37–40. Springer Berlin Heidelberg.

Scott, J. (2012). *Social network analysis*. Sage.

Sing, J. K., Sarkar, S., and Mitra, T. K. (2012). Development of a Novel Algorithm for Sentiment Analysis Based on Adverb-Adjective-Noun Combinations. In *Emerging Trends and Applications in Computer Science*, volume 1, pages 38–40.

Singh, V. K., Piryani, R., Uddin, A., and Waila, P. (2013). Sentiment Analysis of Movie Reviews: A New Feature-Based Heuristic for Aspect-Level Sentiment Classification. In *Proceedings of the IEEE International Multi Conference on Automation, Computing, Control, Communication and Compressed Sensing*, iMac4s 2013, pages 712–717.

Son, S. W., Christensen, C., Grassberger, P., and Paczuski, M. (2012). PageRank and Rank-Reversal Dependence on the Damping Factor. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 86(6).

Statista Inc. (2016a). Number of Monthly Active Twitter Users Worldwide from 1st Quarter 2010 to 3rd Quarter 2016. *Social Media & User-Generated Content*. Retrieved 15 January 2017, from https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users.

Statista Inc. (2016b). Number of Social Network Users Worldwide from 2010 to 2020. *Social Media & User-Generated Content*. Retrieved 15 July 2016, from http://www.statista.com/statistics/278414/number-of-worldwide-social-network-users.

Sterne, J. (2003). *Web Metrics: Proven Methods for Measuring Web Site Success*. John Wiley & Sons.

Suganya, R. (2014). Adapting Hits Algorithm For Image Search In Favour of User Profile. *International Journal for Innovative Research in Science and Technology*, 1(6):305–310.

Sun, Y., Councill, I. G., and Giles, C. L. (2010). The Ethicality of Web Crawlers. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, WI '10, pages 668–675.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267–307.

Takama, Y., Matsumura, A., and Kajinami, T. (2007). Interactive Visualization of News Distribution in Blog Space. *New Generation Computing*, 26(1):23–38.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Ter Wal, A. L. J. and Boschma, R. A. (2009). Applying Social Network Analysis in Economic Geography: Framing Some Key Analytic Issues. *The Annals of Regional Science*, 43(3):739–756.

Terpstra, T., De Vries, A., Stronkman, R., and Paradies, G. (2012). *Towards a Realtime Twitter Analysis During Crises for Operational Crisis Management*. Simon Fraser University.

The Internet Archive (2016). *Internet Archive: Digital Library of Free Books, Movies, Music & Wayback Machine*. Retrieved from https://archive.org.

Thelwall, M. (2007). Blog Searching: The First General-Purpose Source of Retrospective Public Opinion in the Social Sciences? *Online Information Review*, 31:277–289.

Thelwall, M. (2008). Bibliometrics to Webometrics. *Journal of Information Science*, 34(4):605–621.

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

Thelwall, M. and Buckley, K. (2013). Topic-Based Sentiment Analysis for the Social Web: The Role of Mood and Issue-Related Words. *Journal of the American Society for Information Science and Technology*, 64(8):1608–1617.

Thelwall, M., Buckley, K., and Paltoglou, G. (2011). Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418.

Thelwall, M. and Sud, P. (2012). Webometric research with the bing search api 2.0. *Journal of Informetrics*, 6(1):44–52.

Thet, T. T., Na, J.-C., Khoo, C. S., and Shakthikumar, S. (2009). Sentiment Analysis of Movie Reviews on Discussion Boards Using a Linguistic Approach. In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, TSA '09, pages 81–84, New York, NY, USA. ACM.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1 of *NAACL '03*, pages 173–180.

Turek, W., Opalinski, A., and Kisiel-Dorohinicki, M. (2011). Extensible Web Crawler - Towards Multimedia Material Analysis. In Dziech, A. and Czyżewski, A., editors, *Multimedia Communications, Services and Security*, volume 149, pages 183–190. Springer Berlin Heidelberg.

