TREND MINING

Dissertation
zur Erlangung des akademischen Grades
der Doktorin der Naturwissenschaften
(Dr. rer. nat.)

eingereicht von
Dipl.-Inform.
Olga Katarzyna Streibel

beim Institut für Informatik
am Fachbereich Mathematik und Informatik
Freie Universität Berlin

Tag der Disputation: 28.10.2013

Gutachter:

Prof. Dr.-Ing.
Robert Tolksdorf
Netzbasierte Informationssysteme
Institut für Informatik
Freie Universität Berlin
Germany

Prof. Ph.D
Danilo Montesi
Dipartimento di Informatica -
Scienza e Ingegneria
Università di Bologna
Italy

23. Juli 2013

Abstract

In terms of Information Retrieval (IR), a *trend* is defined as *a topic area that is growing in interest and utility over time*. An example of a trend would thus be the general topic *financial crisis* that started to appear on the market in late 2007 and early 2008, or the *Arab Spring* that started to appear on the news in 2011. Several approaches based on methods from text mining and machine learning can be successfully applied to the problem of mining trends in text collections. Among others, the most popular are probabilistic topic models and diverse clustering methods.

The weakness of the existing research in automatic trend detection in texts lies in:

1. inconsistency in the definition of a *trend*
2. lack of a general scientific approach for *trend mining*
3. lack of the integration of explicit knowledge and therefore the difficulty in the interpretation of algorithm's results.

The scientific contribution of this research is contained in the suggestion to deal with the trend detection from the perspective of *trend mining* that is being defined here.

As a solution for the problem of difficulty in the interpretation of the results from the common trend detection techniques, this research proposes the *trend template* that is a knowledge-based trend mining approach. Based on this trend template, two directions of implementation are introduced: trend ontology and trend-indication (the trend weighting method).

The trend ontology works as an a-priori model and enables the discovery of a trend structure in the web documents corpus. Tests with this method on a test corpus show that mining trends with an a-priori model while integrating explicit knowledge leads to a better quality of results considering their interpretability.

The trend-indication approach is based on time-incorporating weighting methods for selection of trend features from web documents. It enables the reduction of features that are considered in the process of trend mining, and therefore reduces the data so that only time-relevant information is considered for further analysis. This method's results on our web document corpus show that time-based weighting functions alone can help in discovering trend-relevant features.

Both the trend ontology and the trend-indication approaches are implemented in the *tremit* tool (TREnd MIning Tool), a test tool developed for this thesis, and are tested on a test corpus. The test corpus consists of 35,635 business news and 4,696 DAX (Deutscher Aktienindex – German stock market) reports from German web sites in a late 2007 and early 2008. The results are compared with the

standard method results of a LDA-based topic model and the k-means clustering algorithm on the same test corpus. Discussion of the results is contained in the experimental part of the thesis.

Motivation

Several years ago, I was about to finish my diploma thesis on "*Classification* and *Generating* of *user-based Information Profiles* using *Machine Learning Methods* and *Algorithms*", focusing on interpretation of results from C4.5 [Quinlan, 1993], PART, REP-Tree, and Self-Organizing Map (SOM) [Witten and Eibe, 2005] – algorithms applied to a set of test documents from a research project PIA (Personal Information Assistant) [PIA, 2013]. I was wondering that some of the results were straightforward and easily interpretable, whereas the others were difficult to understand for human. Following the decision path of J48 (the WEKA [Hall et al., 2009] implementation of the C4.5 algorithm), I asked myself what if, the decision tree algorithm could somehow *know* what it would be learning. What if, at every step in which it takes the decision on which attribute to take for splitting the set of learning objects it could be aware of its own decision, provided the necessary knowledge?

It is important to first define what this *knowing* means and how it can be represented. And second, we should be aware that making a learning algorithm *really knowing* what it is about to learn is most probably a 'typical' AI (artificial intelligence) science-fiction. However, I think that my need for somehow putting knowledge into the algorithm in the hope that it will help to make it *intelligent*, make it producing more *comprehensive* results for human users, emerged after finishing my diploma.

A year after my graduation, I worked for a research project on Trend Mining: Analysis and Fusion of multimodal Data (TREMA). In general, the project focused on exploring the algorithms for analysis of combined data: textual data from web documents and numeric data from the stock exchange market, in order to predict trends. The term *trend mining* appeared during this project constantly alongside *data* and *text analysis* and *trend detection methods*. After 12 months of research with a strong focus on industrial applications and a successful project completion, many questions remained open for me. The main one: What do we actually mean by *mining trends* – could we be more scientific about it? This question is the basis for this thesis. For anybody who would like to know right away how this thesis handles the topic, I recommend looking at the visualization of the thesis as shown in Figure 1.

Berlin, January 2013

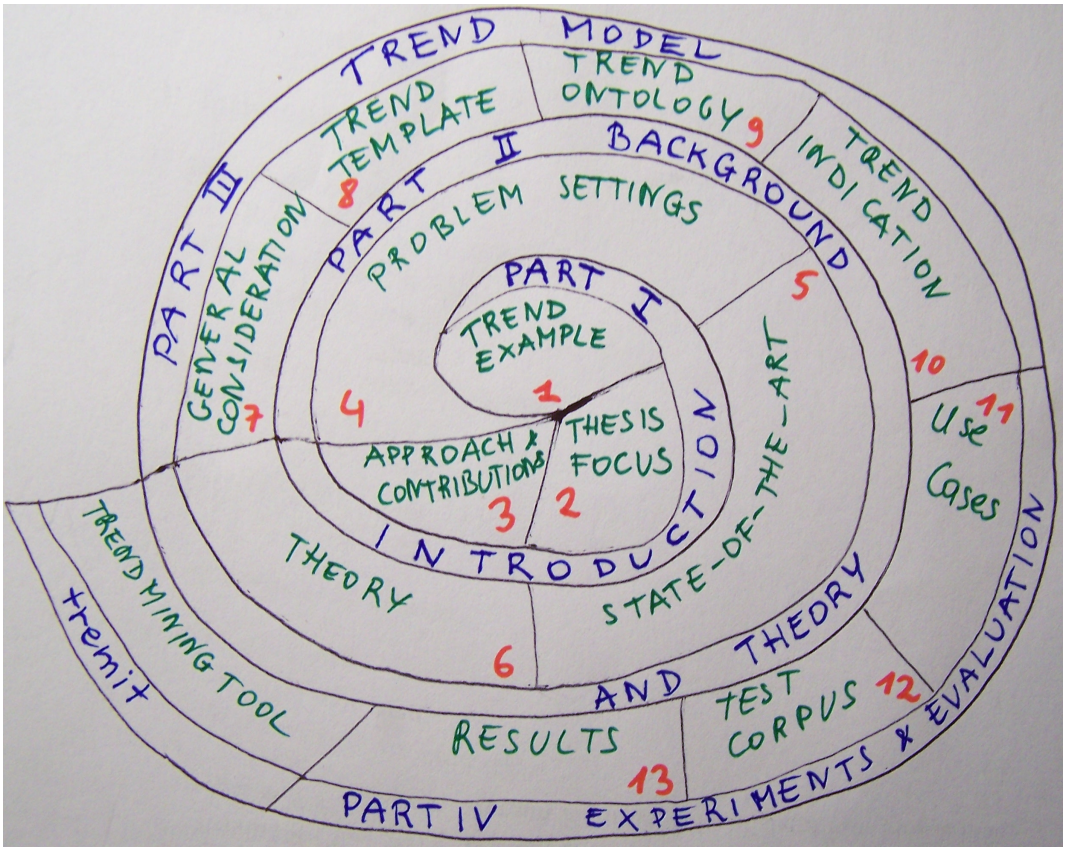


Figure 1: Thesis visualization. Source: *author*.

Acknowledgments

It is imaginable that, while writing a doctoral thesis, one could write a separate doctoral thesis, let us call it a *meta-thesis*, of which the research subject would be the process of writing the doctoral thesis. Besides the chapters on creativity, the beauty of procrastination, and its results, it is imaginable that the *Acknowledgments* could be placed in one entire chapter of the meta-thesis if one would start acknowledging the complete chain of circumstances including people that preceded, accompanied and made it possible – the creation process of the original doctoral thesis. Since there is no thesis without a subject, and there is no subject without a context or environment around it, we can conclude that a doctoral thesis needs a huge number of bytes of other doctoral theses, theories, publications, and tools that somehow allowed for understanding and studying a given research direction and for developing a given research idea within the given context. And first of all it needed people. We make an obvious assumption that there is no need for thanking the creator of TeX for triggering the development of the helpful LaTeX and there is a tendency for appropriately limited and personal acknowledgments. Being a grateful part of my environment, I would like to bring it to the point here and have a beer with you in *real life* ;)

```
FOR (support && discussions && patience &&  great time){
//für die Unterstützung, die Gespräche, die Geduld und den Spaß
//za wsparcie, akademickie dysputy, cierpliwość i cudowny czas
  THANK_YOU_ALL(){ //danke, grazie, dziękuję!
  Prof. Robert Tolksdorf, Prof. Danilo Montesi;
  Petra Ristau and ALL from the TREMA project;
  My colleagues from NBI Group@Freie Universität Berlin;
  My colleagues from CSW Group@Freie Universität Berlin;
  My students @Freie Universität Berlin;
  My colleagues @Universita di Bologna;
  Dr. Malgorzata Mochól, Magnus Niemann, Arne Handt;
  Prof. Adrian Paschke, Radoslaw Oldakowski, Ralph Schäfermeier,
  Dr. Markus Luczak-Rösch, Ralf Heese, Gökhan Coşkun, Kia Teymourian,
  Prof. Elfriede Fehr, Prof. Claudia Müller-Birn;
  Monsz ;) Piotr Majchrzyk, Mama Janina Streibel, Konrad/Dorota/Monika Streibel,
  Dr. Rehab Alnemr, Vlad Tanasescu, Dr. Stephan Sigg,
  Kelly Callahan, Ferda Egritag, Elina Freigang,
  Dr. Sandra Boichman latv. Boihmane, Jeannette Schüler,
  Dr. Esther Manya Ndjeka, Monika Kwiatkowski, Antje Schlieder-Parma,
  Mario Mauritsch, Matthias Scheide, Daniela Berndt, my cat ;), ... }}
```

Dla mojego taty, Ryszarda Streibel (1947-2010) oraz mojego drogiego przekornego przyjaciela, Axla Dietze (1968-2013), którzy obaj odeszli tak szybko, zmieniając mi jakoś mój pogląd na wszystko.

Für meinen Vater, Ryszard Streibel (1947-2010) und meinen lieben spezifischen Freund, Axel Dietze (1968-2013), die so zeitig gegangen sind und ließen mich irgendwie alles anders betrachten.

Contents

List of Figures

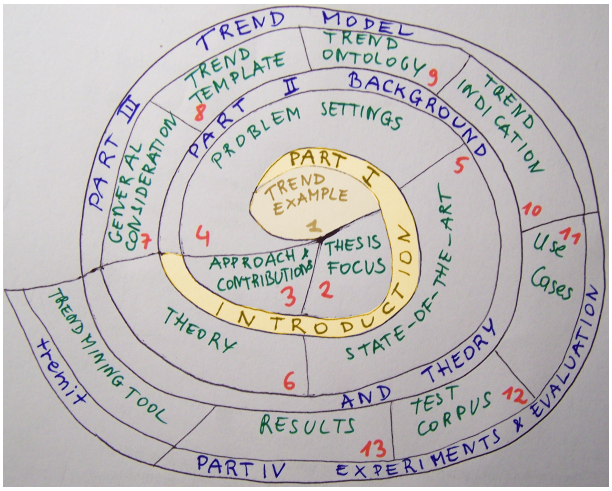
[0] On pages 117-169 figures' source: *author*

List of Tables

[0]On pages 118-146 tables' source:*author*

Part I

Introduction



CHAPTER 1

Trend example

This chapter delivers an example of a trend from which we learn the relations between texts, the web and an emerging topic.

The reports of political developments in North Africa in the period of January to February 2011 were dominated by the 'breaking news' about protests and revolutions, starting with reports of people in Tunisia overthrowing their government, followed by news broadcast via social networks from people taking part in protests in Egypt and by emerging social network updates and reports on protests in Libya and plans for protests in Algeria. Clearly, there was a trend toward a political change from old political systems, sometimes referred to as regimes by many nations, to democracy-based systems in Northern African countries.

Figure 1.1: A timeline-based visualization of selected reports. Source: *author.*

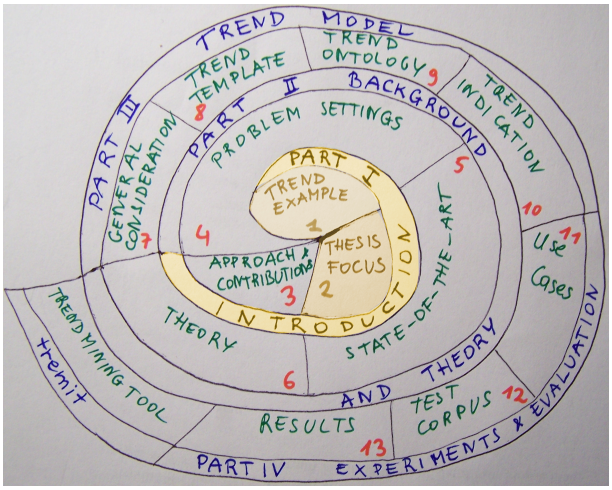
Figure 1.1 illustrates on a timeline selected reports from online available news on protests in North Africa in January and February, 2011. Anyone interested in political events worldwide noticed at some point in January 2011 that the

amount of news reporting on the situation in Egypt was growing rapidly and that the events in Egypt were growing in importance. One could read from different sources the reports of unrest in Egypt and there were different opinions on how the situation was developing and what would be the consequences of this development. Owing to the web-based forms of communication users participating in social networks and interested in politics could follow the events described in Twitter [Twitter, 2013] and Facebook [Facebook, 2013] posts directly [RNN-Users, 2013] from the people taking part in the unrest in Egypt. From January 25th to February 13th, 2011, Twitter was dominated by the trending hashtag [Cyger, 2011] #jan25. #jan25 was used as a tag in almost every message relating to the Egyptian revolution. However, #jan25 that referred to January 25, 2011, the "Day of Revolt" as named in the Wikipedia article [Wikipedia-Users, 2011a] in Egypt started to be used also in many breaking news reports on blogs and news sites (i.e. BBC[1], AlJazeera[2]). Immediately, a Wikipedia [Wiki-Community, 2013] article was written, explaining the chain of events related to the unrest taking place in Egypt starting from January 25th, 2011 [Wikipedia-Users, 2011b]. News articles, blogs, tweets[3], and posts come from different sources but are mostly texts written in a natural language. Most of them are publicly accessible on the web, emerging as a constantly growing, most important information and knowledge platform nowadays. The revolution in Egypt that took place from January 25th until February 13th, followed by the resignation of the former Egyptian president is, among other topics appearing in the same period, an example of a topic that increased in interest and political relevance. Considering the two-month period in the beginning of 2011, the Egyptian revolution is an example of a (political) trend.

[1] `http://www.bbc.co.uk/` online accessed on 03-March-2011

[2] `http://english.aljazeera.net/` online accessed on 30-March-2011

[3] A *tweet* is a post from the micro blogging service Twitter that was the most popular micro blogging service on the web in the beginning of 2011. Tweets are messages no longer than 140 characters and are published online.



CHAPTER **2**

Thesis focus

This chapter delivers an overview of the thesis content. It summarizes the research problem which is the focus of this work and the research questions which are being asked in the beginning of the thesis, and it describes how the answers to these questions will be structured throughout the thesis.

2.1 The problem and solution

This thesis focuses on the problem of mining trends in web documents. Regarding the important aspects of the problem – relevant theory, applicable approaches and algorithms – it seeks appropriate definitions of a 'trend' and 'trend mining'. It also investigates relevant approaches and explores how the integration of knowledge into a trend model influences its outcomes. It searches for the universal trend model that enables mining trends in web documents and helps making the trend mining results interpretable. It suggests the incorporation of knowledge and time as important dimensions for valuing text features in text analysis needed for trend mining.

Several methods have been proposed for the problem of topic detection and tracking (TDT) and of the emerging trend detection in texts (ETD) (see Chapter 5, Section 5.2.1 and 5.3). These approaches bring prototypical solutions for several general as well as specific problems related to trend mining (see Chapter 5, Section 5.2).

Although trend mining is of interest to a variety of researchers and is being used as term in some of current research projects, there is no single definition clarifying what it means scientifically. Regarding its specific characteristics, we propose to define trend mining in a similar way as data mining has been defined. The algorithms applied for detecting topics and tracking them over time mostly

use their own specific definitions of what the trend is. We suggest looking interdisciplinarily at the trend as a phenomenon, and formalizing characteristics of a trend in computer science terms. There are methods that allow for detecting trends by analysing topics over time, but their results are either very specific according to the test data applied in experiments or are difficult to interpret. We seek an universal trend model based on knowledge which allows for generating easily interpretable results. Very few algorithms are available for tests. Most of them are being developed on specific data sets, their implementations are not published online and there is no possibility to test them on different sets. We create the *tremit* – the trend mining tool – for tests based on both our own algorithms and the adapted machine learning algorithms.

2.2 Research questions

While doing this research, many questions regarding mining trends emerged from the beginning:

1. Can a knowledge-based trend model help us in understanding trends?
2. What do the trend and trend mining mean in terms of research in information retrieval and data mining?
3. Are there any research works relevant for trend mining?
4. Which algorithms are suited for mining trends?
5. Which representation models are appropriate for trend mining?
6. Is there a general trend model?
7. While mining trends, can we know in advance whether there is a trend in the data set?
8. Is trend mining a search or a knowledge discovery problem?

While each of these questions is relevant for this work and their answers each contribute to the final results, the first one stands out as the main research question. The purpose of this thesis is to deliver answers to these questions.

2.3 Content structure

The figures presented in the left corner at the beginning of every chapter aim to help the reader to follow the path of the thesis. The thesis is structured in four parts:

↪ Part I: Introduction

↪ Part II: Background and Theory

↪ Part III: Trend Model

↪ Part IV: Experiments and Evaluation

Each part builds on the previous one and the respective chapters, with the content structured as follows:

↪ Chapter 1: By providing an example of a trend, the first chapter explains the relationships between the web, the texts and the emerging topic.

↪ Chapter 2: The research of this thesis is summarized by the second chapter-sketching the problems and the proposed solutions, describing the research questions and giving an overview of the thesis contents.

↪ Chapter 3: This chapter outlines the approach we take in this research and summarizes the contributions being made by this research.

↪ Chapter 4: Here we discuss the general setting of our problem, giving an insight into different perspectives on the trend research. In this chapter we preliminarily define trend mining.

↪ Chapter 5: This chapter describes the state-of-the-art in relevant research fields, to help in answering the second and third questions for the relevant works and suited algorithms.

↪ Chapter 6: Relevant theory is introduced, and we investigate the possible representation models in trend mining.

↪ Chapter 7: In this chapter the general considerations for knowledge-based trend models are described.

↪ Chapter 8: Here we present the trend template – the universal trend model proposed by this research.

↪ Chapter 9: In this chapter, the trend ontology that is one of the possible trend template implementations is described.

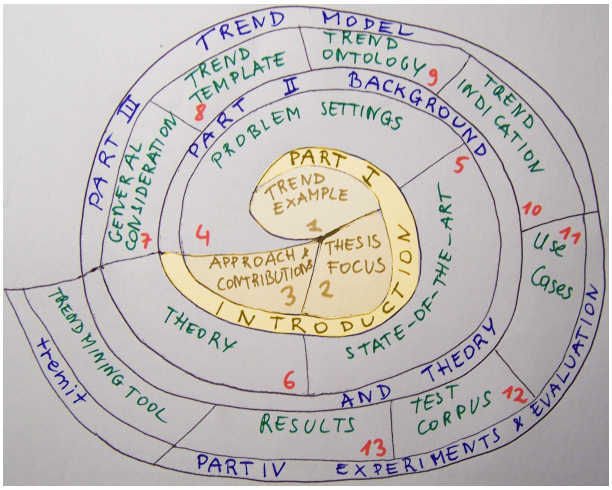
↪ Chapter 10 : This chapter is devoted to the trend indication weighting functions, which represent other possible directions of implementing the trend template.

↪ Chapter 11: In this chapter we describe the possible use cases that help us in understanding in which cases there is a need for trend mining.

↪ Chapter 12: Our test set, the text corpus, is described in this chapter.

↪ Chapter 13: Experiments and results are presented here. This chapter is the conclusive chapter that helps in answering the main research question which was posed beginning with Chapter 7.

This thesis closes with an outlook (14) and an appendix (A). The appendix contains the description of our test tool.



CHAPTER **3**

Approach and contributions

"The myth of methodology, in short form, is the belief that a playbook exists for innovation and (...) it removes the risk from the process of finding new ideas" (p. 37, [Berkun, 2009])

Knowing from Chapter 2 what was asked in the beginning of this research, now we learn how the answers to the questions were found in this thesis. This chapter provides insight into the approach taken while carrying out this research project. It also outlines the contributions of this research.

3.1 Approach

Focusing on the problem of mining trends, we go through the following steps in our research that are visualized in Figure 3.1:

1. **Problem settings (definitions):** First of all, we define the trend mining problem settings. Here we consider the possible perspectives on the trend as phenomenon, and choose the underlying definition of the trend from the Information Retrieval (IR) research. We determine the preliminary definition of trend mining and look at the problem of mining trends in web documents from the perspective of the current web, showing the nuts and bolts of this problem.

2. **Representation models:** Next, we focus on the possible representation models for trend mining while taking the IR view. While looking in general at the given representation models, we deal with trend mining as with a process of information retrieval under the time constraint. The different possibilities in IR – probabilistic approach by the example of topic models, vector space model and statistical approach by the example of clustering method – help

us to understand the expectations and limits of given algorithms that can be applied for mining trends. We focus in particular on the possibilities of a graph-based representation model using an ontology.

3. **Knowledge integration:** Continuing our thoughts on graph-based representation, we deliberate in general on the knowledge-based trend model using an ontology. In this step, we focus on a trend mining example from the market research. Based on our experience in the creation of a preliminary trend ontology, we sketch the general requirements on a knowledge-based trend mining model. Moreover, we start to distinguish the search problem perspective from the knowledge discovery perspective in trend mining.

4. **Model:** Based on the definitions assumed at the beginning of our approach and considering the general requirements on the knowledge-based trend model, we create the trend template that serves as the universal trend model. The idea of integrating knowledge into trend mining now takes two parallel directions – a knowledge-based direction and a knowledge-integrating direction that are realized in the two different implementations of the trend template – the trend ontology and the trend-indicating functions.

5. **Evaluation (Feedback):** We set the evaluation frame in which we experiment with the different algorithms – the LDA-based topic models, the k-means clustering method, and the trend template comparing the results gained on the same test set of documents. In this step we use the experiment's results as feedback for our approach.

3.2 Scientific method

While Figure 3.1 from the previous section shows how we approach the problem of mining trends and how we find the answer to our research question, in Figure 3.2 the general scientific method is presented. "The scientific method is the logical scheme used by scientists searching for answers to the questions posed within science, as well to formulate theories as to assure the means for producing them (instruments, tools, algorithms)."[Dodig-Crnkovic, 2002] Different scientific methods can be considered for research in computer science. A good, deep going discussion about the science and its method is given in [Chalmers, 1999] and another discussion about the scientific methods in computer science by [Dodig-Crnkovic, 2002].

We choose the general scientific method [Schumm, 1991][MIT, 2011], based on six steps: question, observation, hypothesis, experiments, data and conclusion. Figure 3.3 shows how these steps are realized towards the content of this thesis.

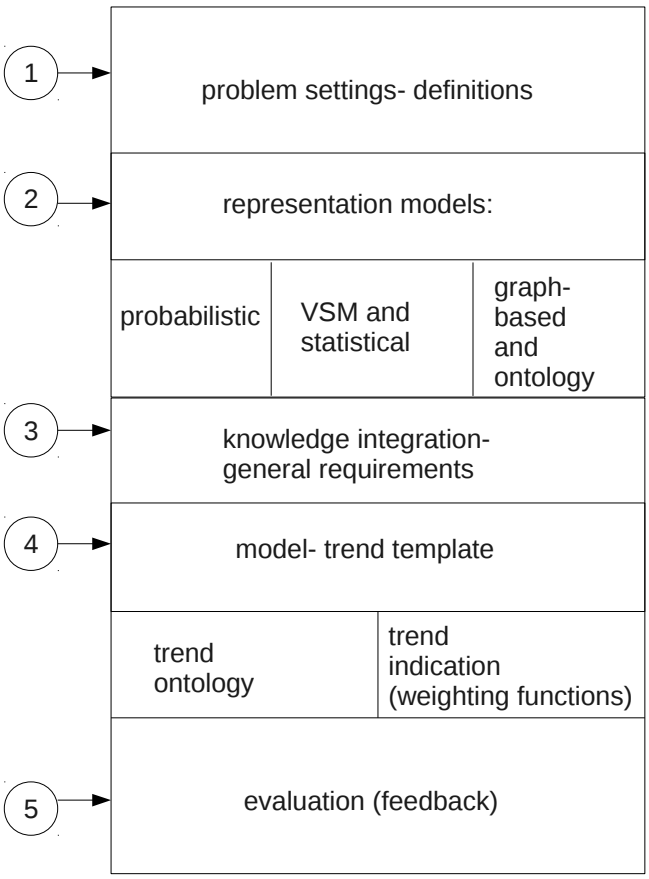


Figure 3.1: Solution approach. Source: *author*.

Step 1, the *question*, is covered by the introductory part of this thesis, where we pose our research questions. The second step, the *observation*, is described in the second part of the thesis, Background and Theory II, which includes the problem settings (Chapter 4), the state-of-the-art (Chapter 5), and the theory (Chapter 6). The third step, the *hypothesis*, is realized in the Trend Model Part III of the thesis – there we present our knowledge-based trend model, which should help in generating interpretable trend mining results. The last remaining steps – *experiments*, *data*, and *conclusion* – are contained in the Experiments and Evaluation, which is Part IV of this research.

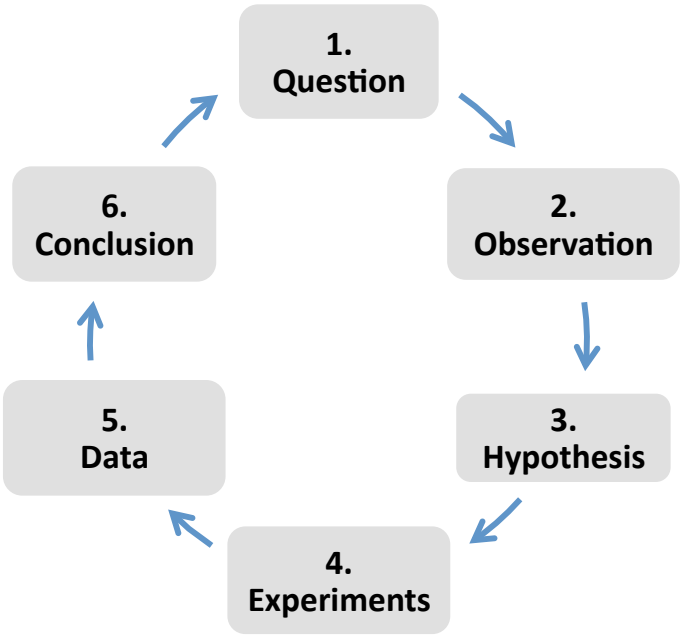


Figure 3.2: The scientific method based on six steps. Source: *author*.

3.3 Contributions

The main contributions of this work lie in:

- the definitions of the trend and trend mining
- the trend template
- the implementations of the trend template: trend ontology and trend indication
- the test corpus based on the (DAX[1]) German Stock Exchange news
- the trend mining test tool – tremit

The research content included in this thesis has been published in the papers (doctoral consortia or symposia, workshop, conference and technical reports) listed as follows (sorted by publication date):

1. Olga Streibel, *Semantic-based Learning Method for Trend Recognition in Simple Hybrid Information Systems*, Conference on Advanced Information Systems CAiSE2008, Proceedings of Doctoral Consortium, pages 106–113, Montpellier, France, June 2008.

[1] http://www.finanzen.net/index/DAX online accessed 17-February-2013

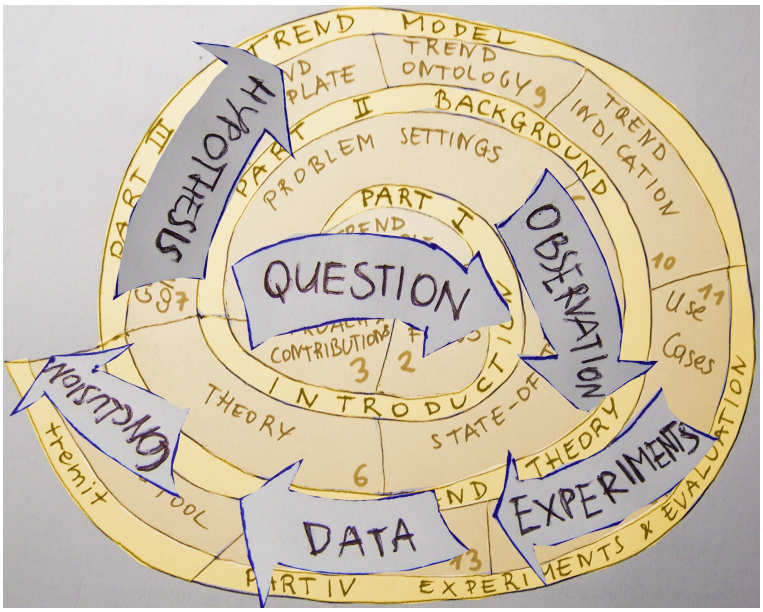


Figure 3.3: The scientific method within the thesis structure. Source: *author*.

2. Olga Streibel, *Trend Mining with Semantic-Based Learning*, European Semantic Web Conference ESWC2008, Proceedings of PhD Symposium, pages 71–72, CEUR. Vol-358, Teneriffe, Spain, June 2008.

3. Olga Streibel, *Semantic learning for Trend Recognition in Text Collections*, pages 8–9, CEUR. Vol-500, Proceedings of STI PhD Seminar, Berlin, September 2009.

4. Olga Streibel and Malgorzata Mochol, *Trend ontology for knowledge-based Trend Mining in textual Information* In IEEE Computer Society Proceedings of 7th International Conference on Information Technology : New Generations, ITNG2010, pages 1285–1288, Las Vegas, U.S., April 2010.

5. Olga Streibel, *Mining Trends in Texts on the Web*, Proceedings of the Doctoral Consortium at Future Internet Symposium 2010 FIS2010, pages 80–90, CEUR. Vol-623, Berlin, Germany, September 2010.

6. Olga Streibel and Rehab Alnemr, *Trend-based and Reputation-versed Personal News Network*, Proceedings of the 3rd International Workshop on Search and Mining User Generated Content SMUC2011 at 20th ACM Conference on Information and Knowledge Management CIKM2011, pages 3–10, Glasgow, U.K., October 2011.

7. Olga Streibel, Lars Wißler, Robert Tolksdorf, Danilo Montesi, *Trend template: mining trends with a semi-formal trend model*, Proceedings of the

3rd International Workshop on Ubiquitous Data Mining UDM2013 in conjunction with 23rd International Joint Conference on Artificial Intelligence IJCAI2013, pages 49–53, Beijing, China, August 2013.

8. Olga Streibel, Alexa Schlegel, Robert Tolksdorf, *German Finance Text Corpus: Description of the German Finance Text Corpus*, corpus data publication at Linguistic Data Consortium (LDC), 2013 *to appear*.

Under the supervision of Prof. Robert Tolksdorf and Prof. Adrian Paschke, from 2010 to 2013 I was an advisor of master and bachelor theses for the following topics:

1. "GUI for knowledge-based Trend Analysis", Bachelor thesis. Author: Diana Olivera Viscarra, 2013.

2. "Texts, Trends, and the Web", Master thesis. Author: Ievgeniia Ozeran, 2011.

3. "Ontologies for Knowledge-based Trend Analysis", Bachelor thesis. Author: Lars Wißler, 2011.

4. "Preprocessing of Documents for Emergent Trend Detection in Text Collections", Master thesis. Author: Iavor Jelev, 2010. (Supervising in cooperation with Institute for Business Mathematics Fraunhofer, Kaiserslautern, Germany).

Discussions with my students as well as their research on the advised topics contributed to this thesis. Some of the content related to this thesis has been preliminary published in the following technical reports:

1. Gökhan Coskun, Ralf Heese, Markus Luczak-Rösch, Radoslaw Oldakowski, Ralph Schäfermeier, and Olga Streibel. Towards corporate semantic web: Requirements and use cases. Freie Universität Berlin, 2008, Technical Report TR-B-08-09, pages 50–56.

2. Gökhan Coskun, Marko Harasic, Ralf Heese, Markus Luczak-Rösch, Radoslaw Oldakowski, Adrian Paschke, Ralph Schäfermeier, and Olga Streibel. Realizing the corporate semantic web: Concept paper. Freie Universität Berlin, 2009, Technical Report TR-B-09-05, pages 26–29 and 31–33.

3. Gökhan Coskun, Marko Harasic, Ralf Heese, Markus Luczak-Rösch, Radoslaw Oldakowski, Adrian Paschke, Ralph Schäfermeier, and Olga Streibel. Realizing the corporate semantic web: Prototypical implementations. Freie Universität Berlin, 2010, Technical Report TR-B-10-05, pages 29–32.

4. Gökhan Coskun, Marko Harasic, Ralf Heese, Markus Luczak-Rösch, Radoslaw Oldakowski, Adrian Paschke, Ralph Schäfermeier, and Olga Streibel. State of the Art Analysis – Working Packages in Phase II. Freie Universität Berlin, 2011, Technical Report TR-B-11-07, pages 19–22.

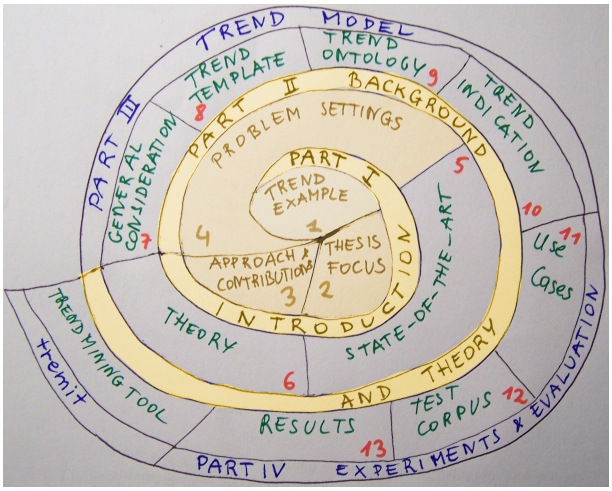
5. Gökhan Coskun, Marko Harasic, Ralf Heese, Markus Luczak-Rösch, Radoslaw Oldakowski, Adrian Paschke, Ralph Schäfermeier, and Olga Streibel. Prototypical Implementations – Working Packages in Project Phase II. Freie Universität Berlin, 2012, Technical Report TR-B-12-04, pages 22–23.

6. Lars Wißler and Olga Streibel. Ontologien im Trend Mining. Freie Universität Berlin, 2012, Technical Report TR-B-12-07.

In the winter term of 2012-2013 I had the opportunity to offer a trend mining seminar for master students at the Freie Universität Berlin. The content of the seminar can be found online at: `https://sites.google.com/site/seminartrendmining`. I would like to mention that this research has been developed during the time I was working in the projects: TREMA (Trends: Mining and Fusion of Multimodal Data) and CSW (Corporate Semantic Web). This doctoral thesis is partially funded by the Investitionsbank Berlin and by the German Federal Ministry of Education and Research (BMBF) and the BMBF Innovation Initiative for the New German Länder - Entrepreneurial Regions.

Part II

Background and Theory



CHAPTER **4**

Problem settings

In Chapter 4, we learn about the general problem settings, starting the search for answers to the questions: What is a trend? What do the trend and trend mining mean in terms of research in information retrieval and data mining? We show a trend from different perspectives. The description of different possible perspectives on the example presents different possible contexts of a research related to trend mining and introduces the information retrieval (and knowledge discovery) point of view on the trend mining problem.We dive into relevant research works searching for the trend definitions and extract the definitions needed for our research. Since the web is an important setting in our research scenario, we close this chapter by discussing the increasing complexity of our research problem while considering the web.

4.1 Different perspectives on a trend

The case of the North African uprising in 2011 presented in Chapter 1 is an example of a trend as well as an example of a sociological phenomenon – a crowd-based movement. The web is currently the biggest open communication platform for users, enabling rapid emergence of crowd-based happenings [Surowiecki, 2004] [Maier, 2008] or movements that can lead to a trend. However, the most instructive way to understand a trend is to look at a trend curve or a diagram and to relate it to the mathematical definitions. For these reasons we look to the fields of sociology and mathematics, and then focus on information retrieval and data mining in order to find their definitions of a trend.

4.1.1 Sociology

There are very few scientific publications which examine trends as a phenomenon in sociology[1]. The most precise one allows for understanding the "anatomy of a trend" [Vejlgaard, 2008]. Detecting trends from the sociological point of view is an analytical method for observing changes in people's behavior over time with regard to "six attitudes towards trends"(p. 30, [Vejlgaard, 2008]). The definition of these six attitudes is based on eight different personality profiles of people who participate in the trend process: *trend creators, trend setters, trend followers, early mainstreamers, mainstreamers, late mainstreamers, conservatives and anti-innovators.* The author visualizes the trend phenomenon itself as a *diamond-shaped trend model* (p. 64, [Vejlgaard, 2008]) with regard to these groups (see Figure 4.1). The source of a trend is always trend creators, and the trend closes with the anti-innovators. Trend setters are hereby "the most open and curious individuals with regard to style and taste" (p. 71, [Vejlgaard, 2008]) whereas the conservatives "prefer styles that have existed for years or even decades. They are the people who are the most skeptical of new styles"(p. 72, [Vejlgaard, 2008]).

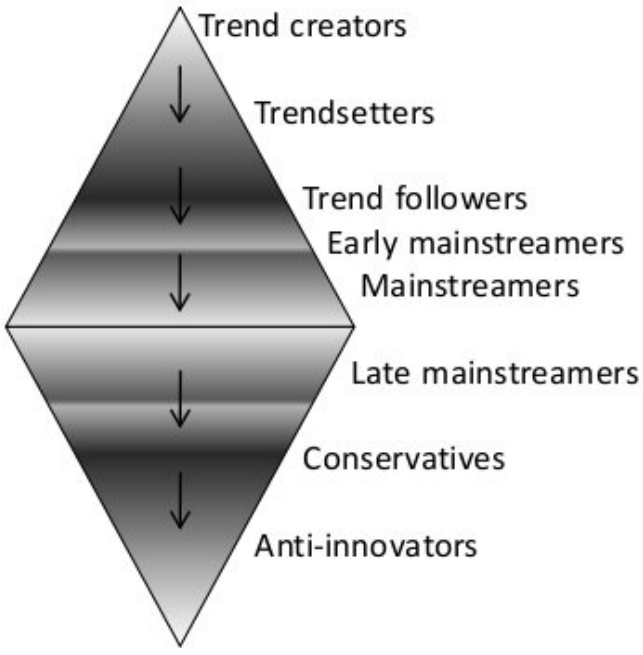


Figure 4.1: Diamond shaped trend model. Source: *[Vejlgaard, 2008].*

Trends from the sociological perspective are certain patterns of people's behavior and lifestyle that have evolved over a focused time interval; the word

[1] as for 2010, and to our best knowledge

trend refers to a process of change. A sociologist's perspective of the North African political trend in the beginning of 2011 would provide an analysis of the political changes based on observations of the different groups involved in the trend: young protesters who most probably became the trend setters while starting the unrest by stating their fresh political view contrary to the people supporting the old political system (most probably the conservatives in the trend process). The sociological perspective sheds light onto the sociological processes involved in the change and brings a deep understanding of the trend itself.

4.1.2 Statistics

Detecting trends from the perspective of statistics is based on analysis of time-series data. There are two goals of this analysis (p. 490, [Han and Kamber, 2006]):

- *modeling time series* (i.e. to gain insight into the mechanisms or underlying forces that generate the time series)
- *forecasting time series* (i.e., to predict the future values of the time-series variables)

Time series, built on real-value measurements, are the observation sequences of a particular phenomenon, such as observations of stock exchange price changes. They can be univariate or multivariate. Other features of time series are: stationary and non-stationary (p. 22, [Mitsa, 2010]). "A stationary time series has a mean and a variance and is not changing over time" (p. 22, [Mitsa, 2010]). A non-stationary time series has no mean and increases or decreases over time. Once a time series has been modeled, two characteristics are interesting for the time series analysis: the trend and the periodicity. The trend analysis process consists of four major components (p. 490-491, [Han and Kamber, 2006]):

1. trend or long-term movements
2. cyclic movements or cyclic variations
3. seasonal movements or seasonal variations
4. and irregular or random movements

A trend in this context is an indicator for a change in the data mean [Mitsa, 2010]. The simple features of time series are: mean, median, mode, and variance (p. 47, [Mitsa, 2010]). The definitions of mean and variance are presented as follows (p. 47, [Mitsa, 2010]):

- The mean "shows the average value of the time series values". For a time series with $X = x_1, x_{2,\ldots,x_n}$ with N values:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \tag{4.1}$$

- The variance "shows the amount of the variation of the time series values around the mean". It is also known as a *second moment*:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N - 1} \tag{4.2}$$

Possible characteristics of time series are: serial correlation, skewness, kurtosis, non-linearity, self-similarity, and chaos. The definitions of skewness and kurtosis are presented as follows [Wang et al., 2006]:

- The skewness measures the asymmetry of the histogram's shape. For the univariate data with the mean Y_t, the standard deviation σ, and n number of data points:

$$S = \frac{1}{n\sigma^3} \sum_{t=1}^{n} (y_t - \overline{y}_t)^3 \tag{4.3}$$

- The kurtosis, also called *heavy tails*, is a measure of the curve's peakedness. For a univariate data with the mean Y_t, the standard deviation σ, and n number of data points:

$$K = \frac{1}{n\sigma^4} \sum_{t=1}^{n} (y_t - \overline{y}_t)^4 \tag{4.4}$$

The self-similarity can be measured by so called Hurst exponent [Mitsa, 2010] [Willinger et al., 1996], and the chaotic behavior by the *Lyapunov* exponent [Mitsa, 2010] [Wolf et al., 1985]. Regarding the example in Figure 4.3 from a statistical perspective, we apply a function to create a time series out of the selected reports. As shown in Figure 4.2, we visualize the function graph that, based on the number of news items appearing in the given time, can be further explored for trend.

While exploring, if there is an upward trend in the news curve, we can calculate the upward trend as follows:
x_t represents the number of articles at the time point $t \in N$, whereas N represents the set of time points that are given in a particular time measure (i.e. an hour, a day, a month), plotted against the update of the real-value data. τ is a particular defined time interval (i.e. three hours if the time measure is one hour or two days if the time measure is based on days). Y_t is hereby calculated as follows:

Figure 4.2: Example of trend estimation from numeric curve. Source: *[Jelev, 2010]*

.

$Y_t := x_{t+\tau} - x_t$. An upward (also called positive) trend in the time interval $[t_1, t_{1+\tau}]$ is a trend that fulfills the following rule [Jelev, 2010]:

$$\frac{\#t : Y_{t_1} < Y_t}{N - \tau < K\%} \tag{4.5}$$

$K\%$ is a percentage value that can be arbitrarily chosen, in relation to the detected amplitudes in the curve progression.

In general, the statistical perspective focuses on the value-time relation of a trend. It provides analysis techniques that bring insight into the trend progress and helps to predict trends based on real values, without necessitating a deep understanding of the trend's background processes.

4.1.3 Information retrieval

Taking the computer science perspective on trend mining leads mainly into the information retrieval research. In particular, into the research on Emerging Trend Detection (ETD) summarized in [Kontostathis et al., 2003] that set the basic frame for most of the research work on trend mining and either for this thesis.

According to [Kontostathis et al., 2003], detecting trends from text collections refers to the *detection of emerging topics in texts*. In terms of ETD a *trend* in texts is defined as *"a topic area that is growing in interest and utility over time"* [Kontostathis et al., 2003].

Closely related to the emerging trend detection is the event detection research that provides methods for event monitoring and event or anomaly detection. The research on Topic Detection and Tracking (TDT) [Allan, 2002] summarized

under event-based information organization, partially applies approaches from event detection research. TDT provides a definition of *topic*, which is "a set of news stories that are strongly related by some seminal real world event" (p. 2, [Allan, 2002]).

From the information retrieval perspective, the initial example shows a topic: *Egyptian revolution* that increased in interest and utility from January to February, 2011. This topic was covered by a set of news stories in the form of tweets, Facebook status messages, and blog entries, that referred to the real world event which was the unrest in Egypt. However, many other similar trends may have emerged at the same time but not with the same visible impact as the Northafrican uprising. In order to discover and analyze these emerging topics, methods, algorithms and tools are necessary. The objective of detecting trends is "to provide an alert that new developments are happening in a specific area of interest in an automated way" [Kontostathis et al., 2003].

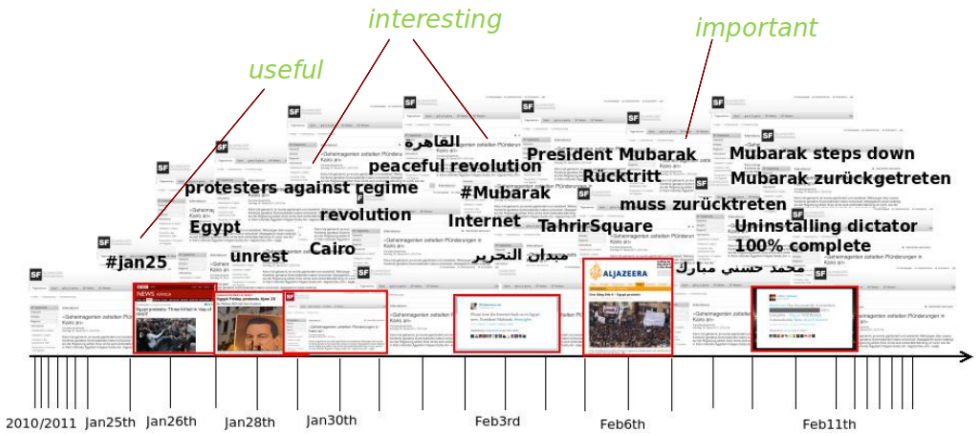


Figure 4.3: Interesting, useful and important keywords in reports. Source: *author*.

4.2 Definitions

In the following subsections we introduce definitions that are essential for further reading of this thesis. Starting with the common definition of a *trend* in 4.2.1, the examples of the diversity in defining trend in relevant research are introduced in 4.2.2. Based on the introductory definitions, we specify in 4.2.3 the definition of a *trend* and *trend mining* in the context of our research.

4.2.1 A common-sense trend definition

trend (Oxford dictionary [Oxford, 2013])

noun

1. a general direction in which something is developing or changing:
   an upward trend in sales and profit margins

2. a fashion:
   the latest trends in modern dance

verb
(no object, with adverbial of direction)

1. change or develop in a general direction:
   unemployment has been trending upwards

   - (of a topic) be or become popular on a social networking site:
     I've just taken a quick look at what's trending on Twitter right now

2. (especially of a geographical feature) bend or turn away in a specified direction: *the Richelieu River trending southward to Lake Champlain*

4.2.2 Trend definition diversity in the relevant research

Trend has been defined in a variety of ways in recent research. A sample of this definition diversity is presented below.

[Engel et al., 2010] define an emerging trend as "a change in topic for an extended period of time as illustrated by the jump discontinuity or the slope discontinuity". The *jump* and *slope discontinuity* correspond to an abrupt change in the content of the text stream (jump) or to a ramping up (or down) in a topic for that text stream (slope). However, in providing the definition of an emerging trend, the authors refer also to the definition of an event. They refer to "the instantaneous discontinuity types (point or jump) as a *surprise* event".

[Naaman et al., 2011] refers to *significant events* and *temporal trends* while focusing on users' interests and the events reflected in so called *social awareness streams*. They state that "trends may reflect a varied set of occurrences, including local events, global news events, televised events Internet-only and platform specific memes, and hot topics of discussion".

[Morinaga and Yamanishi, 2004] address the problem of discovering topic trends and analyzing their dynamics in real-time. In particular they define topic as "a seminal event or activity".

[Kawamae and Higashinaka, 2010] frame that "each trend can be presented as a mixture of topics and localization over time" whereas in [Kawamae, 2011] the author refines this definition by assumption that "each trend can be presented as a mixture of temporal words, terminology words, and localization over time".

[Goorha and Ungar, 2010] generalize the definition of trend detection proposing the use of term discovery that they frame as "automatic identification of

emerging topics associated with products of interests", and the research presented in [Mathioudakis and Koudas, 2010] refers simply to *emerging topics* while using the word trend.

This variety of refinements for *trend* and *trend detection* definitions in the relevant publications is characteristic of the trend mining research. Our present hypothesis is that there is no consensus in the trend mining research about what precisely a trend is. The lack of consensus about trend is comprehensible since the trend mining is a multifaceted problem: it depends on the frequency of trend (i.e. short-term, long-term), on the art of data in which trend occurs (i.e. real valued, textual), on the fact if trend has been directly triggered by an event or not. Trend definition depends often on the application scenarios for which given trend approaches are being developed. In every case it describes the occurring *change* in data. On the other hand, several (i.e. [Engel et al., 2010] [Goorha and Ungar, 2010]) works on trend mining still refer to trend definition in [Kontostathis et al., 2003], where trend is in general "a topic area that is growing in interest and utility over time". This general trend definition inspired our research.

4.2.3 Definitions proposed in this thesis

With regard to [Kontostathis et al., 2003] we propose the following definition of a *trend*:

Definition 4.2.1. ***Trend (in texts)***
is a topic area that is growing in interest and utility over time.

As in [Witten and Eibe, 2005]: data mining "is the extraction of implicit, previously unknown, and potentially useful information from data". Regarding it, we propose the following definition a trend mining:

Definition 4.2.2. ***Trend mining***
is the extraction of implicit, previously unknown and potentially useful information from time-ordered texts or data, i.e. the information that financial and economic crisis is emerging. Trend mining techniques can be used for capturing a trend in order to support users in providing previously unknown information and knowledge about the general or specific development in users' field of interests in a given time frame, i.e. information about the forthcoming market movements

The main reason for introducing the trend mining definition is the intention of using one standard term for many research works on similar or the same subject. Methods on *trend analysis*, *trending analysis*, *emergent trend detection*, *topic detection* and *tracking*, *tracking changes in topics*, follow the same common idea: to capture general, temporal changes (in text or data) in order to support user in providing previously unknown information and knowledge about the general development in users' field of interests. Doing so, they extract the implicit, previously unknown and potentially useful information from time-ordered text or data.

4.3 Stepping up on complexity

The web has become the main source of textual information for many people who are willing to learn about the major world events. Web users, while collaborating over social networks, blogs and micro-blogging services also contribute to news coverage worldwide.