Vilares, D., Thelwall, M., and Alonso, M. A. (2015). The Megaphone of the People? Spanish SentiStrength for Real-time Analysis of Political Tweets. *Journal of Information Science*, 41(6):799–813.

Viswanath, B., Mislove, A., Cha, M., and Gummadi, K. P. (2009). On the Evolution of User Interaction in Facebook. In *Proceedings of the 2nd ACM Workshop on Online Social Networks*, WOSN '09, pages 37–42, New York, NY, USA. ACM.

Wadia, Z., Saleh, H., and Christensen, A. (2014). *Pro JSF and HTML5: Building Rich Internet Components*. Apress.

Wang, H., Lu, Y., and Zhai, C. (2010). Latent aspect rating analysis on review text data: A rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 783–792, New York, NY, USA. ACM.

Wang, H., Lu, Y., and Zhai, C. (2011). Latent Aspect Rating Analysis Without Aspect Keyword Supervision. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 618–626, New York, NY, USA. ACM.

Wasserman, S. and Faust, K. (1994). Social Network Analysis: Methods and Applications. In *Methods and Applications*, page 825. Cambridge university press.

White, H. D. and McCain, K. W. (1989). Bibliometrics. *Annual Review of Information Science and Technology*, 24:119–186.

Wu, G., Wang, Y.-C., and Jin, X.-Q. (2012). A Preconditioned and Shifted GMRES Algorithm for the PageRank Problem with Multiple Damping Factors. *SIAM Journal on Scientific Computing*, 34(5):A2558–A2575.

Xu, J. and Chen, H. (2005). Criminal Network Analysis and Visualization. *Communications of the ACM*, 48(6):100–107.

Yang, D., Zhang, D., Yu, Z., and Wang, Z. (2013). A Sentiment-enhanced Personalized Location Recommendation System. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, HT '13, pages 119–128, New York, NY, USA. ACM.

Yang, T. (2006). Large Scale Internet Search at Ask.com. In *Proceedings of the 5th International Conference on Scalable Information Systems*, SIS '06.

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., and Liu, B. (2015). Combining Lexicon-Based and Learning-Based Methods for Twitter Sentiment Analysis. *International Journal of Electronics, Communication and Soft Computing Science & Engineering*, 89:1–8.

Publications of the Author

Publications in Peer-Reviewed Journals

Malinský, R. and Jelínek, I. (2016b). The Application of Modern Webometric Methods on the Evaluation of Trends in the Social Network Sphere. *IADIS International Journal on WWW/Internet*, 14(2):58-71. ISSN 1645-7641. [50%]

Malinský, R. and Jelínek, I. (2015b). The Visualizer for Real-Time Analysis of Internet Trends. *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 9(12):2174–2178. ISSN 2010-376X. [50%]

Malinský, R. and Jelínek, I. (2012). Sentiment Analysis: Popularity of Candidates for the President of the United States. *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, 6(12):3679-3681. ISSN 2010-376X. [50%]

Malinský, R. and Jelínek, I. (2011b). A Novel Web Metric for the Evaluation of Internet Trends. *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 5(9):1008-1011. ISSN 2010-376X. [50%]

Publications Excerpted from ISI Web of Science

Malinský, R. and Jelínek, I. (2014). Comparing Methods of Trend Assessment. In Casteleyn, S., Rossi, G., and Winckler, M., editors, *Web Engineering*, volume 8541, pages 554–557. Springer International Publishing, Switzerland. doi:10.1007/978-3-319-08245-5_49. [50%]

Malinský, R. and Jelínek, I. (2010). Improvements of Webometrics by Using Sentiment Analysis for Better Accessibility of the Web. In Daniel, F., and Facca, F. M., editors, *Current Trends in Web Engineering*, volume 6385, pages 581–586. Springer Berlin Heidelberg. doi:10.1007/978-3-642-16985-4_59. [50%]

The paper has been cited in:

- Heil, S., Drechsel, M., and Gaedke, M. (2015). Supporting the Developmentof Team-Climate-Aware Collaborative Web Applications. In Cimiano, P., Frasincar, F., Houben, G. J., and Schwabe, D., editors, *Engineering the Web in the Big Data Era*, volume 9114, pages 663–666. Springer International Publishing, Switzerland. doi:10.1007/978-3-319-19890-3_52
- Lorentzen, D. G. (2014). Webometrics benefitting from web mining? An investigation of methods and applications of two research fields. *Scientometrics*, 99(2):409–445. doi:10.1007/s11192-013-1227-x
- Joselyn, J. and Surulinathi, M. (2013). *Webometric study of J. R. D. Tata Memorial Library in Indian Institute of Science, Bangalore*, Bharathidasan University Institutional Repository, Department of Library and Information Science. MLIS Project.

Publications Excerpted from Elsevier Scopus

Malinský, R. and Jelínek, I. (2016a). The Evaluation of Node's Power in the Social Network Sphere. In *Proceedings of the IADIS International Conference WWW/INTERNET 2016*, volume 15, pages 136-142, Mannheim, Germany. IADIS Press. ISBN 978-989-8533-57-9. [50%]

Malinský, R. and Jelínek, I. (2015a). Trend Analysis Framework. In *Proceedings of the IADIS International Conference WWW/INTERNET 2015*, volume 14, pages 161-166, Maynooth, Greater Dublin, Ireland. IADIS Press. ISBN 978-989-8533-44-9. [50%]

Malinský, R. and Jelínek, I. (2013b). Trend Classification Methodology. In *Proceedings of the IADIS International Conference WWW/INTERNET 2013*, volume 12, pages 389–393, Fort Worth, Texas, USA. IADIS Press. ISBN 978-989-8533-16-6. [50%]

Other Publications

Malinský, R. (2013a). Webometric Overview. In *Proceedings of the 17th International Student Conference POSTER 2013*, Prague, Czech Republic. Czech Technical University in Prague. ISSN 978-80-01-05242-6. [100%]

Malinský, R. & Jelínek, I. (2011a). Model for Gathering and Processing Data from Web 2.0. In *Workshop 2011*, Prague, Czech Republic. Czech Technical University in Prague. [50%]

Manuscripts Submitted for Publication

Malinský, R. and Jelínek, I. (2017). *Using Sentiment Domain to Eliminate Word Sense Disambiguation*. Manuscript submitted for publication. [50%]

Appendix A

Penn Treebank English Part-of-Speech Tag Set

Tags	Description	Tags	Description
CC	Coordinating conjunction	TO	*to*
CD	Cardinal number	UH	Interjection
DT	Determiner	VB	Verb, base form
EX	Existential *there*	VBD	Verb, past tense
FW	Foreign word	VBG	Verb, gerund or present participle
IN	Preposition or subordinating conjunction	VBN	Verb, past participle
JJ	Adjective	VBP	Verb, non-3rd person singular present
JJR	Adjective, comparative	VBZ	Verb, 3rd person singular present
JJS	Adjective, superlative	WDT	*Wh*-determiner
LS	List item marker	WP	*Wh*-pronoun
MD	Modal	WP$	Possessive *wh*-pronoun
NN	Noun, singular or mass	WRB	*Wh*-adverb
NNS	Noun, plural	#	Pound sign
NNP	Proper noun, singular	$	Dollar sign
NNPS	Proper noun, plural	.	Sentence-final punctuation
PDT	Predeterminer	,	Comma
POS	Possessive ending	:	Colon, semi-colon
PRP	Personal pronoun	(	Left bracket character
PRP$	Possessive pronoun	)	Right bracket character
RB	Adverb	"	Straight double quote
RBR	Adverb, comparative	‘	Left open single quote
RBS	Adverb, superlative	“	Left open double quote
RP	Particle	’	Right close single quote
SYM	Symbol (mathematical or scientific)	”	Right close double quote

Retrieved from Marcus et al. (1993).