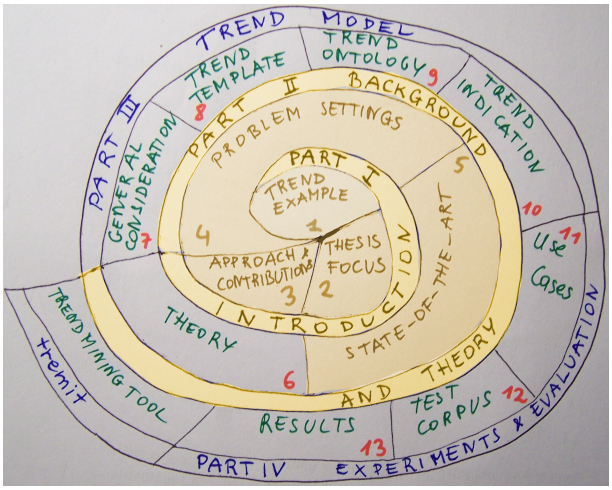
News feeds come from mainstream media as well as social networks. Sometimes feeds from these social networks are more up-to-date than those that come from mainstream media. But the overwhelming amount of information requires a user to personally filter through it until one gets what is really needed.

Social networks like Delicious[2], Diaspora[3], Facebook[4], Flickr[5], LinkedIn[6], Twitter[7], Xing[8], YouTube[9] have become very popular among users on the web. In recent years, Facebook attracted hundred of millions of users worldwide, increasing its membership from over 100 million in 2009 to over 500 million in 2011[10]. Around 175 million[11] web users in 2010 had a Twitter account. Everyday there are 95 million[12] tweets worldwide and "more than 30 billion pieces of content (web links, news stories, blog posts, notes, photo albums) each month" shared on Facebook[13]. Owing to these novel forms of communication, anyone with an internet device could follow the developments during the flood in Rockhampton in Australia in 2010 and 2011 since residents of this town created a public Facebook group reporting in real-time about the flood [14].

[2] http://delicious.com/ accessed 08-Nov-2011
[3] http://joindiaspora.com accessed 08-Nov-2011
[4] http://www.facebook.com accessed 08-Nov-2011
[5] http://www.flickr.com accessed 08-Nov-2011
[6] http://www.linkedin.com accessed 08-Nov-2011
[7] http://www.twitter.com accessed 08-Nov-2011
[8] http://www.xing.com accessed 08-Nov-2011
[9] http://www.youtube.com/ accessed 08-Nov-2011
[10] http://www.facebook.com/press/info.php?factsheet accessed 30-March-2011
[11] http://twitter.com/about accessed 30-March-2011
[12] as for September 2010
[13] http://www.facebook.com/press/info.php?statistics accessed 30-March-2011
[14] http://tinyurl.com/on2k3lj accessed 30-March-2011

In mainstream media, the political events in Iran in 2009 have been described as the Twitter-Revolution[15] since many people communicated about these events using the microblogging service Twitter. Furthermore, the political developments and revolutions in North Africa beginning in January 2011 could be followed on Facebook, Twitter, Flickr, Bambuser, and others. Public Facebook status updates, tweets, bookmarks, and pictures represent immediate knowledge about our world, generated by web users. Among this content, many trends emerge in real time.

[15] http://www.washingtontimes.com/news/2009/jun/16/irans-twitter-revolution/ accessed 30-March-2011



CHAPTER 5

State of the art

Having learned the problem and its overall settings, in this chapter we take a deeper look at the relevant works and tools in the area of trend mining. We start with research areas and review chosen approaches while classifying them into research directions. The discussion about relevant works continues with the description of relevant tools, including offline analysis and web tools. The literature and software review leads to the list of problems related to trend mining research. Finally, a problems summary allows for more understanding of the obstacles to trend mining research and explains how this thesis is positioned into the state-of-the-art works in this field.

5.1 Research areas

In Section 4.1 we narrowed the computer science perspective on trends to the information retrieval research. In the Section 4.2.3 of Chapter 4 we focused our definition on data mining, and in the further text we consider the knowledge discovery perspective on mining trends. Indeed, approaches to trend mining in the literature are often classified into different research areas. This may lead to confusion. In general, when focusing on mining trends from textual data, the following three research areas should be mentioned:

1. emergent trend detection: a sub-area of information retrieval, related to knowledge discovery and text mining
2. topic detection and tracking summarized as event-based information organization: an area related to information retrieval with text, and data mining components applying event detection approaches
3. temporal data mining

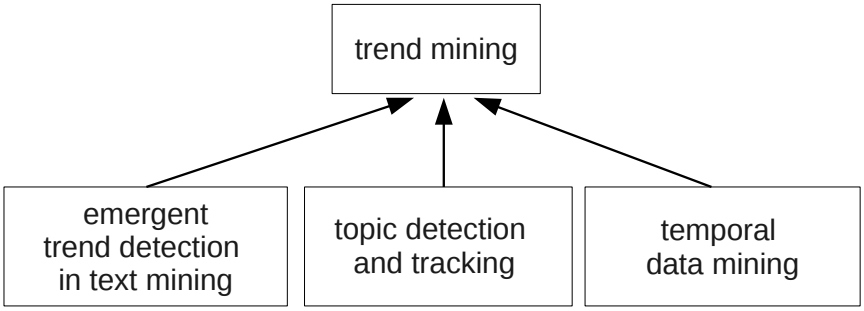


Figure 5.1: Trend mining present in the relevant research fields. Source: *author*.

How these research fields constitute the trend mining is shown in Figure 5.1.

The first research field mentioned above, the emerging trend detection research, is the most relevant for our work. In [Kontostathis et al., 2003] several systems that detect emerging trends in textual data are presented. These so called ETD systems are classified in two main categories: semi-automatic and fully-automatic. For each system there is a characterization based on the following aspects: *input data and attributes*, *learning algorithms* and *visualization*. This comparison includes an overview of the relevant research and projects published in [Allan et al., 1998] [Lent et al., 1997] [Agrawal et al., 1995] [Swan and Jensen, 2000] [Swan and Allan, 1999] [Watts et al., 1997]. However, no sharp distinction has been made between the TDT and ETD research fields, which means that many works, i.e. [Swan and Allan, 1999] or [Lavrenko et al., 2000] can be in fact classified into both fields. The characteristics of a given approach and the research direction lets us classify the given works more into the TDT (event-based field) or more into the ETD (general trend mining).

The second field, topic detection and tracking research, is predominantly related to event-based approaches. Event-based approaches for trend mining rely on the assumption that trends are always triggered by an event, that is often defined as "something happening" or "something taking place"(page 102, [Lita Lundquist, 2000]). Considering a trend from the event research perspective means that trend detection has to be understood as a monitoring task. This is mostly the case for so called short-term trends that are indeed triggered by some

events and in order to detect them we have to monitor the stream in which they occur, i.e. the occurrence of "Eyjafjallajökull eruption"[1] which actually occurred in Iceland and was reported in social networks and on the news in March 2010. However, so called long-term trends, i.e. "financial crisis" that started to be on-topic in 2008 are not necessarily conjoined with one specific event. It is more a chain of events or even the "soft" indicators in public opinion or news:

Askitas N., Zimmermann K.F Wochenbericht des DIW Berlin Nr. 25/2009:

"In der gegenwärtigen Wirtschaftskrise haben sich Prognosen als besonders schwierig erwiesen. Dies ist ein weltweites Phänomen. In immer kürzerer Folge kam es zu Prognoserevisionen, die letztlich in einem Herdenverhalten der Prognostiker endete – ein typisches Zeichen für mangelnde Informationen im Markt. Dies hatte zunächst mit der Geschwindigkeit zu tun, mit der sich in der globalisierten Welt die negativen Impulse, die aus einem Zusammenspiel einer zyklischen Abschwächung der Weltkonjunktur und dramatischer Krisensignale aus den Finanzmärkten entstanden, über die ganze Welt verbreiteten und insbesondere das Investitionsklima eintrübten. Mit dieser Geschwindigkeit war die traditionelle Konjunkturforschung und die amtliche Statistik überfordert, da sich die Anpassungspro-in Tagen oder Wochen vollzogen, und nicht wie sonst üblich in Monaten oder Quartalen. Deshalb wurde noch mehr als sonst auf "weiche" Indikatorsysteme wie Stimmungsumfragen und Handelsindizes zurückgegriffen. Auch wenn sie wenig über die weitere Zukunft sagen können, so geben sie doch in normalen Zeiten ein robustes Bild über die Lage, in der sich die Wirtschaft befindet"

"Forecasting proved to be particularly difficult during the current economic crisis. This was a global phenomenon. There were adjustments of previously given forecasts which succeeded more and more rapidly one after the other. Finally the forecasters ended up behaving herd-like – a typical sign of the lack of information on the market. This happened in the first place because of the very quick spread of the negative impulses in the globalized world, as they emerged from the interaction between the cyclical weakening of the global economy and dramatic signs of crisis from the financial markets, clouding over the investment climate especially. The processes of adaptation took place in days or weeks and not, as was commonly the case, in months or quarters and such a speed just overtaxed the traditional economic research and official statistics. And that is why the forecasters fell back even more than usual on "soft" indicator systems like sentiment surveys and trade indices. Even if they do not allow them to talk of the far future, they do give them quite a solid image of the current economic situation during normal times."[2]

[1] The eruption an Icelandic volcano in March 2010 that caused air travel chaos in Europe and revenue lost for the airlines

[2] Many thanks to Piotr Majchrzyk for the translation from German to English

Nevertheless, the various definition of a trend as used in different research approaches and presented in Section 4.2.2 often include a definition of an event.

The third research field, temporal data mining research [Mitsa, 2010], offers methods for clustering, classification, dimension reduction and processing of time-series data [Wang et al., 2005]. It addresses in general the temporal data and the techniques of time series analysis on these data. An example definition of temporal data is "time series data which consist of real valued sampled at regular time intervals" [Mitsa, 2010]. Temporal data mining is an interesting research area with a huge relevance to trend mining, but minor relevance to our approach since the focus of temporal data mining lies generally in real valued data – not on textual data as it is our case. On the other hand, temporal data mining applies the data mining methodology and deals with the same approaches, such as classification or clustering, that are relevant also for mining trends in textual data. Furthermore, it considers the idea of *temporal ontologies* (page 12, [Mitsa, 2010]) that is conceptually relevant for our approach on trend ontology (see Chapter 7, Section 7.1). While we are not discussing in detail most common approaches in temporal data mining, we will refer as needed to some of them in the following chapters of this thesis.

5.2 Directions in trend mining research

Before we continue with the discussion about the methods applied in trend mining, let us have a deeper look at the different possible refinements of the trend definition in context of the information retrieval, knowledge discovery and data mining research. Based on these refinements, we can examine the possible research directions and move more towards the characteristics of general trend mining approaches and emergent trend detection algorithms. The definitions mentioned in Section 4.2.2 sketch a general characteristic of the overall directions in trend mining research.

In general terms, the three main directions of trend mining research can be classified as: 1) general trend models including approaches from ETD, 2) event-based approaches including the approaches from TDT, and 3) time-series analysis approaches. Certainly, the three directions can be split into different sub-directions. For example, as for the general trend models it would be different characteristics based on input data or applied algorithms, and for the event-based approaches it would be different tasks addressed in the TDT research.

As already mentioned above, the difference between 1) and 2) is not sharp and many approaches could be classified in both directions. If a given approach focuses more on an augmented definition of an event and concentrates on the monitoring of news/text streams while searching for events, it would be classified

more into the first direction. In case a given approach concentrates on the time-based information processing, takes into account events as an existing but not a necessary trigger for a trend, and deals more with an overall trend mining system or the general algorithms for trend mining, it should be classified in the second direction.

While assigning the approaches into these directions, our goal is to review the trend mining problems that the researchers are focusing on, providing an overview of underlying algorithms and data used for tests.

5.2.1 General models

Some of the general trend mining approaches have been designed as complete trend detection systems, including visualization of the results. Their core components are the machine learning algorithms or a combination of different analysis methods with a learning (classification or clustering) approach. Probably the most important one for the trend mining research is research work done by [Lavrenko et al., 2000] where automatic trend detection based on text has been proved feasible and useful. In their paper about Ænalyst [Lavrenko et al., 2000], the authors describe "a system for predicting trends in stock prices based on the content of news stories that precede trends". Ænalyst has been designed and implemented as a general architecture for the association of news stories with trends. The system collects hybrid data: financial time series and time-stamped news stories, re-describes time series data into"high-level features", called trends, and aligns then each trend with time-stamped news stories. Such news stories serve as training sets for learning the language model which determines the statistics of word usage patterns in the stories. This language model, learned for every trend type, helps to monitor a stream of new incoming news stories. The model processes new news stories due to the learned hypothesis. Authors define here the task of trend detection as a special case of *activity monitoring* as introduced by [Fawcett and Provost, 1999]. The approach has been evaluated on 127 stocks set with resulting news collection of over 38,000 news articles from October 1999 to February 2000.

[Pottenger and Yang, 2001] propose an approach for detecting emerging trends in conceptual content. The detection process is "analogous to the operation of the radar system" in the sense that techniques applied for the detection enable the identification of "regions of semantic locality in a set of collections and screen out topic areas that are stationary in a semantic sense with respect to time". Based on the approach presented in [Pottenger and Yang, 2001], a general trend mining system referred to as Hierarchical Distributed Dynamic Indexing ($HDDI^{TM}$) is created and described in [Pottenger et al., 2001]. The system relies on unsupervised and supervised learning approaches. The detection of emerging trends in text collections is based on semantically determined clusters of terms.

Using linguistic feature extraction, clusters based on semantic similarity of these features are created. At multiple points in time, the algorithm takes a snapshot of the statistical state of the collection. Finally, a neural network that uses as input the rate of change in the size of clusters and in the frequency and association of features classifies topics as emerging or non-emerging. The evaluation is conducted on two test sets from commercial patent database[3].

TimeMines is a system that automatically generates timelines from date-tagged free text corpora [Swan and Jensen, 2000]. Two separate machine learning aspects are presented in TimeMines [Kontostathis et al., 2003]. One is the extraction of the most significant features from the input documents. This is done by the use of a simple statistical model for the frequency of occurrence of features in a stream of text as described in [Swan and Allan, 1999]. The model is based on hypothesis testing, choosing the most relevant features from the date-tagged texts while allowing a reduction of the features set. Another learning method, again based on hypothesis testing, groups the features from the reduced feature set based on the similarity in their distribution within a given time period [Kontostathis et al., 2003]. The system has been evaluated on a chosen subset of the TDT corpora[4] using 6,683 stories over 175 days (January 7th - June 30th 1995). TOA System is one of the early systems for technology opportunity analysis. It is a semi-automatic trend detection system which enables mining of text files using bibliometrics [Watts et al., 1997]. TOA relies on the expertise of the user who is researching a given area. There are no inherent learning algorithms present in the system [Kontostathis et al., 2003]. TOA and a few more of the general ETD systems concerned with the detection of trends are summarized in [Kontostathis et al., 2003] based on the following characteristics: *input data and attributes*, *learning algorithms* and *visualization*, that are important for creating a trend analysis system.

In [Bolelli et al., 2009] a generative model is proposed based on latent Dirichlet allocation (LDA [Blei et al., 2003b]) that "integrates the temporal ordering of the documents into the generative process in an iterative fashion". This so called segmented author-topic model (S-ATM), based on the author topic model as presented in [Bolelli et al., 2007], integrates temporal characteristics of the document collection. Authors conduct their experiments on a subset of CiteSeer publications from 1990-2004 published in ACM conferences having a total number of 41,540 documents and 35,314 authors. They present the trends for five topics among which *machine learning*, *processor architectures* and *digital libraries* appear to be on top in the tested time period.

The trend detection model over Twitter data as described in [Kawamae and Higashinaka, 2010] relies on three steps of trend detection: identi-

[3] `http://dli.grainger.uiuc.edu/` accessed 01-August-2011
[4] `http://projects.ldc.upenn.edu/TDT/` accessed on 08-Nov-2011

fication of bursty keywords, grouping of bursty keywords into trends and trends analysis. The proposed approach has been realized as an application delivering a "real-time news bulletin assembled automatically from Twitter"[5]. No information about the test data set applied for the approach could be found in [Kawamae and Higashinaka, 2010].

5.2.2 Event-based and TDT

Introduced by the topic detection and tracking research, five tasks are in the focus of TDT field (page 3, [Allan, 2002]):

- story segmentation: the problem of dividing the transcript of a news show into individual stories
- first story detection: the problem of recognizing the onset of a new topic in the stream of news stories
- cluster detection: the problem of grouping all stories as they arrive, based on the topics they discuss
- tracking: requires monitoring the stream of news stories to find additional stories on a topic that was identified using several sample stories
- story link detection: the problem of deciding whether two randomly selected stories discuss the same news topic

Most research works addressing one of the tasks listed above belong to the event-based research direction of trend mining research. New event detection and event tracking are part of the TDT initiative [Allan et al., 1998]. The detection approach for new event detection is based on a single pass clustering algorithm and a thresholding model that incorporates the properties of events as a major component. This system adapts a sequential pattern matching technique used in data mining systems. Looking for frequently occurring patterns of words allows for the identification of frequently co-occurring terms which can be treated as a single topic. The topic comes up from the resulting words that are defined as a "phrase" [Lent et al., 1997]. The phrase frequency counts represent a data store that can be mined.

In [Petrović et al., 2010] an approach for detecting new events out of the Twitter stream is presented. The authors compare a classic method for first story detection (FSD), based on a nearest-neighbor search in an inverted document index to their locality sensitive hashing-based approach. Using a corpus of Twitter data collected over 6 months (163,5 million tweets) and applying the evaluation

[5] http://www.blicqtimes.com/ accessed on 08-Nov-2011

methods from TDT, they found out that "celebrity deaths are the fastest spreading news on Twitter".

In [Zeng and Zhang, 2009], which is inspired by the TDT research, a problem of topic transition is in focus. Instead of applying topic models based on latent Dirichlet allocation (LDA) as proposed in [Blei et al., 2003a] or probabilistic latent semantic indexing (pLSI) [Hofmann, 1999], they propose to ground on a topic transition model based on hidden Markov models (HMM). They compare their methods performing experiments on two corpora: Reuters-21578[6] (sampled and categorized by Reuters Ltd. in 1987, formatted and published in 1991-1992, containing 21,578 documents in 120 topics) corpus and BBS-1544 corpus which contains 1,544 documents from time period of July 2006 to March 2007 of BBS website in China.

An interesting work on tracking dynamics of topic trends, presented in [Morinaga and Yamanishi, 2004], is an example of trend mining research that can be classified into both event-based and general model directions. The authors [Morinaga and Yamanishi, 2004] focus on an online framework for tracking dynamics of topic trends while concentrating on three tasks: *topic structure identification*, *topic emergence detection* and *topic characterization* that are somehow relevant to the tasks from TDT research. [Morinaga and Yamanishi, 2004] apply a probabilistic model which is the finite mixture and propose a time-stamp based discount learning algorithm (a variant of an incremental expectation maximization clustering as presented in [Neal and Hinton, 1996]) for topic structure identification. They show their results from tests with a set of "contact data of a help desk for an internal email service" with 1,202 records from February to May of 2004.

5.3 Tools

5.3.1 ETDS Tools

Summarizing, the different tools that emerged from ETD research were:

PatentMiner: adapts a sequential pattern matching technique used in data mining systems. Looking for frequently occurring patterns of words, it allows to identify frequently co-occurring terms and to treat them as a single topic. The topic comes up from the resulting words that are defined as a 'phrase' [Lent et al., 1997]. The phrase frequency counts represent a data store that can be mined. Mining is done in this case by using a shape query processing learning tool borrowed from data mining [Agrawal et al., 1995]

TOA: a semi-automatic trend detection system for technology opportunities analysis [Kontostathis et al., 2003]. It enables mining of text files using bibliometrics [Watts et al., 1997]. TOA relies on the expertise of the user who is researching

[6] `http://tinyurl.com/len8xc2` online accessed on 30-July-2011

a given area. There are no inherent learning algorithms present in the system [Kontostathis et al., 2003].

TimeMines: a system that automatically generates timelines from date-tagged free text corpora [Swan and Jensen, 2000]. Two separate machine learning aspects are presented in TimeMines [Kontostathis et al., 2003]. One is the extraction of the most significant features from the input documents. This is done by the use of a simple statistical model for the frequency of occurrence of features in a stream of text (as described in [Swan and Allan, 1999]). The model is based on hypothesis testing, choosing the most relevant features from the date-tagged texts and allowing one to reduce the features set. Another learning method, again based on hypothesis testing, groups the features from the reduced feature set, due to the similarity in their distribution within a given time period [Kontostathis et al., 2003].

New Event Detection: including event tracking is part of the Topic Detection and Tracking (TDT) initiative [Allan et al., 1998]. The detection approach for New Event Detection is based on a single pass clustering algorithm and a thresholding model that incorporates the properties of events as a major component.

HDDI: relies on unsupervised and supervised learning approaches. The detection of emerging trends in text collections is based on semantically determined clusters of terms. Using linguistic feature extraction, clusters based on semantic similarity of these features are created. At multiple points in time, the algorithm takes a snapshot of the statistical state of the collection. Finally, a neural network that uses as input the rate of change in the size of clusters and in the frequency and association of features classifies topics as emerging or non-emerging.

The systems described above give an overview of approaches used for emergent trend detection in text mining with a focus on the learning methods used for detecting trends in text collections. We conclude that there is no inherent "best" approach for emerging trend detection in text collections and that systems' authors are mostly combining their own methods with supervised and unsupervised learning methods from machine learning.

Focusing more on prototypes for trend mining in financial news, we compared [Wüthrich et al., 1998] [Peramunetilleke and Wong, 2002]
[Mittermayer and Knolmayer, 2006]. The authors of [Wüthrich et al., 1998] describe a software application which uses the daily news of major publishers to predict the closing values of various indices in Asia, Europe and the USA. The indices are the Dow Jones Industrial Average (Dow), Nikkei 225 (Nky), Financial Times 100 Index (Ftse), Hang Seng Index (His), and the Singapore Straits Index (Sti). [Peramunetilleke and Wong, 2002] proceed very similarly with the development of their algorithm, however they focus on FOREX, particularly the USD/DEM and USD/JPY course. The goal of their system is also to predict price

changes. The authors try to improve the usual methods, which work only with series of numbers representing the price changes over a timeline, by processing press releases, which supply the causes of the change and thus the context for more accurate forecasts.

5.3.2 Algorithms, the web and the functionality tools

According to the system description in [Kontostathis et al., 2003] and regarding the prototypes [Wüthrich et al., 1998] [Mittermayer and Knolmayer, 2006] [Peramunetilleke and Wong, 2002], the following learning algorithms have been proven to be useful for the problem of trend detection:

- combined "hypothesis testing"-based methods [Swan and Jensen, 2000]
- single-pass clustering [Allan et al., 1998]
- sequential pattern matching and shape query processing [Lent et al., 1997] [Agrawal et al., 1995]
- feed-forward, backpropagation NN, C4.5 and SVM [Pottenger and Yang, 2001], [Wüthrich et al., 1998]
- k-NN classifier, regression analysis [Wüthrich et al., 1998]

In general we noticed that relevant works include different algorithms that have been applied to the problem of trend detection, but two general approaches are more popular than the rest:

- probabilistic topic models
- statistical learning (combined with text mining techniques)

Besides the academic research related to trend mining, there are some tools relevant for any analysis or experiments regarding trend mining. We chose to mention three of them: Rapid Miner[7], WEKA with Pentaho-Weka-widget[8], and GoogleTrends[9], which are the most important regarding mining trends.

Figure 5.2 shows an example of how GoogleTrends visualize common understandable examples of trends. That are *the financial crisis* and *insolvent companies*, which emerged in 2008 in the news on the web. The graph shows a search volume index for the terms "financial crisis" (blue curve) and "insolvent" (red curve) in Germany from 2006 to 2011.

[7] `http://rapid-i.com/content/view/182/192/lang,en/` accessed online 01-Feb-2013

[8] `http://wiki.pentaho.com/display/DATAMINING/Weka+Server` accessed online 20-April-2013

[9] `http://www.google.com/trends` accessed online 02-January-2013

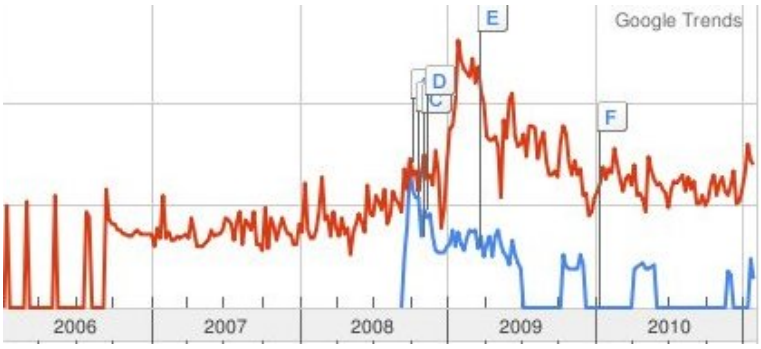


Figure 5.2: Example of trends based on Google search. Source: *[Google, 2011]*

5.4 Conclusions

Considering the ETD systems [Kontostathis et al., 2003] we notice that none of the methods proposed concentrate on the integration of knowledge (see Table 5.1).

	Approach	Learning Method	Knowledge Integration	Semantic Analysis
SEMI-AUTOMATIC SYSTEMS				
TOA	NLP, computational linguistic, LSI, principal component analysis	none	none	partial
TimeMines	statistical, probabilistic	not expl.	none	none
Patent Miner	query-based	seq.pattern matching	none	none
AUTOMATIC SYSTEMS				
Ænalyst	time series analysis, language models	Bayes-classificator	none	none
New Event Detection	clustering	single pass clustering	none	none
HDDI	statistical	back-propagation neural networks	none	partial

Table 5.1: ETD-systems. Source: *author*.

The research works focus on different tasks, algorithms and data sets. Regarding the sets used for tests, there are no specific recommendations for the size of the test set. Several chosen examples are summarized in Table 5.2

TEST SETS AND ALGORITHMS			
Test set	Time period	Task	Algorithm
163,5 million twitts	6 months	inverted index, hashing	locality sensitive hashing [Petrović et al., 2010]
41,540 documents, 35,314 authors	1990-2004	topic trends	segmented author-topic model [Bolelli et al., 2007]
21,578 documents, 120 topics (Reuters-21578)	1987	topic transition	HMM [Zeng and Zhang, 2009]
1,202 helpdesk records	Feb.to May 2004	finite mixture model	probabilistic– finite mixture [Morinaga and Yamanishi, 2004]

Table 5.2: Several algorithms and data sets. Source: *author*.

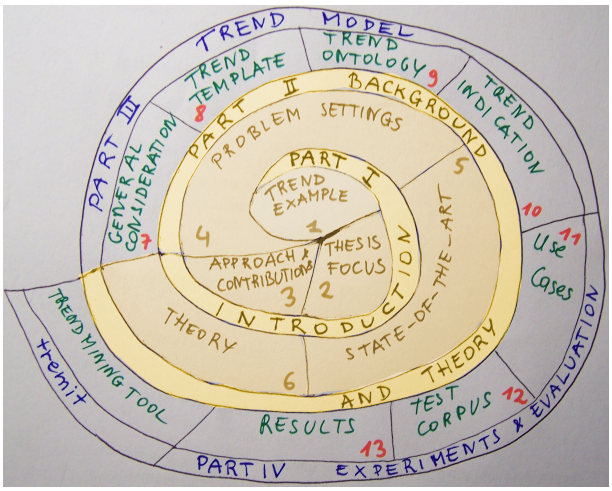
What we learn from the state of the art:

- the methods underlie one general assumption of what is a *trend* without an explicit definition. This implicit definition of a trend can be redefined with regard to the specific use case and data.
- there is no fundamental "best" approach to trend mining, rather a combination of text analysis methods and statistical learning approaches regarding the specific trend mining task/problem is the common technique
- probabilistic approaches (e.g. probabilistic topic models) and clustering methods are in general useful and often applied for topic detection tasks
- many different corpora varying in size, time period, labeling (automatic part-of-speech tagging or human made labels) and language are used for the tests and evaluation; there is no benchmarking corpora at hand[10]
- results of the specific methods often require validation from human experts and the "development and use of effective metrics for evaluation of ETD systems is critical" (page 1, [Kontostathis et al., 2003])

Furthermore, this review of the relevant research shows that there is still room for improvements regarding many issues. The most important once are listed as follows:

- *Time-based information retrieval makes a difference* – "time-based aspect is different from much other work in information retrieval, and specific handling of that is likely to be helpful." (page 13, [Allan, 2002])
- *The web as a source of texts is not just a "flat" document corpus* – "Although many datasets, such as TREC .GOV collection [NIST, 2013], have been built for research purposes, they are usually small and biased, and cannot represent the characteristics of the real-world web graph." [Dai and Davison, 2010]
- *Knowledge and context are important for the trend analysis* – "... we need to further analyze context, i.e. relations among words in order to more deeply analyze the semantics of topics." [Morinaga and Yamanishi, 2004]

[10] We are aware of the TDT-corpora available at `http://projects.ldc.upenn.edu/TDT/` accessed-online 01-June-2013. These are the test corpora as generated by the TDT project, and suited for tests within the TDT-formulated tasks.



CHAPTER 6

Theory

Chapter 6 helps in understanding the theory behind the methods applied to mining trends in this thesis. It starts with an introduction into the overall approach of IR, showing possible adoptions of it to the problem of trend mining. We go through possible relevant theoretical backgrounds for the trend mining research in general and continue with the representation models from IR. Three different representation models that underly three different approaches: probabilistic, statistical and graph-based are presented in this chapter. Showing the general differences between them, two known algorithms are introduced and described in detail, the topic models and the k-means clustering. This chapter closes with the description of an ontology in the sense of Semantic Web research and relates the ontology approach to the graph representation model mentioned in the beginning of this chapter.

6.1 Introduction

The theory behind trend mining can be interpreted in a general way by describing the existing theoretical approaches from relevant research fields, data mining or information retrieval respectively. Moreover, it can be discussed by explaining the theoretical background of specific algorithms applied for the experimental part of this thesis. Our goal here is the understanding of the difference between theoretical concepts relevant for trend mining in general and the theory behind the algorithms that we choose to apply to mining trends. However, the previous is connected to the latter – the consideration of a general trend mining approach is important for the decision on theoretical representations of the trend model, hence for the understanding of applied algorithms.

Since the focus of mining trends in this thesis lies on the textual data and we have to consider the document analysis, the concepts from text mining [Engel et al., 2010] are hereby relevant for trend mining in any case. However, the common text mining techniques, based on natural language processing of text and statistical text analysis, are just a part of the problem. Relating the trend mining to the data mining as shown in our definition from 4.2.3 leads to theoretical concepts of machine learning methods used in data mining. However, we do not exactly apply machine learning theory to trend mining. Hence, the question remains: what do we need for mining trends? In any case, having the textual data, we need a representation model and an overall idea about the process of mining trends. A visualization of the information retrieval process from [Göker and Davies, 2009] as shown in Figure 6.1 represents a general IR approach, starting with the information need and resulting in a ranked number of documents. This general approach is summarized by the authors: "There are three basic processes an information retrieval system has to support: the representation of the content of the documents, the representation of the user's information need, and the comparison of the two representations." (p. 2, [Göker and Davies, 2009]) Accordingly, indexing refers to representing the documents, whereas the resulting query is the user's formulating of her/his information need. If we adopt the information retrieval process to trend mining, we can illustrate the trend mining process as shown in Figure 6.2[1], 6.3[2] and 6.4. Figures 6.2 and 6.3 take the overall information retrieval process and transform it into the relevant issues regarded in trend mining by adding time as a component into the IR. Figure 6.4 visualizes the interpretation of trend mining that is being suggested by this thesis.

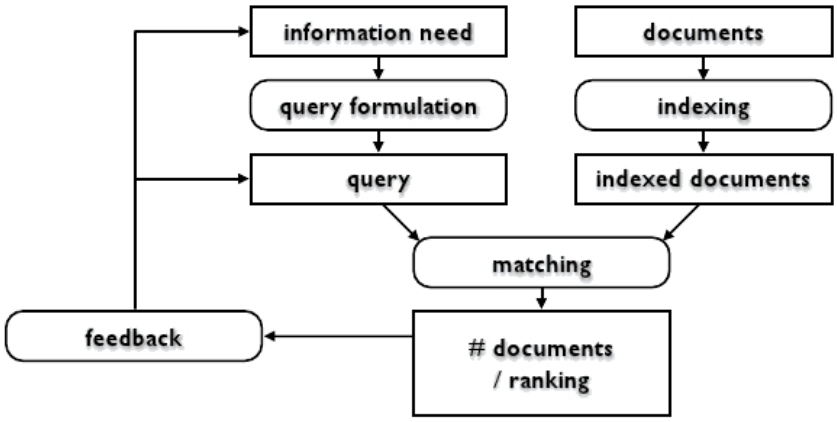


Figure 6.1: Information retrieval process. Source: *[Göker and Davies, 2009]*.

[1] Axis 'Zeit' represents the time.
[2] 'Themengebiete' means 'topic areas'.

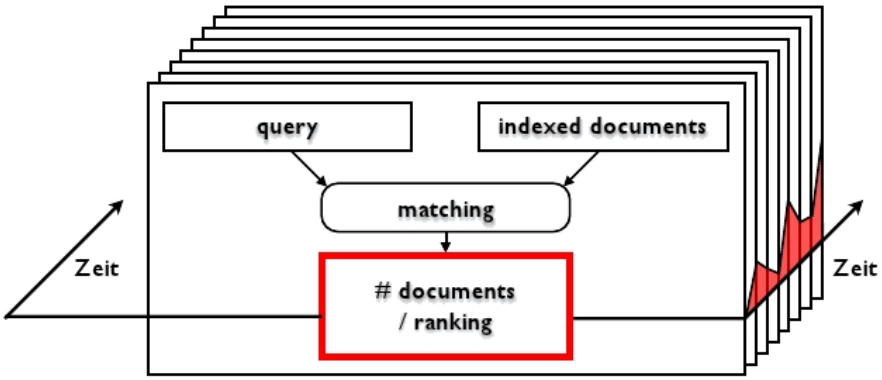


Figure 6.2: Trend mining process, results' focus. Source: *[Streibel et al., 2013a]*.

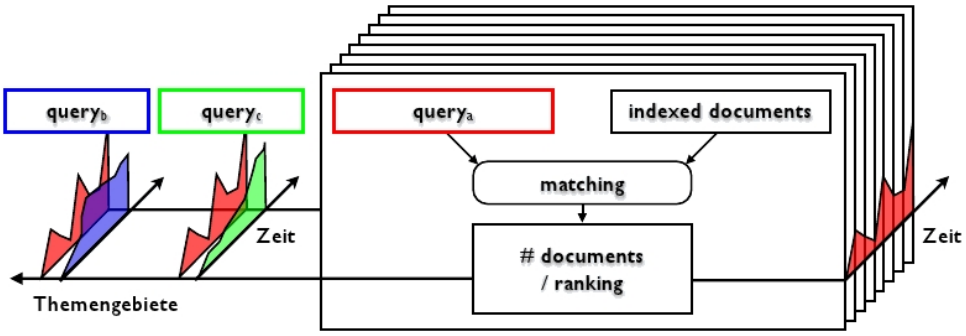


Figure 6.3: Trend mining process focusing on query. Source: *[Streibel et al., 2013a]*.

The idea of the trend mining process as visualized by Figure 6.2 concentrates on the results representation. In this case, the time component is being added into the process of information retrieval after the matching of a query with indexed documents. The resulting documents are presented according to their time relevance and the trend is contained in their content, changing over time in regard to the same query.

In Figure 6.3, the time component is relevant for the query itself. A trend is interpreted in this case as the change in the queries over the time.

Finally, Figure 6.4 presents our interpretation of trend mining from documents. We propose to understand it as an overall process of information retrieval based on time and interest. An important step in the process is that the indexing of documents depends on time and interest selection. The query remains optional.

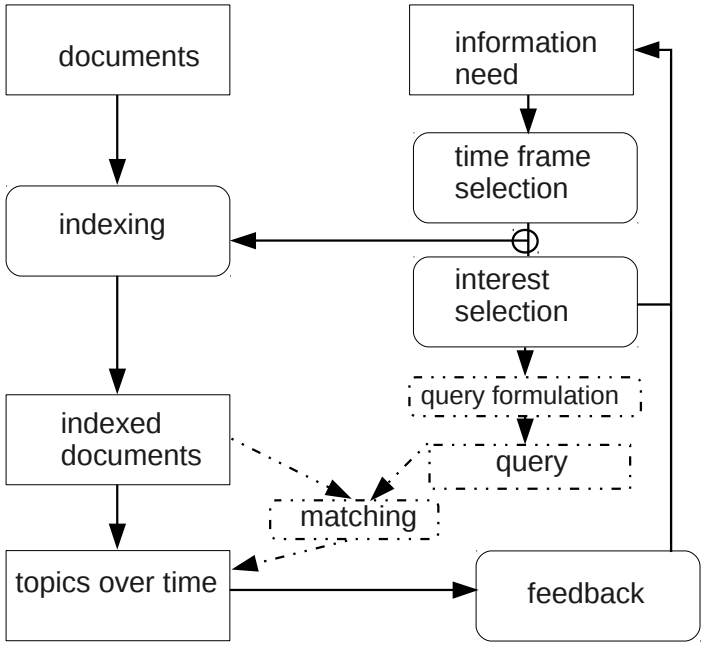


Figure 6.4: The overall trend mining process. Source: *author*.

6.2 Different approaches

In information retrieval, the representation of information contained in documents is based on a given representation model, examples of which are listed below:

- *exact match models*: boolean model and region model
- *vector space model (VSM)*
- *probabilistic models*: bayesian model and language model

The boolean model is "the first model of information retrieval" (p. 3, [Göker and Davies, 2009]) and it is a very straightforward model. It assumes that a query, represented as a simple term, would match with a set of documents indexed exactly with this given term in an unambiguous way. The query and the corresponding documents are combined into new sets using Boole's logic, e.g.

operators: *AND*, *OR*, *NOT*. This model does not enable any ranking of retrieved documents.

More sophisticated models are the vector space and the probabilistic model. In the following subsections we discuss VSM and probabilistic models that will help to explain the theoretical background from two algorithms introduced in sections 6.5 and 6.3. We also introduce the basics of graph-based models of which exemplary approach we interpret *Ontology* in Section 6.6.

6.2.1 Vector space model

This IR approach is based on Luhn's similarity criterion, generally formulated as follows:
"The more two representations agreed in given elements and their distribution, the higher would be the probability of their representing similar information" (p. 5, [Göker and Davies, 2009]). The idea behind VSM is to represent the query as well as the documents in the form of vectors of terms in Euclidean space. Representing documents as vectors in high-dimensional space in which each term corresponds to one dimension allows for applying the similarity metrics in order to find the best result which means the highest similarity between the query (also represented as a vector) and the corresponding similar documents. Therefore, VSM allows for ranking the query results according to their similarity.

If we transfer the idea of VSM for trend mining, considering the problem of trend mining from the search problem perspective (more on this in Section 7.4 of Chapter 7), we shall imagine the trend as an emerging topic area in a document collection. This emerging topic area shall be represented by a vector of terms. For each document represented as a vector, the similarity value can be determined between the vector defining trend and the given document. The resulting list of documents most similar to the vector of trend (terms from emerging topic area) represent the set of documents that are potentially trend indicating and can be further analyzed.

6.2.2 Probabilistic approach

One of the problems of vector space model is the term weighting problem (p. 7, [Göker and Davies, 2009]). Since the vector based representation of the document does not provide per se the values that have to be contained in the vector, term weighting methods are needed. The most known method is Salton's *tf-idf* term weighting function [Salton et al., 1982] – based on considerations of term and document weighting from [Jones, 1972] – that does not always perform well. Another problem is the incremental update of the document index. When a new document is added into the collection, all of the vectors have to be updated. The

solutions for these problems leads to the probability theory in which probabilistic approaches are grounded.

The probability based approach in general takes into account the estimation of probability instead of counting the unambiguous values for terms and documents in both, the relevancy of a term for defining the given document and in relevancy of a document as a result for the given query. The probabilistic indexing model suggests assigning probabilities for index terms, creating the set of possible terms for each document. These terms are weighted by the probability $P(T|D)$ that dictates the level of the probability that the given document D contains the given information that can be simply defined by the term T. The ranking of documents is then given by $P(D|T)$. Other probabilistic approaches are: *probabilistic retrieval model*, *2-Poisson model*, *Bayesian network models*, and *language models* (p. 8-15, [Göker and Davies, 2009]).

6.2.3 Graph-based approach

In the graph-based model, documents are represented as graphs. This representation allows for preserving the document structure and its semantics. Different approaches for the graph-based IR model have been proposed, i.e. representing documents by conceptual graphs. The graph-based approaches deal with the query as a graph and the document as a graph. The goal is to find the perfect match between them (the graph isomorphismus). Regarding the Semantic Web with its *ontology* approach, documents can be represented as directed, labeled resource description framework (RDF) [W3C, 2004] graphs and the query can be represented using a query language, such as SPARQL [W3C, 2008]. If we would, regarding the semantics, transform the trend represented as an emerging topic area into a SPARQL query and model the documents as RDF graphs, and then apply graph matching algorithms, we could retrieve all document parts fitting to the query. Of course, it doesn't make sense to model the emergent topic area completely as a graph since it would be too inefficient. But the modeling of documents partially as graphs (i.e. applying schemas like FOAF [Brickley and Miller, 2010]) enables retrieval of more knowledge about the document content.

6.3 K-Means clustering

As a statistical method, the clustering algorithm k-means, applied to the problem of document clustering, uses the VSM paradigm. Documents that have to be clustered are represented as vectors and the k-means method applies different distance metrics to find the similarity between these vectors.

The k-means clustering method is also called centroid-based technique and partitioning method [Han and Kamber, 2006] [Engel et al., 2010]. This algorithm takes as input the parameter $\mathcal{K}$ and partitions a given set of objects n, one of

each represented by a vector a, into $\mathcal{K}$ clusters. The resulting similarity between the objects of a respective cluster is higher than the resulting similarity between the clusters.

6.3.1 General description

The procedure of k-means first randomly selects k of the objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function converges. Typically, the **square-error criterion** is used, defined as:

$$\mathcal{E} = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{6.1}$$

where $\mathcal{E}$ is the sum of the square error for all objects in the data set; p is the point in space representing a given object; and m_i is the mean of cluster C_i (both p and m_i are multidimensional). In other words, for each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed. This criterion tries to make the resulting k clusters as compact and as separate as possible (p.402, [Han and Kamber, 2006]).

6.4 K-means: batch and incremental

Vector[3] a is described as (p. 82-83 in [Engel et al., 2010]):

$$a \in \mathbf{R}^n$$

Elements of the vector a are:

$$(a[1], \ldots, a[n])^T$$

The set $\mathcal{A}$ is a finite set of a size $|\mathcal{A}|$ and describes the set of m vectors by:

$$\mathcal{A} = \{a_1, \ldots, a_m\} \subset \mathbf{R}^n$$

C is a prescribed subset of $\mathbf{R}^n$, d is distance-like function $d(x,a)$ The centroid c of set $\mathcal{A}$ is notated with $c = c(\mathcal{A})$ and represents the solution of the minimization problem:

$$c = argmin \left\{ \sum_{a \in \mathcal{A}} d(x,a), x \in C \right\}$$

[3]This paragraph is the K-Means description as on p.82-82 in [Engel et al., 2010]

In the case of the squared Euclidean distance: $d(x,a) = ||x_a||^2$ the set $\mathcal{C}$ may be the entire space. When the relative entropy is used:

$$d(x,a) = \sum_{i=1}^{n} a[i]log(a[i]/x[i])$$

the set $\mathcal{C}$ of housing centroids x should be restricted to vectors with at least nonnegative entries: $a[i] \geq 0$
$\mathcal{Q}$- the quality of the set $\mathcal{A}$ is defined by:

$$\mathcal{Q}(\mathcal{A}) = \sum_{i=1}^{m} d(c,a)$$

where $c = c(\mathcal{A})$ and $\mathcal{Q}(\emptyset) = 0$
Let $\Pi = \{\pi_1, \ldots, \pi_k\}$ be a partition of $\mathcal{A}$, i.e.

$$\bigcup_i \pi_i = \mathcal{A}$$

, and $\pi_i \cap \pi_j = \emptyset$ if $i \neq j$. The quality of the partition Π is defined by:

$$\mathcal{Q}(\Pi) = \mathcal{Q}(\pi_i) + \ldots + \mathcal{Q}(\pi_k) = \sum_{i=1}^{k} \sum_{a \in \pi_i} d(c_i,a)$$

where $c_i = c(\pi_i)$
The goal is to find a partition $\Pi^m in = \{\pi_1, \ldots, \pi_k\}$ that minimizes the value of the objective function $\mathcal{Q}$
Partitions and centroids are associated in the following way:

1 Given a partition $\Pi = \{\pi_1, \ldots, \pi_k\}$ of the set $\mathcal{A}$ one can define the corresponding centroids $\{c(\pi_1), \ldots, c(\pi_k)\}$ by

$$c(\pi_i) = argmin \sum_{a \in \pi_i} d(x,a), x \in \mathcal{C}$$

2 For a set of k centroids $\{c_1, \ldots, c_k\}$ one can define a partition $\Pi = \{\pi_1, \ldots, \pi_k\}$ of the set $\mathcal{A}$ by
$\pi_i = \{a : a \in \mathcal{A}, d(c_i,a) \leq d(c_l,a)$ for each $l = 1, \ldots, k\}$

The classical batch k-means algorithm iterates between two steps described in 6.3 to generate a partition $\Pi^{'}$ from a partition Π. While step 2 is straightforward, step 1 requires to solve a constrained optimization problem. The degree of difficulty involved depends on the distance-like function d and the set $\mathcal{C}$. The entire procedure is a gradient-based algorithm.

Incremental k-means is an iterative algorithm that seeks to change the cluster affiliation of one vector per iteration.

6.4.1 Distance metrics

Distance-like functions are:

- euclidean distance
- Kullback-Leibler divergence (relative entropy)
- Manhattan (city block) metric

Euclidean distance and Manhattan metric are both generalized by the Minkowski metric:

$$d(i,j) = (|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \ldots + |x_{in} - x_jn|^p)^{1/p} \tag{6.2}$$

where p is a positive integer. This distance represents Manhattan distance when $p = 1$ and Euclidean distance when $p = 2$. There is also the possibility of assigning weight to each variable according to its importance. Then the *weighted Euclidean distance* can be computed:

$$d(i,j) = \sqrt{w_1|x_{i1} - x_{j1}|^2 + w_2|x_{i2} - x_{j2}|^2 + \ldots + w_m|x_{in} - x_{jn}|^2} \tag{6.3}$$

6.4.2 Algorithm

In general, the k-means algorithm functions as follows:
Input:

- k: the number of clusters
- D: a data set containing n objects

Output: a set of k clusters

Method:

1. arbitrarily choose k objects from D as the initial cluster centers;
2. repeat
3. (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
4. update the cluster means, i.e., calculate the mean value of the objects for each cluster;
5. until no change;

6.4.3 Geometrical interpretation

The geometrical interpretation of the k-means method is illustrated in Figure 6.5 (p. 403, [Han and Kamber, 2006]). It shows the steps of clustering the different objects according to their similarity and their distance of the cluster centres.

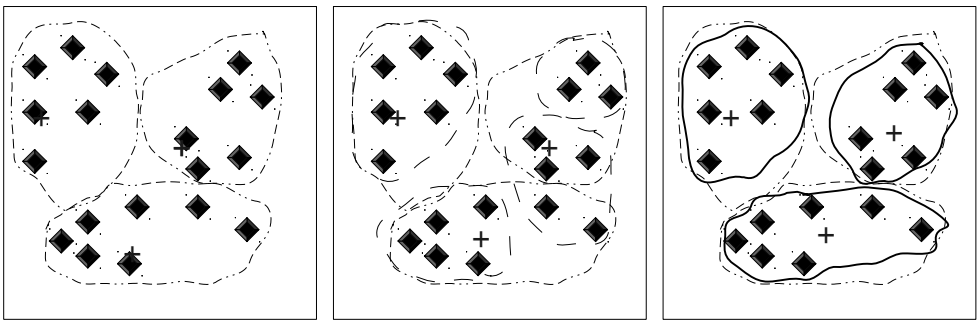


Figure 6.5: Geometrical interpretation of k-means. Source: *author*.

6.5 Topic modeling

An example of a probabilistic model is the topic modeling that can be applied to a set of documents. Topic modeling is a computational tool that helps in organizing, searching and understanding "vast amount of information". It belongs to a group of *probabilistic topic modeling* - "a suite of algorithms that aim to discover and annotate large archives of documents with thematic information"[Blei, 2011]. Without the need for prior annotation or labeling of the documents, topic modeling as a statistical method enables us to organize and summarize documents while discovering the themes that run through them [Blei, 2011].

6.5.1 General description

In general, the topic models offer a way to understand a given document collection as a mixture distribution of topics. Documents in a collection can be described by topics. The topics can be described by words. Every document in a collection contains different topics; it is a set of topics with different probability values that express how much a given topic characterizes the given document. In fact, documents consist of words. The distribution of words over the topics is defined

by the probability values that express how much a given word describes the given topic.

Given that, a *collection* of documents is a set of a mixture of topics, a *document* is a distribution over topics, a *topic* is a probability distribution over the word, and a *word* is the basic unit of discrete data.
"The goal of topic modeling is to automatically discover the topics from a collection of documents" [Blei, 2011]. The topic model can be understood as 1) a generative model: given the topics, generate the documents or 2) inverted, as a statistical inference problem: given the documents, infer the topics. In the following, we describe latent Dirichlet allocation (LDA) in order to understand topic models in greater detail.

6.5.2 Latent Dirichlet Allocation

A topic model is a generative model and it specifies a simple probabilistic procedure by which documents can be generated. The data in generative probabilistic modeling arises from a generative process that includes *hidden variables*. "This generative process defines a *joint probability distribution* over both the *observed* and *hidden random variables*. We perform data analysis by using that joint distribution to compute the *conditional distribution* of the hidden variables given the observed variables. This conditional distribution is also called the *posterior distribution*." In the case of probabilistic topic modeling, the observed variables are the words of the documents and the hidden variables are the topic structure [Blei, 2011].

The simplest kind of topic model is the latent Dirichlet allocation. LDA is a generative probabilistic model of a corpus. "The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words." Formally, the following terms are defined [Blei et al., 2003b]:

- A word is defined as the basic unit of discrete data. Furthermore, it is an item from a vocabulary indexed by $1, ...,V$. The words are represented using unit-basis vectors that have a single component equal to one, all other components equal to zero. The vth word in the vocabulary is represented by a V-vector w such that $w^v = 1$ and $w^u = 0$ for $u \neq v$.

- A document is defined as a sequence of N words: $\mathbf{w} = (w_1, w_2, ..., w_N)$, where w_n is the nth word in the sequence.

- A corpus is a collection of M documents: $D = \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M$.

The documents themselves are observed, while the topic structure — the topics, per-document topic distributions, and the per-document per-word topic

assignments — are hidden structure [Blei, 2011]. The generative process for a document collection D under the LDA model is as follows [Darling, 2011]:

1 For $k = 1...K$:

(a) $\phi^{(k)} \sim Dirichlet(\beta)$

2 For each document $d \in D$:

(a) $\theta_d \sim Dirichlet(\alpha)$

(b) For each word $w_i \in d$:

i. $z_i \sim Discrete(\theta_d)$

ii. $w_i \sim Discrete(\phi^{(z_i)})$

where K is the number of latent topics in the collection, $\phi(k)$ is a discrete probability distribution over a fixed vocabulary that represents the kth topic distribution, θd is a document-specific distribution over the available topics, z_i is the topic index for word w_i, and α and β are hyper parameters for the symmetric Dirichlet distributions that the discrete distributions are drawn from. The generative process described above results in the following joint distribution:

$$p(\mathbf{w}, \mathbf{z}, \theta, \phi|\alpha, \beta) = p(\phi|\beta)p(\theta|\alpha)p(\mathbf{z}|\theta)p(\mathbf{w}|\phi_z) \tag{6.4}$$

6.5.3 Gibbs Sampling

Gibbs Sampling it the most used algorithm based on a Markov Chain Monte Carlo approximate inference. It is a sampling method in which the missing values of the variables are randomly generated and the probability distribution of the variables for which the value was simulated, are exchanged with other variables. The process continues until an acceptable value for the conditioned probability is found (more detailed description can be found in [Darling, 2011]). The Pseudocode 6.5.1 shows the implementation of LDA Gibbs sampling.

6.5.4 Algorithm

LDA Gibbs Sampling [Darling, 2011]

Algorithm 6.5.1: LDAGibbsSampling(w,d)

comment: **w**ords $\in$ **d**ocuments

comment: Output: topic assignments z and counts $n_{d,k}$, $n_{k,w}$, and n_k

comment: randomly initialize z and increment counters

for each *iteration*
do
- **for** $i \leftarrow 0$ **to** $N-1$
  do
  - $word \leftarrow w[i]$
  - $topic \leftarrow z[i]$
  - $n_{d,topic} -= 1; n_{word,topic} -= 1; n_{topic} -= 1$
  - **for** $k \leftarrow 0$ **to** $K-1$
    do
    - $p(z = k|\Delta) = (n_{d,k} + \alpha_k)\frac{n_{k,w}+\beta_w}{n_k+\beta x W}$
  - $topic \leftarrow sample$ **from** $p(z|\Delta)$
  - $z[i] \leftarrow topic$
  - $n_{d,topic} += 1; n_{word,topic} += 1; n_{topic} += 1$

return $(z), n_{d,k}, n_{k,w}, n_k$

6.5.5 Geometrical interpretation

The geometrical interpretation as given in [Blei et al., 2003a] is illustrated in Figure 6.6. It shows so called *topic simplex* for three topics embedded in a word simplex for three words.

6.6 Ontology

"...in its most prevalent use in AI, an ontology refers to an engineering artifact, constituted by specific vocabulary used to describe a certain reality, plus a set explicit assumptions regarding the intended meaning of the vocabulary words. This set of assumptions has usually the form of a first order logical theory, where vocabulary words appear as unary or binary predicate names, respectively called concepts and relations. In the simplest case, an ontology describes a hierarchy of concepts related by subsumption relationships; in more sophisticated cases, suitable axioms are added in order to express other relationships between concepts and to constrain their intended interpretation." [Guarino, 1998]

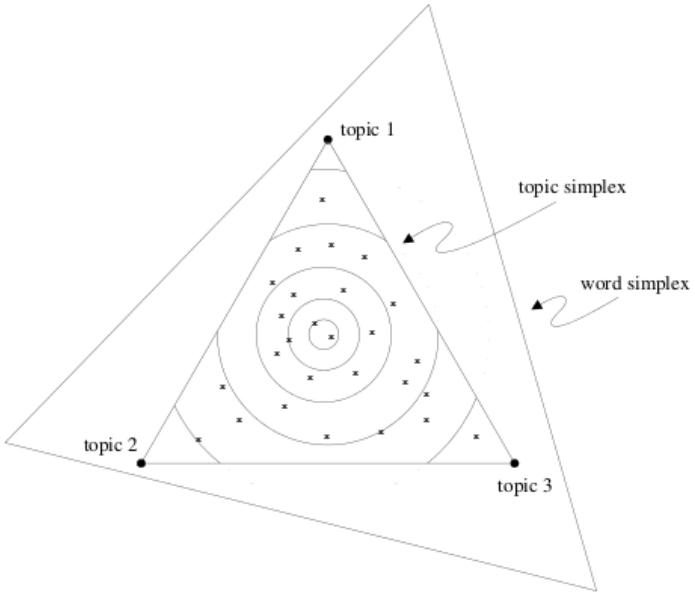


Figure 6.6: Topic model geometrical interpretation. Source: [Blei et al., 2003b]

6.6.1 General description

The idea of the ontology is described differently in the literature, however always referring to one of the following – *an artifact*, *a conceptualization*, *a specification*, *an agreement*. An ontology is " a specification of the conceptualization" according to [Gruber, 1993] and "Every ontology is a treaty–a social agreement–among people with some common motive in sharing."(p. 439, [Peter Norviq, 2003][Gruber, 2004]). In general, being a conceptualization, an ontology can be defined in terms of the AI research[Genesereth and Nilsson, 1987][Guarino, 1998] as a structure:

$$< D,R > \tag{6.5}$$

where D is the *domain* and R are the *relations* on D. [Guarino, 1998]. There is a problem with this general understanding of an ontology. [Guarino, 1998] argues that an ontology is more than a the context-less conceptualization and shows the need for a context of a given domain, calling it the domain's *world*. Proposing to specifically define the domain as a *domain space*: $< D,W >$, and the relations on the domain as a *conceptual relations*, he concludes that a *conceptualization* is actually defined by:

$$C =< D,W,\Re > \tag{6.6}$$

where $\Re$ is a *set of conceptual relations* on the domain space $< D,W >$. The *conceptual relation* ρ^n of arity n on the domain space $< D,W >$ is defined as a

total function:

$$\rho : W \rightarrow 2^{D^n} \tag{6.7}$$

Having defined the conceptualization, [Guarino, 1998] defines *intended world structures* S_c on C, a language L, an *ontological commitment* K, and a set $I_k(L)$. The latter is the set of *intended models* of L according to K, namely all models of L that are compatible with K. Given a language L with ontological commitment K, an ontology is defined as a *set of axioms designed in a way such that the set of its models approximates as best as possible the set of intended models of L according to K*. Figure 6.7 illustrates the definition of what an ontology is, relating it to

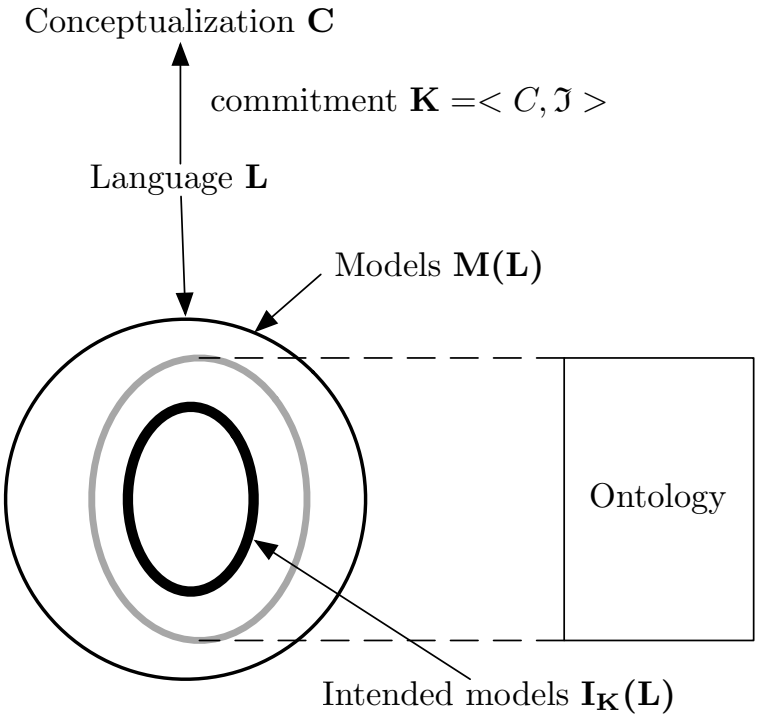


Figure 6.7: Definition of *ontology*. Source: *[Guarino, 1998]*.

the language L and the conceptualization C mentioned above. "It is important to stress that an ontology is language-dependent, while a conceptualization is language-independent" [Guarino, 1998].

6.6.2 Expressivity levels

The general description of an ontology as illustrated at the beginning of this section helps in understanding on the abstract level what an ontology is. On the less abstract level and from the practical point of view, an ontology allows for defining knowledge about a given domain in a formal way. The formal way means, that there are formal languages which we can apply for defining *semantics* of our

conceptualization of a given domain. The development of the research within the Semantic Web [Berners-Lee et al., 2001] caused in particular the creation of the respective ontology languages. Following ontology languages have been created: DAML-OIL [4], DAML-ONT [5], RDF[6], RDFS[7], OWL[8], OWL2[9]. While DAML-OIL is not in a wide use anymore, most popular is currently the use of RDF/S and OWL/2. The difference in the respective languages lies in their expressivity levels. RDF, based on the simple XML syntax, is the simplest language for defining the simplest ontology (in the sense of semantic expressivity). [Antoniou and van Harmelen, 2003]

6.6.3 Example

A very simple ontology example can be constructed as follows– for the given description about the persons, their relationship, and their pets, we construct an ontology visualized in the Figure 6.8. As described in (p. 50–51 [Alnemr, 2012]):

- *Classes* in OWL are concrete representations of concepts that describe the domain or are relative to the domain. They are the *sets* that have *individuals* and are represented using formal descriptions that state the requirements for membership of the class. They can be organized in a subclasses-superclasses hierarchy (also known as *taxonomy*). All classes are headed by the `Thing` class. For a class to be a *subclass* of another, it means that all of its instances are instances of the superclass (i.e. necessary implication). For example, class `Animal` is the superclass of `Cat` which means all member of `Cat` are also members of `Animal`.

- *Properties* in OWL are the relations between the individuals or objects, e.g. `hasChild`, `isFriendsWith`, `hasPet`. They correspond to *roles* in DL and *relations* in UML. They can have *inverses* (e.g. `hasOwner` and `isOwnedBy`). They also can be *symmetric* (e.g. `hasSibiling`), *transitive*, *reflexive* or have a single value (*functional*). These are characteristics that can affect their inference behaviour. There are two types of properties: object properties and data properties. An *object property* is a relation between two individuals. A property relates objects in its *domain* to objects in its *range*. A *data property* relates an object to a data value (i.e. XML Schema Datatype value[10] or an rdf literal) not other objects. For example, `Olga hasAge` "30".

[4] `http://www.w3.org/TR/daml+oil-reference` accessed 20-Jun-2013
[5] `http://www.daml.org/2000/10/daml-ont.html` accessed 20-Jun-2013
[6] RDF: `http://www.w3.org/TR/REC-rdf-syntax/` accessed 20-Jun-2013
[7] `http://www.w3.org/TR/rdf-schema/` accessed 20-Jun-2013
[8] `http://www.w3.org/2004/OWL/` accessed 20-Jun-2013
[9] `http://www.w3.org/TR/owl2-overview/` accessed 20-Jun-2013
[10] XML Schema: `http://www.w3.org/TR/xmlschema-2/` accessed 20-Jun-2013

- *Individuals* in OWL are the objects in the domain that we are interested in and which hold a unique name. They are instances of their classes e.g. `Simi` is an instance of class `Cat`.

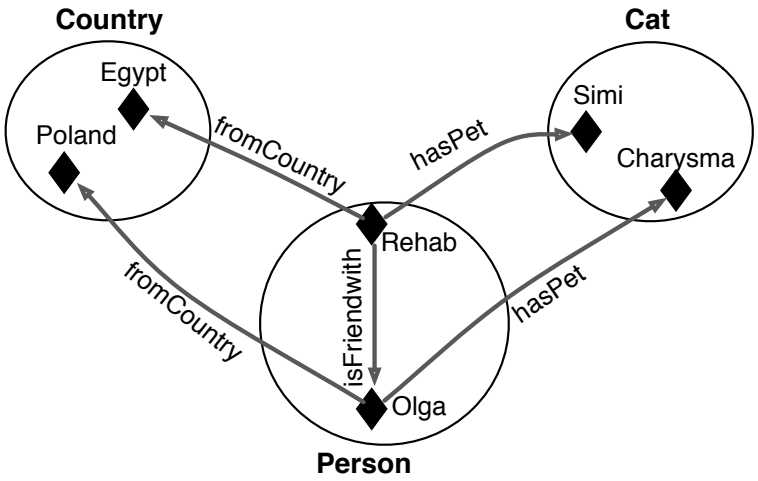
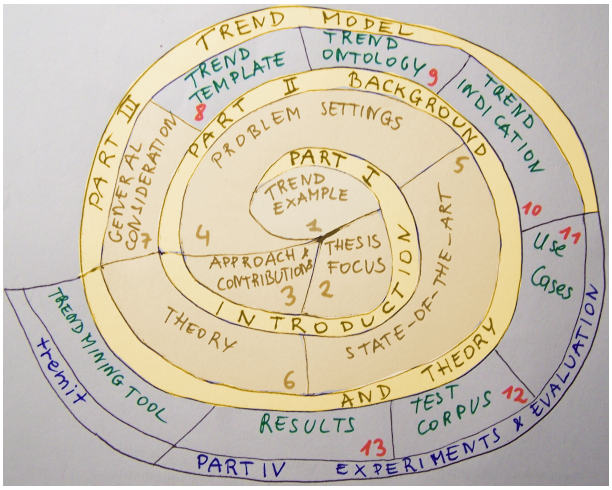


Figure 6.8: A simple ontology. Source: *[Alnemr, 2012]*.

An ontology as a conceptualization of knowledge is applicable for any domain and any knowledge as long as this knowledge can be represented by relevant concepts and relations. In Chapter 7 we describe the creatoin of a trend ontology for the market research, discussing its limits and obstacles and in Chapter 9 we give an overview over the applicable trend ontology for financial market.

Part III

Trend Model



CHAPTER 7

General considerations

Chapter 7 explains which aspects are important of a knowledge-based approach on mining trends. Based on a case from market research, we discuss here the idea and the development process of a trend ontology. The discussion allows for identifying obstacles and issues relevant for a knowledge-based approach in trend mining. This chapter includes an excursion in the problem of trend mining in general, posing the question: is trend mining a search or a discovery problem? We go through possible answers to this question and chose the most appropriate way for dealing with this problem. The search and the knowledge discovery perspective will be mentioned again in Chapter 13.

7.1 Preliminaries

In the previous chapter we discussed possible theoretical aspects of trend mining, including the graph based representation of the documents by the use of an ontology model. Before we deliberate on a search versus knowledge discovery perspective on trend mining, we report in the sections 7.1.1 to 7.3 on the development process of an ontology for trend mining in market research. We describe the preliminary stage of so-called trend ontology, which we created for the market research case presented in Chapter 11. The content of this chapter is the basis for the trend template definition in Chapter 8 and it has been mainly published in [Streibel and Mochol, 2010].

7.1.1 Specifics of the market research case

A trend

A trend in terms of market research is the evolution of customers' opinion referring to a specific topic that can be described by its categories or labels. Customer opinion is strictly conjoined with sentiments used by customers to express linguistically their emotional viewpoint on specific issues. In general, a trend mining method for market studies should enhance the efficiency in the analysis of textual market research data that is generated in primary and secondary research (for more explanation on market research use case, see Figure 7.1 and description in Section 11.1.2).

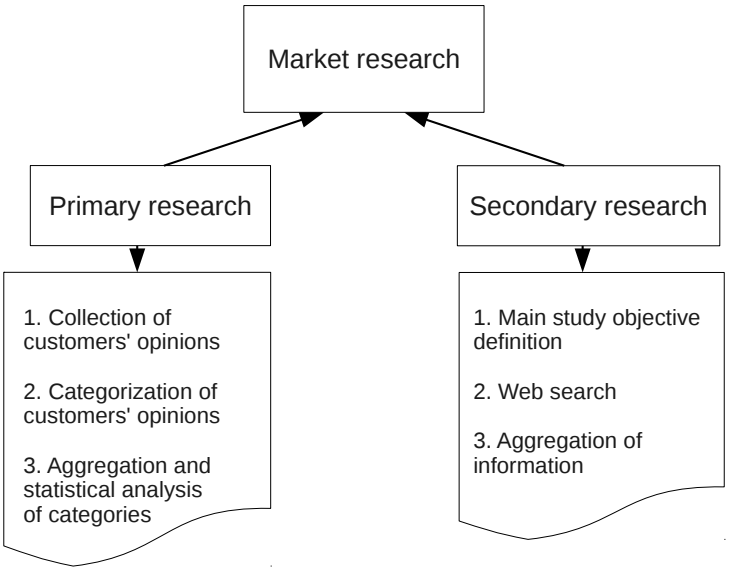


Figure 7.1: Market studies – primary and secondary research. Source: *author*.

Tasks in trend mining

Tasks in trend mining include the automatic categorization of open ended questions and their valuing process, the filtering of relevant information, and the identification of trends. In terms of market studies, a trend ontology should sup-

port the analysis process by providing knowledge regarding main market research concepts that occur in texts of the market research projects (e.g. the concepts *image* or *product quality* in terms of market research). The ontology should cover the concepts' definitions for the main keywords and terms used by customers in order to describe their opinion (e.g. substantives, verbs, and adjectives that help to classify sentences such as "this brand fits to me", "I like the nice logo"). Further, it should include the definition of categories used in terms of market studies on customer opinion (e.g. "overall satisfaction", "level of commitment"). In the best case, the ontology should support the categorization of customers' opinions based on a given list of categories that are relevant for the respective project, therefore it should cover the knowledge about trend indicating features of any given keyword or term (e.g. positive, negative and neutral description keywords).

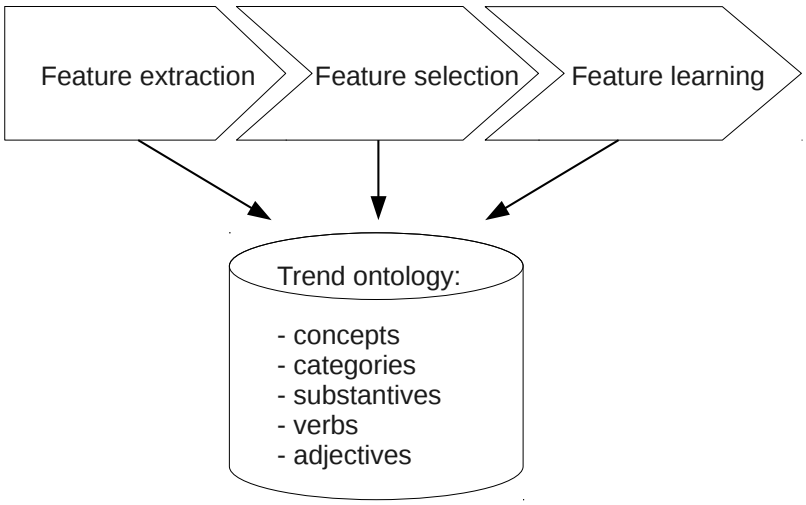


Figure 7.2: Trend mining process in market research. Source: *author*

.

Trend ontology requirements

A trend ontology, that has to be defined as a knowledge model supporting the trend mining process (see Figure 7.2), is an ontology that contains:

- the meta-level knowledge about market research concepts (commonly used in the market research)

- common keywords used in the market research projects (based on market research specific projects)
- knowledge about trend indicating terms and relations between the particular concepts in market research.

Furthermore, the trend ontology should be used as a knowledge base that can be applied in different phases of the trend mining process:

- feature extraction
- feature selection
- learning stage

7.1.2 Engineering methods

According to [Dimitrova et al., 2008], Ontology Engineering (OE) is defined as the "set of activities that concern the ontology development process, the ontology life cycle, and the methodologies, tools and languages for building ontologies". In recent years OE has evolved from a pure research topic being common in scientific domains to real world applications, which was demonstrated by the wide range of projects with major industry involvement and by the increasing interest of small and medium-sized enterprises (SMEs) requesting consultancy in this domain. At the beginning the knowledge engineers managed and controlled the ontology authoring process, but as the ontologies become larger covering more specific domains, the involvement of the domain experts became indispensable and the ontology development could be tackled only through the intensive cooperation of ontology engineers and domain experts in the context of large spatial distributed teams. The authors of [Braun et al., 2007] state that the ontology authoring process requires not only an active participation of domain experts but they should also lead the entire process providing the relevant domain and conceptual knowledge. Furthermore, a number of other aspects like dealing with context or data and web integration become crucial. In order to build and deploy ontologies on a large scale beyond the boundaries of the academic community, there is still a need for technologies to assist the implementation process. Most OE methodologies rely on specialized knowledge engineers but in real world-settings the need for maintenance of domain ontologies emerges in the daily work of its users [Braun et al., 2007].

7.2 Yet another ontology?

In this section, we describe our experience in modeling the trend knowledge and creating the first version of the trend ontology that helped us in further development of the knowledge-based trend mining approach.

7.2.1 Methodology for trend ontology

During our research in the TREMA project, we experienced the difficulty of applying common OE methodologies developed by academia to the practical problem of the trend ontology development. Therefore we used an agile, practical and expert-based method; the prototypes of trend ontologies for market research were developed under active participation of market research experts on the basis of three knowledge models. Our aim was to define a lightweight knowledge base that can be used in real-time as enhancement for statistical learning methods, therefore our trend ontologies do not include any rules. The ontologies are modeled for the German language.

7.2.2 Keyword/concept based trend ontology

Relying on the experience of experts from the market research domain, we identified and modeled with Protégé[1] using RDFS[2] an initial keyword set categorized by the main concepts of the market research (our case considered only the technology market). The main set categories are: *Image* (image), *Produktqualität* (product quality), *Kundenbeziehung* (customer relation), *Service* (service), *Stimmungsbild/Wahrnehmung/Entscheidung* (public opinion/customers' opinion/decision). Each category is implemented as a class consisting of relevant concepts that describe the category. For the product quality category, the concept set consists of Zuverlässigkeit (realiability), Performanz/Leistung (performance/power), etc. We defined the class property included_in, in order to express semantically the category membership of the given keyword/concept. In addition to the categorized concept sets, we modeled synonyms for several keywords/concepts and added the trend-indicating property to each concept that had been classified by experts as trend-indicating ones. Keyword/concept based trend ontology is built on a very simple schema and can be easily applied, for instance, in order to extend the word based feature vector creation as for machine learning methods. Figure 7.3 visualizes a snippet from the ontology, showing the connections between the concepts.

7.2.3 Term field based trend ontology

Extending the keyword-based trend ontology we observed the emergence of so-called term fields in market research, which correspond to the semantic fields from the Semantic Field Theory [Lehrer, 1974]. Relying on the semantic field idea, the extension of concept definition by adding term fields to the concept seemed reasonable. However, defining which term belongs to the concept field and whether a given term is trend-indicating or not is difficult the more terms

[1] http://protege.stanford.edu/, tool version 3.0-3.3, online accessed 01-July-2013
[2] http://www.w3.org/TR/rdf-schema/ online accessed 01-July-2013

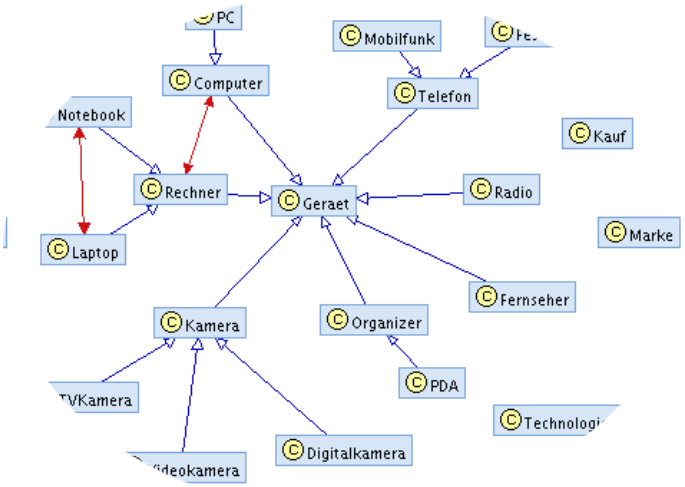


Figure 7.3: Keyword-based trend ontology for market research. Source: *author*.

are used for the term field definition; we searched for the exact definition of trend-indicating features in the texts of market research. Applying statistical methods (e.g. term frequency in documents) supported by manual expertise, we identified adjectives that, according to experts, were significant for a description of customer opinion. The most relevant adjectives were: *vertrauenswürdig* (reliable), *kompetent* (competent), *vielseitig* (all-round), *aktuell* (up-to-date). Conducting the search for semantic fields of these adjectives and their relevance to the main concepts of market research domain, we detected the appearance of so-called satisfier, disatisfier and sensitive[3] categories. Identifiers are adjectives belonging to the concept and describing its features, i.e. an entertainment has entertainment identifier which is described by the adjectives: *abwechslungsreich* (varied), *ansprechend* (attractive), *entspannend* (relaxing), and similar. We defined each main concept as a category with its semantic field and its own identifier that consists of diversificator. Diversificators include the descriptors satisfier, disatisfier and sensitive, which are adjectives grouped by the relevant meaning that refers to the positive, negative and neutral customer's opinion about a given concept. Each identifier consists of a diversificator that refers to more or less positive customer opinion. The customer (dis)satisfaction refers to a negative or positive trend indication. Trend ontology based on term fields adds the meta-level concepts identifier, diversificator, sensitive, satisfier and disatisfier to the keyword-based trend ontology and extends concept sets in term fields.

[3]These terms are used in the marketing Satisfaction Research. In our case we used them to define adjective that express satisfaction or dissatisfaction in language.

7.2.4 (Temporal) invariant scheme based trend ontology

The adjective groups used as satisfier, disatisfier and sensitive are important for the proper sentiment interpretation of a given set of texts. The sentiment interpretation helps with trend detection. However, the validity of diversificators often expires after some time. Assuming that adjectives used for describing customer satisfaction change with time, we looked for an invariant part of trend knowledge. The semi-automatic analysis of relevant market research news done by experts resulted in a structure that seemed to be valid for a long period of time and intuitively applied by experts for analysis of market research texts. This (temporal) invariant scheme based trend ontology consists of three meta-level ontology classes: general, quantification and classification. The general class includes groups of the most important concepts like suppliers and companies. Suppliers, which are important extraction features, are always used in market research projects (regarding our case study) in order to classify the relevance of the texts. The quantification part of our structure contains the idea of identifiers and diversificators, and it adds the amplifier[4] as a new meta-concept. Classification consists of different categories that define the context for the quantifier. Its character is dynamic since it strongly depends on the context at a given point in time. The interesting subcategory of classification is the so-called structure that defines the basic structure for the context. We observed that this category particularly refers to the economic model of the given market. The temporal invariant trend ontology is visualized in Figure 7.4.

7.3 Important issues

We observed the following crucial issues that, in our opinion, need to be considered in further research on trend knowledge modeling.

- Language: Texts relevant for market research include specific language mixed with common words expressing emotional estimation. Customer comments depend on the target group that has been the particular focus of market research studies – even if domain experts often use different descriptions than non-experts, both descriptions have to be considered for trend mining. The synonyms of market research concepts are often hard to describe and have to be weighted since they may exhibit slightly different semantic soundness. The term 'engagement' refers to both involvement and commitment, but engagement in service quality means something different than commitment to service quality Engagement = (Involvement, Commitment, ...). Furthermore, market research studies are conducted in different languages. Semantic relations used for defining concepts dependencies rely

[4] Amplifiers are adverbs, e.g. *sehr* (much), *viel* (many), *wenig* (few), and other similar words.

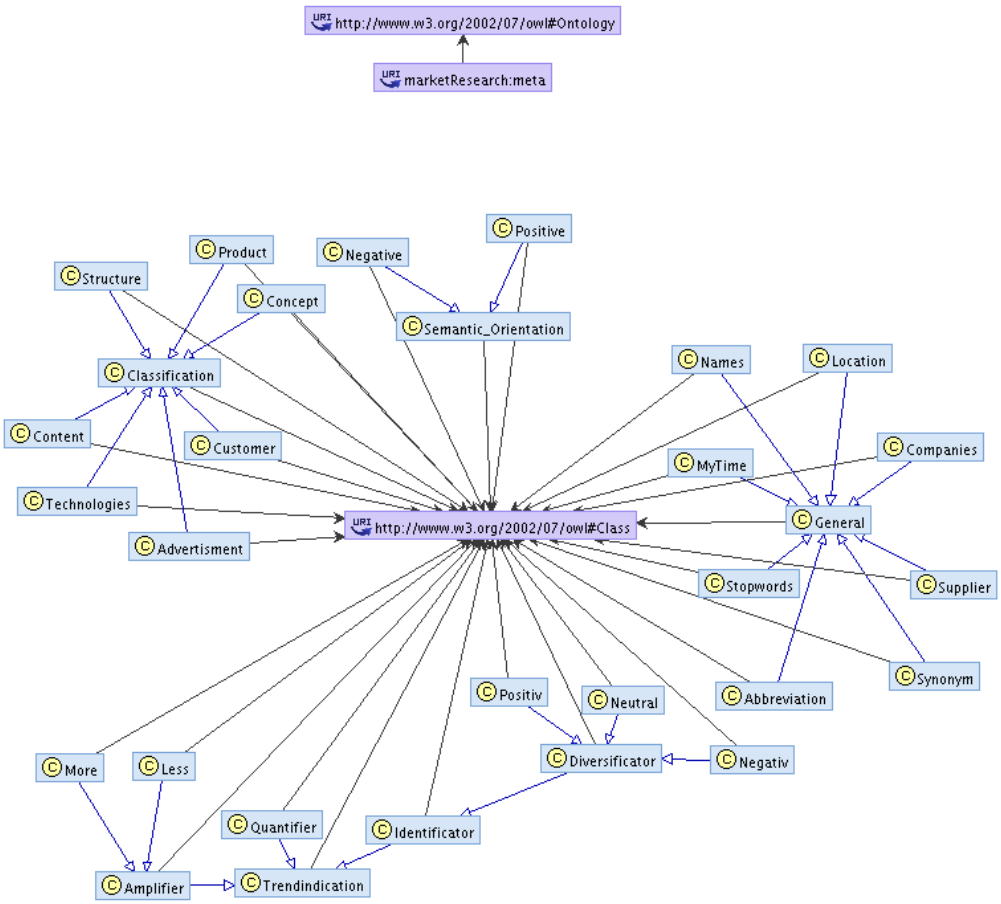


Figure 7.4: Meta level trend ontology for market research. Source: *author*.

strongly on the language used in a trend ontology model. Realizing our trend ontology for the German language, we faced problems such as which relations might be proper for defining a dependency between *Kaufkraft* (buying power/value) and *Kaufentscheidung* (buying decision) in terms of *Stimmungsbild* (the market mood). The modeling of synonyms and the use of relations like included-in or belongs-to, implies the emergence of concept groups rather than taxonomic relations. Trend ontologies have more "fuzzy" structure than ontologies created for more structure fields such as Life Science.

- Time: Concepts used for the modeling of market research ontology and their relevance change in time. There is a need for defining life cycles of categories

modeled in the trend ontology in order to detect if a given instance (i.e. satisfier instance) still belongs to the positive sentiments or if it has drifted to the neutral sentiment; e.g. air bags in a car in the 1980's were used as a positive feature in a car description; however, describing cars nowadays, air bags are an expected feature with a more neutral than positive tone when compared to past decades. Some concepts may fade with time while others can change their meaning or are replaced by a new concept.

- Context: Keywords and terms used in customer comments always depend on context. Regarding the project context, description concepts should refer to the project topic: talking about radio in terms of the Internet may imply services like www.last.fm[5], while radio in the context of a car may imply a concrete hardware device. In texts, the context of sentiments depends on the keywords used for their description. Picking up a concept definition without considering the concept's term field may lead to false conclusions in trend mining.

- Dynamics: Trend ontology covers a very dynamic knowledge. The aspects of time and context affect the ontology structure: meta ontology can be based on the temporal invariant scheme (that is invariant only for a given time period); middle ontology depends on market research topics and must be adapted for every new study; the lowest level of trend ontology is the most dynamic one. Modeling the trend knowledge aspect of dynamics should be considered from both the knowledge level (concepts and their meaning are changing over time) and the abstract level (in terms of knowledge formalization).

- A trend structure: Even if we know that the trend-indicating keywords and concepts are changing in time, and that their positive or negative value differs and depends on the context, we assume that there is an invariant trend structure which contains the three main trend detection parts: general concepts, the trend value concepts, and the classification structure that models the context of the trend.

7.4 Knowledge discovery or search problem?

The section above contains the description of our experience on modeling a trend ontology in an application oriented use case, the case of mining trends in market studies. In Section 7.1.1, we discussed trend mining in market studies as a process of knowledge discovery, based on different steps including classification, filtering, and aggregation of information. In Figure 7.2, we limited the trend

[5] www.last.fm accessed 23-Jul-2013

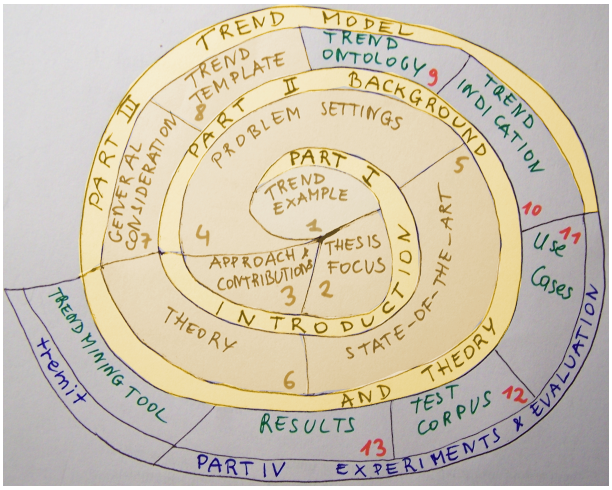
mining process and focused on the classification process that consists of feature extraction, selection, and learning – the general steps in clustering or classification of data. In Chapter 6, we described in general the possibilities of understanding trend mining in terms of information retrieval process. While mining trends, we moreover assume that the information that we are about to mine is most probably containing the trend(s), similar to the case of mining data where we assume that we have valuable data to mine. It is useful to reflect here on the question of whether trend mining in texts is a knowledge discovery or a search problem.

Regarding a given web document corpus that we want to mine for trends, we have at least two different possibilities for how to approach the trend mining task. Taking the search problem perspective, the query is the emerging topic area (the trend) and the answer to the query is the part of the web corpus with similar content to the content of the query (the documents that contain the trend). However, if we already know what the trend is, since we ask for it in the corpus, we do not need to look for it. Here one ends in the famous Meno - Socrates discussion:
"Meno: And how will you enquire, Socrates, into that which you do not know? What will you put forth as the subject of enquiry? And if you find what you want, how will you ever know that this is the thing which you did not know?
Socrates: I know, Meno, what you mean; but see what a tiresome dispute you are introducing. You argue that man cannot enquire either about that which he knows, or about that which he does not know; for if he knows, he has no need to enquire; and if not, he cannot; for he does not know the very subject about which he is to enquire"([Jowett, 1949], re-cited from p. 5, [Witten et al., 2007]).

From the AI learning perspective, the emerging topic area allows for deriving the topic-in-time labels for documents. Regarding the classification approaches, classification within a document corpus aims to estimate whether a document belongs to a given class C1 or C2 (if only these two classes are predefined for the corpus). In this case, the documents are instances that have to be learnt and predefined topics are the class labels to which the documents belong. Bringing this approach to the trend mining problem, the goal of such classification in trend mining would be to estimate if the web document belongs to the given "topic-in-time", discussing an interesting topic that is emerging in interest and utility in other documents and represents the emerging trend. But in this case, we must already have the training instances, the topic-in-time labels in order to perform the classification correctly. Regarding the unsupervised (clustering) learning approach we just have to cluster the documents according to the topic-in-time features they contain in order to discover knowledge.

In the experimental part of our research, trend mining is being handled as a knowledge discovery process that starts with a search. In fact, we always know some aspect of the trend while looking for it – we search for a *specific* term in

order to retrieve valuable results. Looking for the *financial crisis* as a trend, we may want to have information regarding a specific stock or company or to see how the keyword *crisis* is changing its context over time. We shrink the search space starting with a specific term and obtain this part of the web document corpus, which contains the particular term or topic with the context around it – in our case we are applying our trend model, described in Chapter 8, to the corpus of financial news containing information relevant to DAX (see more description in Chapter 12).



CHAPTER 8

Trend template

Having learned from the previous chapter about the general important issues in knowledge-based trend mining, we introduce in this chapter our knowledge-based trend model starting with the definition of a trend template. In the beginning of this thesis, we discussed different trend definitions given by the literature and extracted the general trend definition that underlies this thesis. This chapter formalizes this general trend definition integrating lessons learned from previous chapters. We describe here that a trend is an emerging topic area that grows in interest and utility over time, and that it can be described by the following features: trigger, context, relation, time interval and amplitude.

8.1 Trend template

In Section 4.2.3 we describe a trend as a topic area that is growing in interest and utility over time. From Chapter 4 we learn about the different views on a trend, including the sociological trend diamond model (see Figure 4.1) from which we learn about the different stages of a trend, based on the people involved in the trend process: trend creators, setters, followers, mainstreamers, late mainstreamers and conservatives. From Section 4.1.3 we conclude that a trend in documents is an emerging topic area and that it sometimes can be triggered by an event (see 5.1). Based on our preliminary research experiments and considerations presented in Chapter 7, we learn that a trend always appears under a time constraint, in a given context. In Section 8.1.1, we summarize these conclusions as assumptions about a trend and give an abstract model of a trend by formalizing our assumptions in Section 8.1.3. The content of this chapter is the basis for the implementations presented in Chapter 9 and Chapter 10 and it has been mainly published in [Streibel et al., 2013b].

8.1.1 Assumptions

In the context of our research, a trend can be described by the following features:

- trigger
- context
- amplitude
- time interval
- relation

Moreover, we assume that these features can be extracted from a document corpus and that they are basis of the so-called trend structure (see Section 7.3).

8.1.2 Definitions

A trigger is a *thing*. Triggers may be an event, a person, or a topic: anything that triggers a trend. A trigger can but does not have to cause a trend. A trigger makes the trend visible. An example of a trigger is *Lehman Brothers insolvency* that can be classified as both a topic and an event.

Context is the area, also called the domain of the trigger. If the trigger is a topic then the context is this topic's area, e.g. *Lehman Brothers insolvency* is in the context of *real estate market*.

Amplitude is the strength of given trend. It can be expressed by a number, the higher the number, the more impact the trend has. It may also be expressed by a qualitative value that describes the trend phase, e.g. beginning (setter), emerging (follower), mainstream, fading (conservative).

Time is a necessary dimension while spotting a trend, since there can be no trend without time. It is the interval in which the trend is appearing, independent from the amplitude, e.g. the *real estate crisis* appeared in the years 2008-2011.

Relation expresses the dependency between a trigger and a context, it puts the given trigger, e.g. *Lehman Brothers insolvency* within the given context of the *real estate crisis* in a relation, e.g. *Lehman Brothers insolvency is part of the real estate crisis*.

An abstract conceptualization of the proposed trend definition is shown in Figure 8.1.2. It shows that a trend is defined by a structure that we call a trend

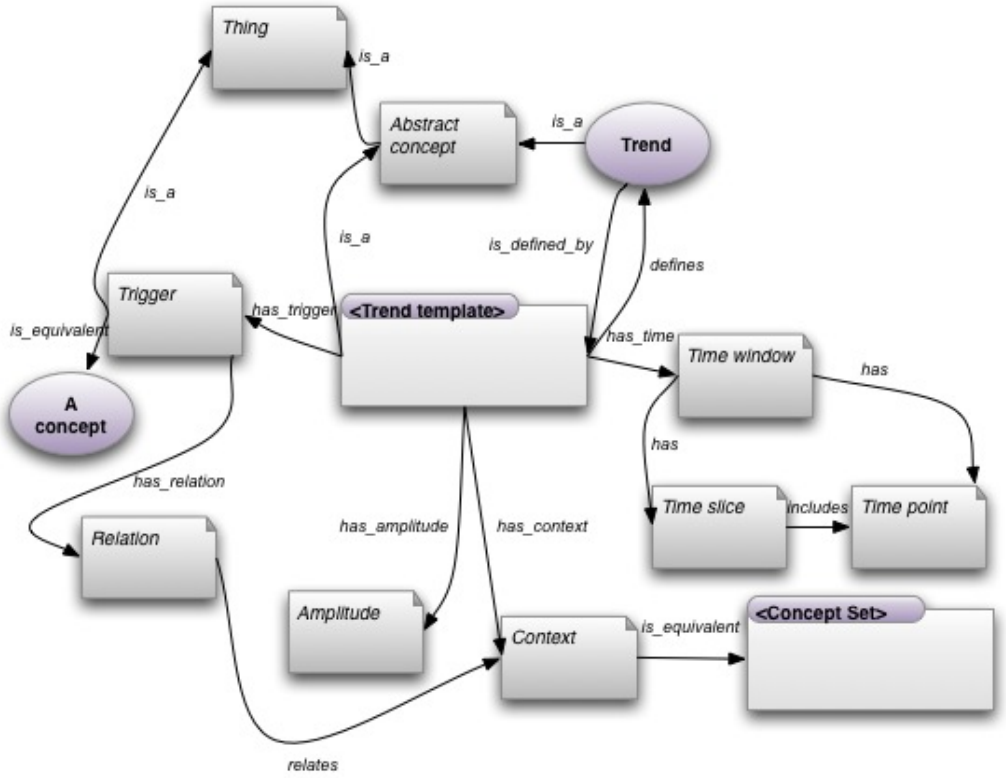


Figure 8.1: An abstract conceptualization of the trend template. Source: *author*

.

template. The trend template consists of the main characteristics as defined above.

8.1.3 Formal description

The trend template is an abstract model that describes the main concepts that are important and necessary for knowledge-based trend mining. In the following, we give the explicit definition of the trend template.

DEF. 8.1.1: Trend template (TT) is a quintuple:

$$TT := \langle T, C, R, TW, A \rangle \tag{8.1}$$

where: T is the trigger, C is the context, R is the relation, TW is the time window, and A is the amplitude.

DEF. 8.1.2: T = Trigger is a set of concepts:

$$T := \{t_0, \ldots, t_n\}, n \in \mathbb{N} \wedge t \in T \tag{8.2}$$

so that if E, P, T_o are the sets defining:
Events:

$$E := \{e_0, \ldots, e_n\}, n \in \mathbb{N} \wedge e \in E \tag{8.3}$$

Persons:

$$P := \{p_0, \ldots, p_n\}, n \in \mathbb{N} \wedge p \in P \tag{8.4}$$

Locations:

$$L := \{l_0, \ldots, l_n\}, n \in \mathbb{N} \wedge l \in L \tag{8.5}$$

Topics:

$$T_o := \{t_{o0}, \ldots, t_{on}\}, n \in \mathbb{N} \wedge t_o \in T_o \tag{8.6}$$

then:

$$T := E \cup P \cup T_o \cup L \tag{8.7}$$

DEF. 8.1.3: C = Context is a union set consiting of a set of concepts and a set of relations between them where c is a context element:

$$C := C_{co} \cup R_{co}, c \in C \tag{8.8}$$

with C_{co} the set of concepts

$$C_{co} := \{c_{co0}, \ldots, c_{con}\}, n \in N \wedge c_{co} \in C_{co} \tag{8.9}$$

and R_{co} the set of relations:

$$R_{co} := \{r_{co0}, \ldots, rco_n\}, n \in N \wedge r_{co} \in R_{co} \wedge R_{co} \subseteq C_{co} \times C_{co} \tag{8.10}$$

whereas r_{co} defines a binary relation:

$$r_{co} : c_{cox}, c_{coy} \longrightarrow r_{co}(c_{cox}, c_{coy}) \wedge c_{cox} \neq c_{coy} \tag{8.11}$$

and the context element is defined by:

$$c = c_{co} \cup (c_{coi} c_{coj}), C = C_{co} \cup C_{co} \times C_{co} \tag{8.12}$$

DEF. 8.1.4: R = Relational is a set of relations:

$$R := \{r_0, \ldots, r_n\}, n \in N \wedge r \in R \wedge R := \{T \times C\} \tag{8.13}$$

with

$$r_i : t_i, c_i \longrightarrow r_i(t_i, c_i) \tag{8.14}$$

DEF. 8.1.5: TW = Time window is a function that assigns time slices to the time points:

$$TP := \{t_{point} | t_{point} = second \vee minute \vee hour \vee day \vee month \vee year\} \tag{8.15}$$

$$TS := \langle t_{point0} \dots t_{pointn} \rangle \tag{8.16}$$

$$TW : TP \longrightarrow TS \tag{8.17}$$

DEF. 8.1.6: A = Amplitude is a function that assigns a value to the quadrupel of $\langle T, C, R, TW \rangle$

$$A : T \times C \times R \times TW \longrightarrow \mathbb{R} \cup V \tag{8.18}$$

where $\mathbb{R}$ is the set of real values and V is the set of categorial values

$$a : (t,c,r,tw) \longrightarrow r_{value} \vee v \tag{8.19}$$

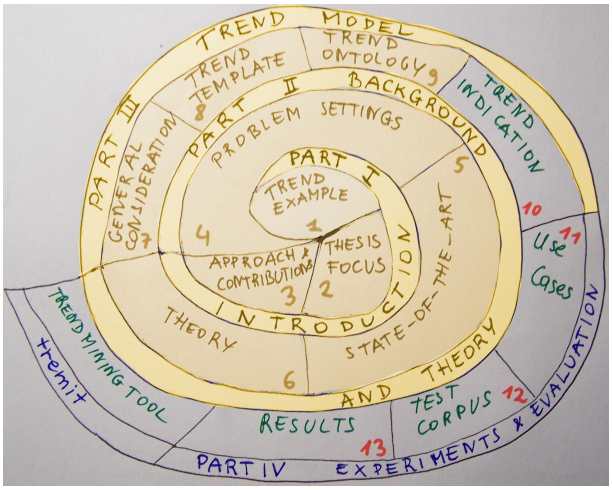
8.2 Trend probability

Understanding a template intuitively leads to the conclusion that the template is something that should be filled out. If we regard the trend template as the basis for knowledge-based trend mining algorithms, we can proceed as follows: the trend template can be filled as far as possible with fitting parts of the document set that has to be analyzed. Yet, having some of the parts filled, we need to estimate and to learn others. Furthermore, we need to somehow estimate the sustainability of the recognized trend that is described by the given trend template, to estimate the probability that the trend will continue. For these reasons, we propose to apply Bayes' formula.

Our general trend estimation method is then:
TT is a trend template that can be found on the given document corpus D. T_D is in general a trend in document set D that is being described by the trend template TT. $P(T_D|TT)$ is the a posteriori probability that TT describes the sustainable trend T_D in D, ergo the $P(T_D|TT)$ estimates the probability that, given TT, we found a sustainable trend T_D in D. $P(T_D|TT)$ is the posterior probability of T in D conditioned on TT. In general, $P(T_D|TT)$ can be calculated as follows using Bayes' theorem:

$$P(T_D|TT) = \frac{P(TT|T_D)P(T_D)}{P(TT)} \tag{8.20}$$

where $P(T_D)$ is the prior probability of a trend T in D and it expresses that any trend in the document set is sustainable. Similarly, $P(TT)$ is the prior probability of the trend template TT and $P(TT|T_D)$ is the posterior probability of TT conditioned on T_D, which says that a sustainable trend T in D is described by the trend template TT. However, the detailed description of the method for the estimation of the probability for the proposed abstract trend model should be better elaborated with the naïve Bayesian classifier and will be the subject of future work. We are not extending this idea in the context of this thesis.



CHAPTER 9

Trend ontology

The previous chapter explained the idea of the trend template. This chapter about trend ontology shows the implementation of a trend template as a trend ontology. It describes concepts, relations and functionality of the trend ontology.

9.1 Trend ontology – general idea

One way of implementing the trend template is the realization of this model in form of an ontology. We can understand the ontology as an instance of the trend template or as its simplified and application oriented implementation. In Section 6.6 of Chapter 6, we discussed what an ontology is and gave an example of modeling the ontology. Based on the trend template presented in previous chapter, we created an applicable model, using given concepts and properties from RDFS/OWL[1] and SKOS[2]. Our model serves as a general model that can be extended regarding the particular application domain and applied for annotating a text corpus in order to retrieve the trend structure out of it. The trend ontology is divided in three levels: meta, middle and low that corresponds to three abstract layers of the model (see Figure 9.1). Whereas the low level and the middle level layers relate to the corresponding application domain (in our case it is the German Stock Exchange, DAX), the meta level is the most interesting one. Meta ontology incorporates the general trend features and can be applied to any application domain. In the following, we focus on its structure. The content of this chapter is

[1] http://www.w3.org/TR/owl-features/ accessed online 01-June-2013
[2] http://www.w3.org/2004/02/skos/ accessed online 01-June-2013

the basis for the experiments with trend ontology presented in Chapter 13 and it has been mainly published in [Wißler and Streibel, 2012].

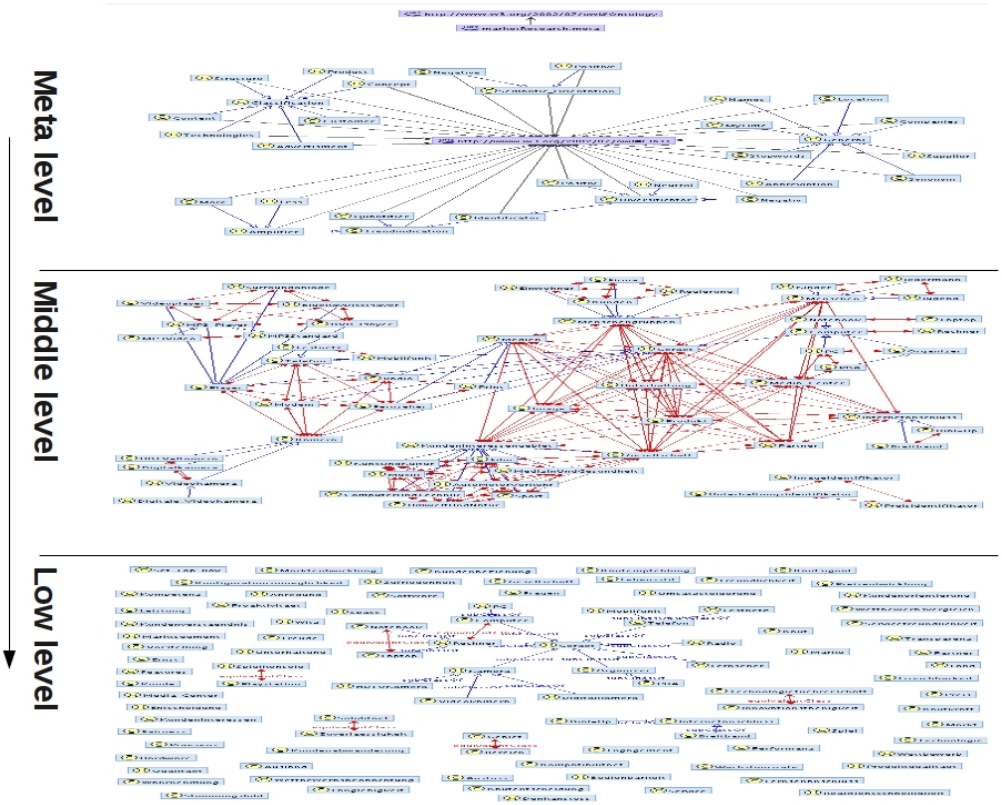


Figure 9.1: The visualization of the 3 levels trend ontology. Source: *author*.

9.2 Meta ontology

The meta ontology is based on other existing ontologies (see the imports shown in Listing 9.1) in order to enhance interoperability and simplify modular domain ontology design.

The central concepts of the ontology are *Trigger*, *TriggerCollection*, *Context*, and *Indication*. The concepts mirror the composition of the trend template. *Trigger* consists of the subconcepts: event, person, location. Concepts: topic and group are . A snippet from the meta ontology in N3 is shown below:

```
@prefix skos:<http://www.w3.org/2004/02/skos/core>
@prefix lode:<http://linkedevents.org/ontology/>
```

```
# ...
@prefic foaf:<http://xmlns.com/foaf/0.1/>
# ...
# trigger
Trigger a owl:Class;
rdfs:subClassOf (
skos:Concept
time:TemporalEntity )
# ...
# event
Event a owl:Class;
rdfs:subClassOf (
 :Trigger;
 lode:Event )
# ...
# person
Person a owl:Class;
rdfs:subClassOf (
 :Trigger;
 foaf:Person )
# ...
```

Listing 9.1: Trend ontology, a fragment.

```
<rdf:RDF
xmlns:skos="http://www.w3.org/2004/02/skos/core#"
xmlns:time="http://www.w3.org/2006/time#"
xmlns:rdf="http://www.w3.org/1999/02/22−rdf−syntax−ns#"
xmlns:foaf="http://xmlns.com/foaf/0.1/"
xmlns:owl="http://www.w3.org/2002/07/owl#"
xmlns:rel="https://sites.google.com/site/trendontology/relation"
xmlns:rdfs="http://www.w3.org/2000/01/rdf−schema#"
xmlns:lode="http://linkedevents.org/ontology/"
xmlns:vann="http://purl.org/vocab/vann/"
xmlns:dcterm="http://purl.org/dc/terms/"
xmlns:dctype="http://purl.org/dc/dcmitype/"
xml:base="https://sites.google.com/site/trendontology/trendontology">
<dc:title xml:lang="en">TRENDO: A trend ontology (META)</dc:title>
```

The Listings 9.1 and 9.2 present more details of the meta trend ontology.

Listing 9.2: Trend ontology, a fragment.

```
<!−− Trigger −−>
```

```
<owl:Class rdf:about="#Trigger">
<dcterm:issued rdf:datatype="http://www.w3.org/2001/XMLSchema#date">
  2012-04-15</dcterm:issued>
<rdfs:comment xml:lang="en">
Trigger is a thing. Examples of triggers are: an
event, a person, a group, or a topic- anything that triggers the
trend. A trigger can but does not have to cause a
trend. A trigger makes the trend visible. An example
of a trigger is Lehman Brothers' (a name of a company from 2007) insolvency that can
be classified as both: a topic and an event. Trigger is anything that can trigger a trend:
   a person or group, an event, location,
or a topic. It is defined as a subclass of skos/core/concept - an idea or notion; a unit
  of thought.
</rdfs:comment>
   <rdfs:subClassOf rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/
     >
    <rdfs:subClassOf rdf:resource="http://www.w3.org/2006/time#TemporalEntity"/
      >
<owl:Restriction>
        <owl:onProperty>
          <owl:ObjectProperty rdf:about="#keyword"/>
        </owl:onProperty>
        <owl:someValuesFrom>
          <owl:Class rdf:about="#Keyword"/>
        </owl:someValuesFrom>
      </owl:Restriction>
    </rdfs:subClassOf>
    </owl:Class>
<!-- Event, person, group, location and topic are the subclasses of trigger -->
<owl:Class rdf:about="#Event">
<dcterm:issued rdf:datatype="http://www.w3.org/2001/XMLSchema#date">
  2012-04-15</dcterm:issued>
<rdfs:comment xml:lang="en">
An event is a happening of something as might be reported on the news.
</rdfs:comment>
<rdfs:subClassOf rdf:resource="https://sites.google.com/site/trendontology/
  trendontology/#Trigger">
<rdfs:subClassOf rdf:resource="http://linkedevents.org/ontology/#Event">
</rdfs:subClassOf>
</owl>
<owl:Class rdf:about="#Person">
<dcterm:issued rdf:datatype="http://www.w3.org/2001/XMLSchema#date">
  2012-04-15</dcterm:issued>
<rdfs:comment xml:lang="en">
The person is an important person or relevant person in the context of the given
  domain for which trend should be detected. In the terms
of sociologist viewpoint on trends, it could be the trend setter or the early adopter.
```

```
40 </rdfs:comment>
41 <rdfs:subClassOf rdf:resource="https://sites.google.com/site/trendontology/
     trendontology/#Trigger">
42 <rdfs:subClassOf rdf:resource="http://xmlns.com/foaf/0.1/Person">
43 </rdfs:subClassOf>
44 </owl>
```

9.3 Relational

In the case of the trend ontology for the financial market, we extended the meta ontology by an additional ontology that allows the definition of so-called dynamic relations, *Relational*. The relational ontology consists of two ontology classes *RelationalThing* and *RelatingThing*, as well as a relation *scalableRelated*. A RelationalThing is something that can be connected, and a RelatingThing is something that connects.

An example of the relating expression from the trend ontology is " The oil price has a strong influence on the gasoline price". Hereby, the *oil price* and the *gasoline price* are concepts belonging to RelationalThings. The *strong influence* belongs to the RelatingThing. The relation *scalableRelated* allows for connecting the RelationalThing with the RelatingThing.

Listing 9.3: Trend ontology, a snippet from Relational.owl.

```
1  <owl:Class rdf:ID="RelationalThing">
2    <rdfs:subClassOf rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
3    <rdfs:subClassOf>
4      <owl:Restriction>
5        <owl:allValuesFrom>
6          <owl:Class rdf:ID="RelatingThing"/>
7        </owl:allValuesFrom>
8        <owl:onProperty>
9          <owl:ObjectProperty rdf:ID="scalableRelation"/>
10       </owl:onProperty>
11     </owl:Restriction>
12   </rdfs:subClassOf>
13 </owl:Class>
14 <owl:Class rdf:about="#RelatingThing">
15   <rdfs:subClassOf rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
16   <rdfs:subClassOf>
17     <owl:Restriction>
18       <owl:onProperty>
19         <owl:ObjectProperty rdf:about="#scalableRelation"/>
20       </owl:onProperty>
21       <owl:someValuesFrom rdf:resource="#RelationalThing"/>
22     </owl:Restriction>
```

```
23      </rdfs:subClassOf>
24    </owl:Class>
25    <owl:ObjectProperty rdf:about="#scalableRelation">
26      <rdfs:domain>
27        <owl:Class>
28          <owl:unionOf rdf:parseType="Collection">
29            <owl:Class rdf:about="#RelationalThing"/>
30            <owl:Class rdf:about="#RelatingThing"/>
31          </owl:unionOf>
32        </owl:Class>
33      </rdfs:domain>
34      <rdfs:range>
35        <owl:Class>
36          <owl:unionOf rdf:parseType="Collection">
37            <owl:Class rdf:about="#RelationalThing"/>
38            <owl:Class rdf:about="#RelatingThing"/>
39          </owl:unionOf>
40        </owl:Class>
41      </rdfs:range>
42    </owl:ObjectProperty>
```

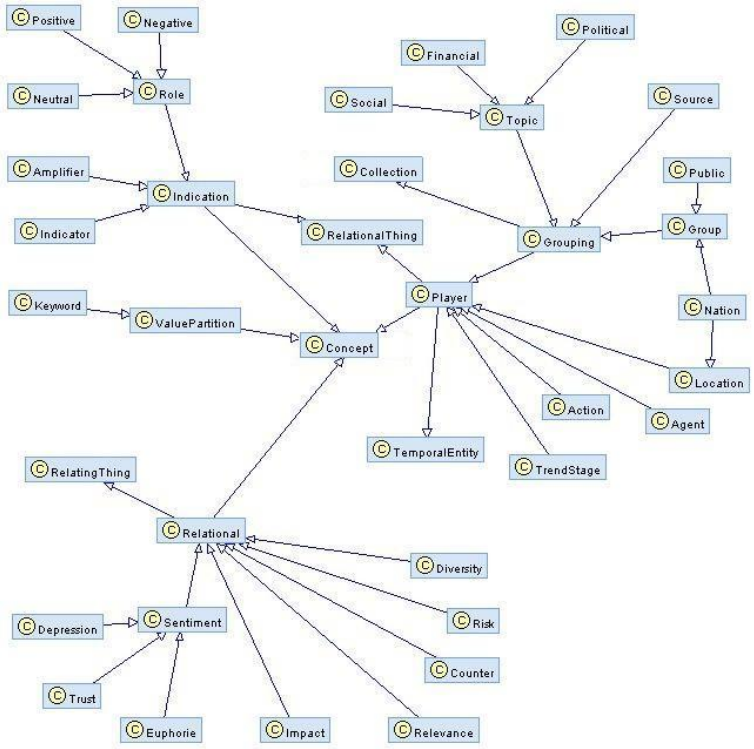


Figure 9.2: The applied meta ontology. Source: [Wißler and Streibel, 2012].

9.4 Applying the meta ontology

The trend ontology allows for the parsing of a document corpus. The ontology is filled by the respective terms from the documents, classifies them into the concepts, captures the relations between them and therefore creates a meaningful structure of concepts from the corpus.

The application of the ontology to the particular test corpus requires some adjustments to the specifics of the given corpus, such as omitting the corpus specific stop words, adding some corpus specific relations. Also the respective concepts can be completed by some use case dependent sub-concepts. In Figure 9.2 we visualize the meta level ontology applied on our test corpus.

Which relations are used in the specific case of a particular corpus depends on the application scenario. Three basic relations of our ontology allow already for the extraction of useful meaning for further trend analysis and can be extended on demand:

- **skos:related:** manually created relation
- **rel:countableRelated:** statistic relation, counts the appearance of two concepts together.
- **skos:member:** semantic relation, shows through the integration in topic in which topics a given concept emerges

9.4.1 Topic categories

While analyzing the test corpus we focused on the different categories of topics that emerge from the documents. In particular, the web news about the different companies and their stocks on the market offer a good opportunity for the analysis of the specific emerging topics. We divided the news of the test corpus into groups that are listed below.

- **BusinessVolume:** information about the technical market position of companies such as sales news, quarterly reports, debts
  - **Topic creating classes:** sales, company
- **ConcernNews:** information about corporations such as their market position, customer development, management board changes
  - **Topic creating classes:** concern_Type, company
- **FinancialEvent:** events relevant for the stocks such as bankruptcy, natural disasters, war

  - **Sub-classes:** insolvency (keyword: Insolvenz), takeover (keyword: Uebernahme)

  - **topic creating classes:** location

- **MarketNews:** general market reports, economic downturn, market sentiment, crises, bubbles

  - **topic creating classes:** group, political, sentiment

- **Recommendation:** investment recommendation

  - **topic creating classes:** analyst, share, financial_Instrument

- **ShareNews:** general news about stocks

  - **topic creating classes:** company, share

- **SharePriceChange:** specific news about stock estimation

  - **Sub-class:** target

  - **topic creating classes:** share, share_factor

The documents from the test corpus can be classified by their keywords into the groups listed above. The respective combination of the concepts from a given document determines the topic groups to which the document belongs.

9.5 Algorithm

The Pseudocode 9.5.1 illustrates our *trendDescription* method.

Algorithm 9.5.1: CREATETRENDDESCRIPTION(c,o)

comment: parse $\forall$ **d**ocument $\in$ **c**orpus

comment: into **o**ntology

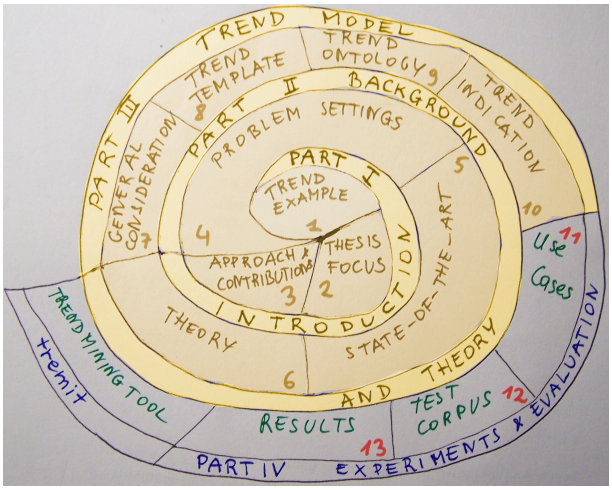
$parse(c, inO, outO)\{$
$model.read(inO)$
$create.reasoner(inO)$

do {
$parse(keywords)$;
$match.model(keywords,inO)\{$
for $keyword \leftarrow 0$ **to** i
if $inO.concept.label==keyword$ **or**

$keyword.prefix$ **or** $keyword.postfix==$
$inO.concept.label.prefix$ **or** $.postfix$
then $matches.add(keyword)\}$

$relate.model(matches, inO)\{$
if $model.getRelation(matches).isEmpty$
then $model.createRelation(matches)$
else $model.incCounter(matches)\}\}$

$model.write(outO)\}$



CHAPTER **10**

Trend indication

We have learned one possibility of trend template interpretation in the last chapter. This chapter focuses on selection and extraction of trend indicating features from text corpus, showing that the knowledge inspired model can be realized in a statistical way. We learn here the weighting methods for trend features and discuss the possibilities of knowledge integration into the feature extraction process.

10.1 Preliminaries

In the following we describe the definitions of outliers, interestingness, utility and trend indication that are used later in the definitions of topic, topic area, emerging topic and emerging topic area. The content of this chapter is the basis for the experiments with the weighting functions presented in Chapter 13. The trend indication functions and the trend estimation algorithm have been mainly published in [Streibel and Alnemr, 2011] in combination with the reputation approach. The reputation approach, presented more detailed in [Alnemr and Meinel, 2011], is a very useful extension to the trend estimation approach but it lies outside of our thesis' focus. The methods explained in this chapter are based on the use case described in 11.1.1.

10.2 Definitions

Continuing with the example of informing oneself by reading users' posts from Twitter reporting on unrests in Egypt from January to February 2011 (see also the example in Chapter 1 and the use case description in 11.1.1), we introduce our trend estimation approach. According to the general definition as provided

by related research on trend mining, trends in texts are defined as emerging topic areas and an emerging topic in texts is a topic that "increases in interest and utility over time"[Kontostathis et al., 2003].
In order to estimate a trend we define in the following a time window, outliers, interestingness, utility and the trend indication.

Definition 10.2.1. ***Time window and time slice***
t_{window} is a time interval in which trends can occur. A day is an example of time window.
t_{slice} is a subinterval of time window. If its starting point lies at t_0, the end point has to lie at $t_k < t_n$. Regarding a day as an example of time window, an hour is an example of a time slice.

$$t_{window} = [t_0...t_n] \wedge t_{slicek} = [t_0...t_k]$$
$$t_{window} :=< t_{slicek},...,t_{slicen} >$$
$$|t_{slicek}| = |t_{slicen}| \wedge k,n \in \mathbb{N} \wedge k < n \tag{10.1}$$

Definition 10.2.2. ***Outliers***
An outlier is a term that appears, compared with the whole time window, significantly often in a given time slice. An outlier as a value can be determined for every term by calculating:

$$outlier(w)_{t_{slice}} := TF_{(w,|P|_{t_{slice}})} * IPF_{(w,|P|_{t_{window}})} \tag{10.2}$$
$$IPF_{(w,|P|_{t_{window}})} := \log \frac{|P|_{t_{window}}}{PF(w)_{t_{window}}}$$

whereas: $TF_{(w,|P|_{t_{slice}})}$ says how frequent a term w appears in the posts of particular time slice[1]*. $|P|$ expresses the total number of given posts. $IPF_{(w,|P|_{t_{window}})}$ determines the appearance of the particular term w in the whole window.*

If we consider the beginning of the reports on Egyptian unrests on 25th of January and the $t_{window} = day$,$t_{slice} = hour$, the terms as '#jan25', 'Egypt', 'revolution' were outliers on this day and would have the most significant outlier values among other terms of chosen time window.

Definition 10.2.3. ***Interestingness***

$$interest(w)_{t_{slice}} = f(w)_{t_{slice}} := \log \frac{TF_{(w,|P|_{t_{slice}})}}{|W|_{t_{slice}}} \tag{10.3}$$

where $|W|$ is the number of all terms considered.

$$interest(w)_{t_{window}} :=$$

[1] the calculation is based on the principle of the weighting method TFIDF[Salton et al., 1982] by including time as an calculation dimension

$$< f(w)_{t_{slice}k}, f(w)_{t_{slice}k+1}, \ldots, f(w)_{t_{slice}n} > \tag{10.4}$$

expresses increasing interest if:

$$f(w)_{t_{slice}k} < f(w)_{t_{slice}k+1} < \ldots < f(w)_{t_{slice}n}$$

The interest values of our example terms "#jan25", "Egypt", "revolution" were constantly increasing over the time slices of time windows beginning from January 25th.

Definition 10.2.4. ***Utility***
The utility can be described by the number of resources that, in a given time window, have been tagged with the term w divided by the number of all resources tagged in this time window:

$$util(w)_{t_{window}} := \log \frac{|R|_{(tag=w)t_{window}}}{|R|_{(tag)t_{window}}} \tag{10.5}$$

where $|R|$ is the number of resources (posts, status messages, tweets) in the given system.
Similar to the interestingness, we can identify increasing and decreasing utility for every term while looking slice for slice in whole time window, of the utility value increase or decrease.

Regarding our example, the utility values of '#jan25', 'Egypt' were significantly high since more and more messages were tagged with these terms.

Definition 10.2.5. ***Trend indication***
Only terms with significant high outlier values have to be considered as trend indicating. Terms with their outlier values below a certain threshold can be omitted.

$$trendind(w)t_{window} = \frac{interest(w)t_{slice} * util(w)t_{window}}{ratio(t_{window})} \tag{10.6}$$

whereas:

$$ratio(t_{window}) = |t_{window}|$$

is the size of time window given by the number of its time slices.

A feature in texts (e.g. a term or a term pair) is *trend indicating* if:

a) it has a significant outlier value and

b) its interest and utility values, in relation to the frequency of the time window, are increasing

Let the $\mathbb{C}$ be a category and S topic (subject). Further, let c describe concept, let t describe time, t_{slice} time slice, t_{window} time window, f frequency, u utility, and w word. Further, let call an arbitrary tagging system a TS.

Definition 10.2.6. ***Topic***
A topic is a set of words from a vocabulary.

Definition 10.2.7. ***Emerging topic***
An emerging topic in a given time window is the set of all trend indicating words in this time window.

Definition 10.2.8. ***Topic area***
A topic area in a given time slice is the intersection of the subset of all words that appear frequently enough to be detected and rare enough to be important in given time slice with the set of words used as tags (e.g. Twitter's hashtags) in a TS in this time slice.
Let call $w_{t_{slice}} \in \mathbf{W_{t_{slice}}}$ *the set of all words in a given time slice and* $w_{(d,t_{slice})} \in \mathbf{W_{(d,t_{slice})}} \wedge d \in D$ *the set of words in a document in given time slice.*

Definition 10.2.9. ***Emerging topic area***
Let $w_{t_{slice}} \in \mathbb{W}_{t_{slice}}$ *where* $\mathbb{W}_{t_{slice}}$ *be the set of all words in a given time slice.* $w_{(d_k,t_{slice})} \in \mathbb{W}^{d_k}_{t_{slice}}$ *where* $\mathbb{W}^{d_k}_{t_{slice}}$ *is the set of words in a document k in given time slice with* d_k *as an arbitrary document in given time slice* $d_k \in D$.
$O_{d_k,t_{slice}}$ *is the set of all outlying and interesting words in an arbitrary document of given time slice* $w^{outlier \wedge interest}_{d_k,t_{slice}} \in O$ *whereas* $O_{d_k,t_{slice}} \subseteq \mathbb{W}^{d_k}_{t_{slice}}$ *and emerging topic area is the conjunction of all:*
$O_{d_k,t_{slice}} \cup O_{d_{k+1},t_{slice}} \cup \ldots \cup O_{d_n,t_{slice}} \cup TS^{tag}_{t_{slice}}$ *Emerging topic area in given time slice is the conjunction of all subsets of all outlying and interesting words in all documents of this time slice with the chosen set of words used as tags in a TS in this time slice.*

10.3 Trend estimation

The stepwise weighting method based on functions as presented above allows us to select the trend features out of a given text corpus in a chosen time window. Trend features are in particular, the trend indicating terms in texts. Regarding the trend template defined in Chapter 8, the terms with certain outlier values that are of increasing interestingness value can be understood as the triggers. The trend indicating terms from the given text corpus can be understood as the context of the triggers.

The particular terms of high trend indicating values are, if treated separately, without any meaning. Only if we know that 'Cairo' is a capital of 'Egypt', 'Tahrir

Square' is a place in Cairo, we can conclude that these terms are also semantically connected and probably describe one particular place.

In order to verify any connections between the trend indicating terms, we use knowledge from an ontology. A trend ontology as presented in Chapter 7 or any ontology that describes the domain in which we are looking for trends is applicable for enhancing the calculated terms (in our example: political domain). In our example, we propose first to look up in Dbpedia[2] since it is the most popular source of structured knowledge on the web.

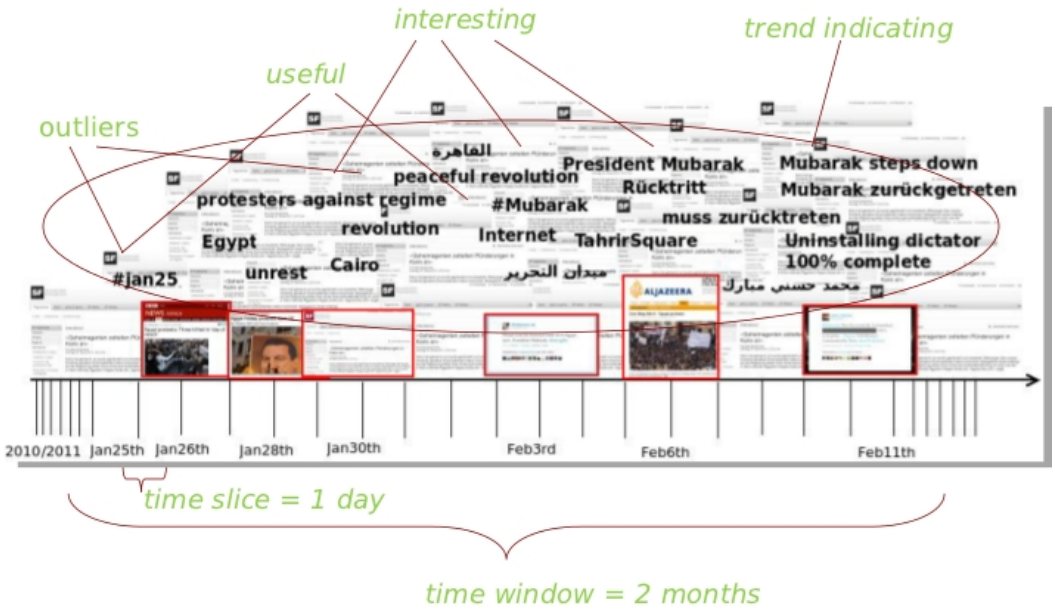


Figure 10.1: Trend indicating keywords. Source: *author*.

Figure 10.1 shows a timeline-based visualization of selected reports on protests in North Africa in January to February 2011 with reference to the features that we look for: outliers, interesting words, useful words, trend indication.

The Pseudocode 10.3.1 describes the general algorithm for selection of trend features from status messages *stMessages*, and tweets *twMessages*.

[2] `http://dbpedia.org` accessed 01-June-2013

Algorithm 10.3.1: selectTrendFeatures($stMessages, twMessages$)

comment: PRE: stMessages, twMessages

comment: POST: tMessages, noMessages, noPeers

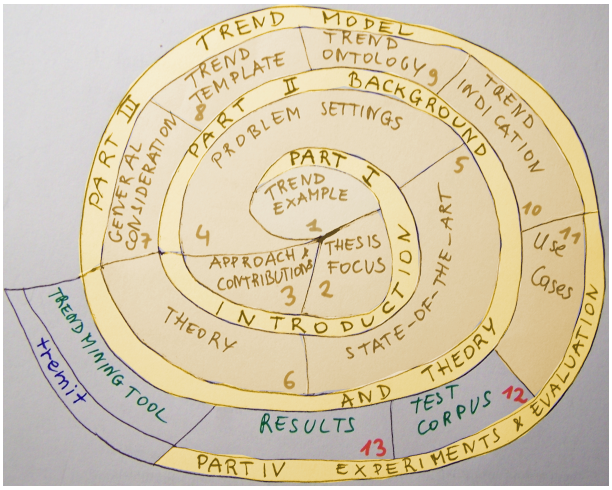
$PREPROCESS:$
$findBestTimeSlice(PARAM : timewindow)$;
$parsMessage()$;
$calculateStopWordList(PARAM : timewindow)$;
for each $stMessage$,$twMessage$
$removeStopwords()$;
$stemm()$;
$tokenize()$;
$tsVectors = createTimeStampedMessageVectors()$;
$END\_PREPROCESS$;

$TREND\_FEATURE\_SELECTION:$
for each $term \in tsVectors$
$calculateOutlierV()$;
$calculateInterestingnessV()$;
$calculateUtilityV()$;
for each $term \in tsVectors$
if $(term.OutlierV > treshold \&\&$
$term.InterestingnessV \leftarrow upper \in tsVectors \&\&$
$term.UtilityV > threshold)$
$calculateTrendindication(term)$;
$tTermList = addTermToTrendingList()$;
$END\_TREND\_FEATURE\_SELECTION$;

$TREND\_FEATURE\_ENHANCEMENT:$
for each $term \in tTermList$
$if(lookupOntology(term))$
$createRDFdescription(term)$;
for each $message \in tsVectors$
$if(message \supset termfromtTermList)$
$createRDFdescription()$;
$END\_TREND\_FEATURE\_ENHANCEMENT$;

Part IV

Experiments and Evaluation



CHAPTER **11**

Use cases

The goal of this chapter is to show three different cases in which the trend mining approaches presented in this thesis can be applied. In order to derive the general requirements on trend mining, we introduce the different fields of application discussing their particular characteristics. Based on the trend example from the first chapter of this thesis, we describe the first use case that illustrates mining trends from social network messages. A following example describes mining trends in a market research application. The chapter closes with the description of mining trends in financial markets – the showcase that serves as the running example for the evaluation part of this thesis.

11.1 Three application fields

While talking about trends, one can think of changes in the political preferences of a given country before elections, and the shrinking or growing of percentages of followers for each party in the three months before presidential elections. An enthusiast of technology gadgets could associate trends with the newest technology products emerging on the market in the last year. In many cases one thinks about the ups and downs of financial markets, i.e. NASDAQ[1] or DAX[2] curves over stock values in recent months. Indeed, there are many common examples of the word trend and popular conceptions of application fields in which trend mining approaches are useful which were discussed in Chapter4. As for this research, in every kind of problem which involves the reading of textual content that is necessary for the purpose of understanding the development in the particular

[1] http://www.nasdaq.com/ online accessed 10-August-2012
[2] http://www.finanzen.net/index/DAX online accessed 10-August-2012

field of interest, trend mining methods presented in this thesis are useful. The following sections illustrate cases of mining trends in different tasks:

- mining political trends in a particular field of interest while reading social network updates
- mining market trends while conducting a typical market research study
- mining stock exchange trends on a particular financial market

11.1.1 Mining trends from social network messages

There are many possibilities for informing oneself about what is happening in the world. One method is by reading users' updates on social networks, such as Twitter messages and Facebook status updates as illustrated in Figure11.1:

Figure 11.1: Informing oneself from timelines. Source *[Streibel and Alnemr, 2011]*.

We analyzed the process of informing ourselves while reading daily Twitter[3] timelines and Facebook[4] "top news" during the unrest in Egypt from January 26th to February 11th in 2011 and discovered interesting issues about the process itself. We noticed that the process mainly involves *filtering out* the *relevant information*, and can be summarized in 5 steps:

1. Estimating *trending* messages:

[3] https://twitter.com/ online accessed 10-August-2012
[4] http://www.facebook.com/ online accessed 10-August-2012

1.1 Which topics are emerging in the timeline?

1.2 What are trending[5] hashtags[6] in general?

1.3 Which tweets include the trending hashtags and what are they about? What updates appear as the top stories on Facebook?

2. Choosing trending and *interesting* tweets: Which timeline messages fit into my field of interest and piques my curiosity today?

3. Estimating information's *reputation*: Who are the authors of trending and interesting messages? Which of them are interesting according to my own field of interests and according to my own *subjective criteria*:

   3.1 Is the author a real person or a web robot?

   3.2 Is this person trustworthy?

   3.3 What does this person write about in general?

   3.4 Does this person write a lot of messages interesting to me?

4. Extending the list of trending and interesting messages by messages written by the authors of high reputation

5. Reading the information and external links in the tweets that are trending, interesting and trustworthy: linking to external news (blogs, mainstream news portals)

Regarding the time period January 26th through February 11th 2011, an example of trending messages on Twitter were messages marked with the hashtag #jan25 and #Cairo. Most users *interested in the political developments worldwide* could notice an increasing number of messages in their timeline containing information about the situation in Egypt. Terms like "Egypt", "revolution", "President Mubarak", "protest", "Tahrir Square", etc. seemed to appear more frequently than usual. And more Twitter users started to retweet posts containing these words. However, misleading and irrelevant information was also posted using these trending terms for purposes other than the informing others about what was happening in Egypt during those days. An example of situation relevant tweet from February 11^{th} 2011 that uses trending terms: #Egypt #Jan25 #Mubarak #Tahrir is shown in Figure 11.2. In order to distinguish the useless from the valuable information, one had to search for more information about the authors of the tweets by reading their Twitter profiles or their past posts. In cases where it was not possible to find more information about the authors, the statistics were significant in estimating whether the given author is a person who could post

[5] based on Twitter's own trend estimation for trending tags in messages

[6] `http://hashtags.org/` online accessed 01-June-2011

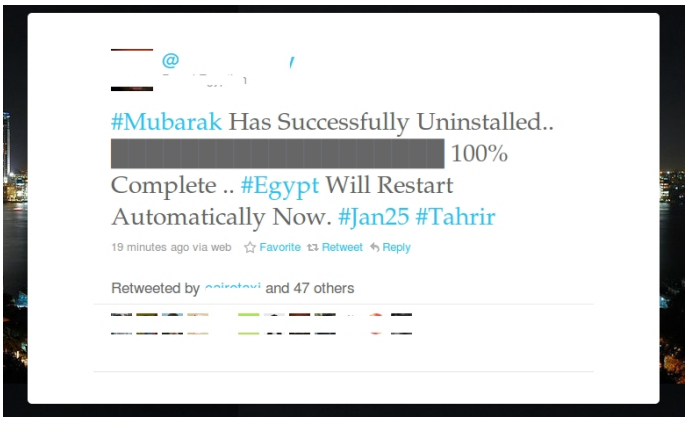


Figure 11.2: An example of a situation relevant tweet. Source: *Twitter*.

trustworthy information: how many messages on similar topics did the author post, how many followers does the author have, how many other users retweeted the posts of this author and how did they comment on this post. After estimating the reputation value of the authors and hence the trust value of the trending and interesting posts, one could continue with reading the chosen tweets and the information contained within them (often the tweets contained links to other social networks posts: blogs, pictures, videos).

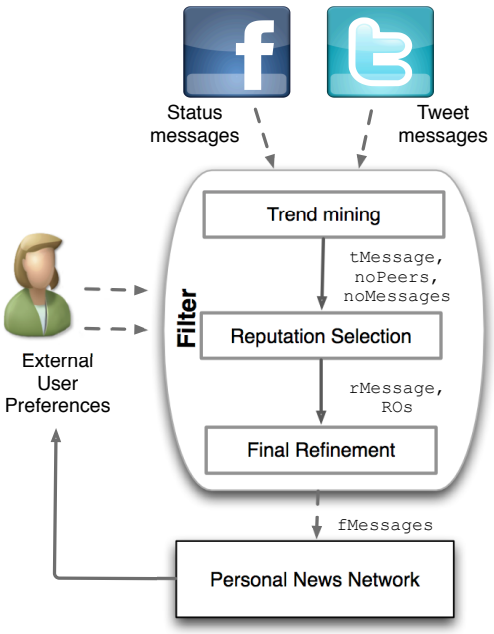


Figure 11.3: Trend mining in a PNN. Source [Streibel and Alnemr, 2011]

The Figure 11.3 illustrates the process of trend mining in social network

messages.

11.1.2 Trends in market research studies

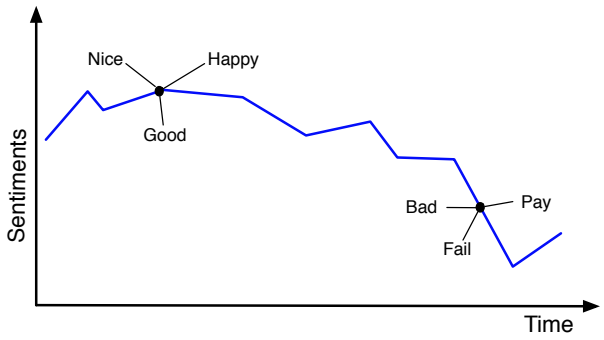


Figure 11.4: A trend in terms of market research. Source: *author*

The objectives of market research projects are to identify market trends as well as to analyze consumer preferences and consumer behavior in the market. A general example of how a trend can be understood in terms of market research is illustrated in Figure 11.4. In general, market research studies are accomplished with projects focused on a certain topic, i.e. on IT market products, and based on two main types of questions:

- quantitative questions (scaled questions): simple choice questions and multiple choice questions related to the topic
- open ended questions: results of primary research (e.g. customers reasons or motivations, their comments, etc.), results of secondary research (e.g. results based upon internet research in order to analyze general trends in a specific market, such as the IT market)

In primary research, the open ended questions are systematically integrated into a market research questionnaire, complementing the quantitative questions. The processing of those open ended questions includes the following steps:

- collection of respondents' opinions
- back translation in a common language: if a study is conducted in a given country, in customers' native language, it is often helpful to translate it into English or German when it is relevant for the market research topic
- categorization
- aggregation and statistical analysis of categories (frequencies and percentages of a given topic)

In the secondary research an explorative process follows additionally to the primary research. It is used in certain types of studies where a broader input and orientation is needed in addition to the study data. The secondary research can be divided into the following stages:

- definition of main study objective, i.e. smartphones
- development of hypotheses that provide a structure for the information research and a focus for the research based upon the required and relevant information, i.e. Samsung Galaxy[7] is growing in popularity in the IT mobile smartphone market. A hypothesis can be answered with a 'yes' or a 'no'.
- collection and screening of relevant internet links
- qualitative and explorative expert interviews (optional)
- aggregation of information

Types of secondary information can be:

- user and buying experiences
- reports about products or markets
- test reports
- predictable or unpredictable events
- regulations and laws
- sales channels

Both question types are crucial for trend detection. While the analysis of quantitative questions is based on the examination of numeric data and can be done automatically using appropriate statistical tools, the analysis of open ended questions still requires human involvement since it is based on the opinion analysis – text analysis where the steps based on categorization, generalization, and interpretation of information are mostly conducted manually. Categorization involves the analysis of positive or negative tagged customer comments that are written in form of unstructured text. Furthermore, the secondary research in market studies includes, in general, the analysis of Internet sources like reports, comments, and news articles that are relevant for the topic of the market research study. Secondary research, like primary research, aims at identifying the customers' opinion trends by categorizing news regarding customer sentiments hidden in texts due to the categories given by the project topic. In general, limitations

[7] http://smartphones.samsung.de/ online accessed 10-August-2012

of the current approaches are mainly based on the difficulty of automatic trend discovery in textual information (customer opinions, articles, reports, news).

Regarding our case study, the main goal of market research is the analysis of market and buying patterns by processing a broad amount of text-based information. The core task in such text processing is the *evaluation of customer opinions* which is based on enhanced text analysis. This includes the *detection of relevant statements*, evaluation of statements and text categorization due to the given project category list regarding the dependencies between sentiments and categories.

11.1.3 Trends on German Stock Market (DAX)

Figure 11.5: DAX curve in September'07 and May'08. Source: *[Economics, 2011]*.

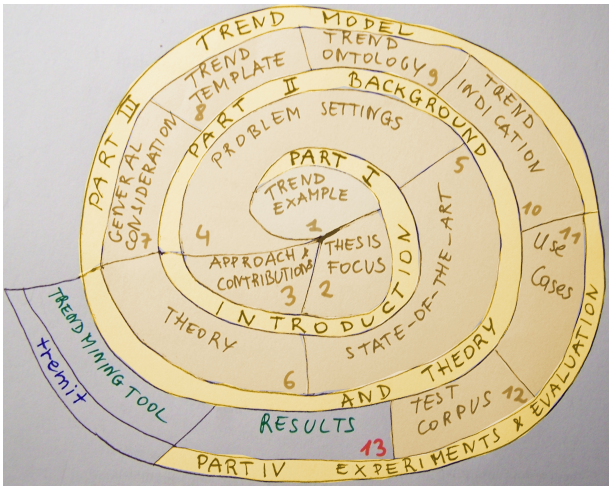
Figure 11.5 presents DAX, a major stock market index which tracks the performance of large companies based in Germany. Curve based on DAX points in September 2007 to May 2008.

The ability to follow the developments of a given financial market and the possibility of deriving trends from such developments is crucial for many people. Traditionally, world events as reported by financial and economic news sources are important indicators for the strength and consistency of the value of trading instruments, i.e. company stocks, and currencies, i.e. the Euro. Such reported events can have far reaching implications to the value of investments made by institutions. Other examples of market moving news are the reported performance of a company or a key speech by a leading political figure or a company chief executive. A clear correlation can be observed between news reports and financial

market fluctuations with structures of names and movements. Crucially, this correlation is causative: the information precedes the fluctuations. Information about critical events may be presented through a variety of media. Conventional information sources are newswire systems such as Reuters, but online news portals also play an increasingly important role. These systems deliver several thousands of pieces of information every day to the desktop of the investor. It is the task of the investor to filter through these data, supported by keyword based filters of the site, for the relevant headlines. While headline information may be of immediate value, the required information such as *predictions*, *expectations* and other *indications of change* may not be as immediately obvious. These content features of the full news text are our main concern.

Furthermore, the first indications of changes in events are likely to be evidenced at the source of the event. This may be evident elsewhere – in a company's ad-hoc announcement, or even in a blog or discussion forum. The need to *extract the significant information* from the massive quantity available is now known to be a prime need for all forms of knowledge and information management in business. The strategic and timely delivery of such content in a form that (human or mechanical) decision makers will be able to react to can be considered as a significant requirement for information systems.[8]

[8]Many thanks to Petra Ristau from JRC GmbH Berlin for the use case explanation.



CHAPTER **12**

Test corpus

Data mining books recommend that you "get to know your data" before analyzing it. In this chapter we describe the data, which is in the case of this thesis the text corpus used for the evaluation. Sketching the historical context of the news contained in the text corpus, we provide the reader with the description of the corpus' content and sources. By presenting different text mining techniques that have been applied to the corpus, we provide a summary of a possible simple analysis of the documents by making use of state-of-the-art NLP routines. Throughout the chapter there are graphics that help to explain what particular kind of information results from each text mining technique. After reading this chapter, there should be a sufficient understanding of our test data as well as an idea of how far a simple text analysis can bring us.

12.1 The historical background

We choose to start the description of the corpus by explaining the context of the corpus content. The historical content helps to explain why we mainly experiment with the news stories which occurred between September 2007 and May 2008. At the end of this section, there is a brief description about the origin of the corpus, to help in understanding initial obstacles in getting interpretable results presented in Chapter13.

12.1.1 The crisis in 2007-2008 in the news

In 2007 the so-called worldwide financial crisis started by the emergence of the sub-prime mortgage crisis on the real estate market in the U.S.. From 2007 to

2008 several important events took place on the financial markets worldwide, among others: Bear Stearns Inc.[1] collapse, Citigroup[2] losses, Lehman Brothers[3] insolvency, as well as several other events as i.e. bank failures. A good retrospective view of these events regarding 306 companies worldwide from the corporate governance perspective on economic crisis is described in [Erkens et al., 2012].

In the emerging economic crisis, September 2007 is just another month in which two important events are being reported on economic news; On the one hand, U.S. dollar is reaching its deepest value counted from 1990. On September, 14th, in one of its news, German economic news provider Godmode Trader reports:
"US-Dollar: Vorsicht, Vorsicht...:
Der Dollar steht am Scheideweg. Der Dollar-Index, der den Wert des US-Dollars gegenüber anderen Weltwährungen abträgt und als Richtschnur bei der Orientierung darüber gilt, ob der Dollar jetzt eben stark oder schwach tendiert, ist auf dem niedrigsten Niveau seit dem Jahr 1990. "[4]
On the other hand the run on Northern Rock [Shin, 2009] in London follows on September 15th:
"The rush of customers taking money out of Northern Rock continued for a second day on Saturday, amid concerns over its emergency Bank of England loan." writes BBC news[5].
The Figure 12.1 shows the USD curve falling since 2002, "DAX, positiver Trend wäre nur noch Zufall" [Godmode, 2011]. In March 2008 JP Morgan plans the takeover of Bear Stearns Inc. . New York Times publishes on March, 17th:
"After a weekend of intense negotiations, the Federal Reserve approved a $30 billion credit line to help JPMorgan Chase acquire Bear Stearns, one of the biggest firms on Wall Street, which had been teetering near collapse because of its deepening losses in the mortgage market."[6] BBC writes:[7]
"Rescue for troubled Wall St bank
(...) The deal values Bear Stearns, which has been at the centre of the US mortgage debt crisis, at just $236m (£116m). Its shares have lost 98% of their value since their high of $158 in April one year ago, when the bank was worth $18bn. (...)"
Shortly after, in April 2008 German Stock Exchange DAX value are falling:
"DAX, positiver Trend wäre nur noch Zufall
"Nach den sehr ungünstigen Vorgaben aus dem Späthandel der Wall Street vom Freitag sowie aus Fernost habe der DAX gestern mit einem Gap nach unten eröffnet und weitere Verluste eingefahren. Flankiert von negativen Nachrichten aus dem Unternehmenssektor (die fünftgrößte US Bank Wachovia vermeldete

[1] until 2008 an investment bank based in New York
[2] http://www.citigroup.com/citi/ online accessed 29-Sep-2012
[3] investment bank based in U.S.
[4] http://tinyurl.com/mnx83ea online accessed 29-Sep-2012
[5] http://news.bbc.co.uk/2/hi/business/6996136.stm online accessed 29-Sep-2012
[6] http://tinyurl.com/k3xb3lm online accessed 29-Sep-2012
[7] http://news.bbc.co.uk/2/hi/business/7299938.stm online accessed 29-Sep-2012

Figure 12.1: A falling curve for USD since 2002. Source: *[Godmode, 2011]*.

nach Angaben der Marktbeobachter einen Quartalsverlust sowie die Planung einer Kapitalerhöhung), wären somit zum Wochenauftakt die Bären am Ruder gewesen." The events mentioned above are only a tiny snippet of information related to the financial and economic crisis which appeared in the news late 2007 and early 2008. However, their importance confirmed our first intuitive impression that online news articles that appeared between September 2007 and April 2008 are part of the bigger, global trend of so called financial crisis. We assumed that this period between September 2007 and April 2008 is an interesting time window in which definitely a trend occurs and therefore trend mining techniques can be applied. Figure 12.2 draws the "crisis anniversary" graphic showing the global events on the Dow Jones curve. Source: Bloomberg, shown at BBC news[8].

12.1.2 Original news corpus

During the TREMA (Trend Mining, Analysis and Fusion of Multimodal Data) research project funded by Investment Bank Berlin, conducted in 2007 and 2008 in Berlin under the cooperation between Free University Berlin, JRC GmbH, neofonie GmbH, and Metrinomics GmbH, a test corpus of over 300,000 web news

[8] http://www.bbc.co.uk/news/business/economy/ online accessed 01-June-2013

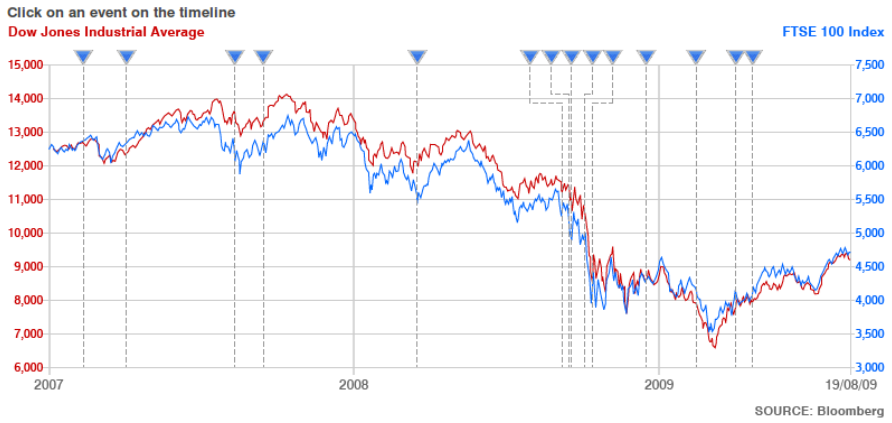


Figure 12.2: The "crisis anniversary" graphic on the Dow Jones. Source: *Bloomberg*.

were created for research purposes. Business news in the German language from German news providers had been stored, anonymized, preprocessed and generally classified into two groups: the articles relevant to general financial market topics, and web news relevant to market research issues regarding the study of technical products and information technology markets. The preprocessed original news corpus was handed out to Free University Berlin for the purpose of further research.

The corpus data was stored in the form of XML files in two directories on a DVD. The original corpus consisted of 276,587 documents of financial news (1.5 GB directory on a hard disc) and 74,145 documents of IT market products (331,6 MB directory on hard disc). A very large part of the corpus consisted of spam articles, articles with encoding errors (encoding label differed from encoding in text body and text title), and articles without a time stamp or with an error in their time stamp. The initial tests could not be performed on this corpus and in order to make it useful some time had to be invested into cleaning the data while following the steps outlined below:

- manual overview of the usefulness of the text content: sorting out chat boards and forums comments that weren't related to the business news
- sorting encoding: we focused on UTF-8 encoded documents
- trustworthiness of the source: we chose to focus on the main news providers and sorted out blogs that concentrated on the long discussions of smaller financial events (such as rising or falling of specific stocks on DAX)
- checking time stamps: time stamps in the files incorporated many errors and had to be sorted out
- sorting out empty articles

- sorting out spam articles

We used automatic as well as manual methods to clean the data. At the end of our cleaning procedure the corpus consisted of 90,000 documents.

Listing 12.1: One of documents as XML file from the original corpus

```
<Document>
 <Meta>
  <BaseUrl>http://de.biz.yahoo.com/</BaseUrl>
  <MimeType>text/html</MimeType>
  <ContentEncoding>iso-8859-15</ContentEncoding>
  <Keywords>adva,adva akkumulieren,akkumulieren,
  deutschland,deutschland netzwerktechnik, deutschland,hardware,hardware ,
    netzwerktechnik,netzwerktechnik hardware,&amp;quot;akkumulieren&amp;quot;,&amp;quot;
    akkumulieren&amp;quot; gesel,&amp;quot;halten&amp;quot;,&amp;quot;halten&amp;quot; jetzt,2007,2007
     bekannt,2007 vergleich,5 millionen,</Keywords>
  <Description>Westerburg aktiencheck de AG - Der Analyst Henning Wagener von
    AC Research erhöht sein Rating für die Aktien von ADVA ISIN DE0005103006/
    WKN 510300 von zuvor &amp;quot;halten&amp;quot; auf jetzt &amp;quot;akkumulieren&amp;quot;
    Die Gesellschaft habe Zahlen für das abgelaufene vierte Quartal und das
    Gesamtjahr 2007 bekannt gegeben Demnach habe das Unternehmen im vierten
    Quartal 2007 im Vergleich zum entsprechenden Vorjahreszeitraum einen Umsatzrü
    ckgang um 7 5 Millionen Euro auf 53 8 Millionen Euro hinnehmen müssen </
    Description>
  <Abstract shortend="true" source="meta">Westerburg aktiencheck de AG - Der
    Analyst Henning Wagener von AC Research erhöht sein Rating für die Aktien von
     ADVA ISIN DE0005103006/ WKN 510300 von zuvor &amp;quot;halten&amp;quot; auf
    jetzt &amp;quot;akkumulieren&amp;quot; Die Gesellschaft habe Zahlen für das [..]</
    Abstract>
  <Title source="head">ADVA akkumulieren - Y! Finanzen</Title>
  <LastVisitDate unit="s">1210454130</LastVisitDate>
  <FirstVisitDate unit="s">1205967991</FirstVisitDate>
  <LastModifiedDate unit="s">1210454130</LastModifiedDate>
  <CreationDate unit="s">1205967991</CreationDate>
  <Size unit="b">16782</Size></Meta>
 <Title>ADVA akkumulieren</Title>
 <Body>Aktienkurse Adva Optical Network... ADV.DE 2.55 -3.41% Westerburg (
   aktiencheck.de AG) - Der Analyst Henning Wagener von AC Research erhöht sein
   Rating für die Aktien von ADVA (ISIN DE0005103006/ WKN 510300) von zuvor "
   halten" auf jetzt "akkumulieren". Die Gesellschaft habe Zahlen für das abgelaufene
   vierte Quartal und das Gesamtjahr 2007 bekannt gegeben. Demnach habe das
   Unternehmen im vierten Quartal 2007 im Vergleich zum entsprechenden
   Vorjahreszeitraum einen Umsatzrückgang um 7,5 Millionen Euro auf 53,8 Millionen
    Euro hinnehmen müssen. </Body>
 <Date>Mittwoch 19. März 2008, 16:17 Uhr</Date>
 </Document>
```

12.2 Content and sources

For our experiments we use a document corpus consisting of business news only in the German language. The corpus originates from the following web sources: comdirect[9], derivatecheck[10], Handelsblatt[11], GodmodeTrader[12], Yahoo[13], Financial Times Deutschland[14], and finanzen.net[15]. The corpus provides news from January 2007 to May 2008. In general, the content of the corpus is focused on finance and business information concerning German companies and stocks. It focuses on the situation at DAX, as well as on reviews and ratings of German companies and shares. It is not only about German companies, but also what is discussed in the German finance and business world. In addition, the corpus contains company-related news regarding Allianz, Bayer, Siemens, ThyssenKrupp, Volkswagen, Apple, Oracle, Starbucks, Lanxess, Fraport, Novartis, Meditec, Google, Ebay, Deutsche Telekom, Thyssen, Daimler, Adva, Yahoo, Porsche, E.ON, CentroSolar, Solarworld, Commerzbank, Citigroup. Initially, the preprocessed and pre-cleansed corpus contained about 90,000 XML files. While performing more tests, we had to extract a cleaner part of this document corpus. The final test documents are divided into 35,650 articles about general business news and over 5,000 articles with specific DAX reports.

12.3 Techniques for preprocessing

In order to get a more detailed overview of the corpus content, we applied ad hoc different NLP analysis techniques.By sorting out other error files, the overall amount of files was reduced from about 42,300 to about 40,500 (5% loss). In order to provide more relevant information about every document, we use mainly (1) named-entity recognition, (2) part-of-speech tagging, and (3) stemming, and add the tags: `NER`, `POS`, `StemmedBody` in addition to the already given XML tags like `meta` (e.g. URI, keywords, CreationDate, ...), `title`, `abstract`, `body` and `date` in the test corpus files.

12.3.1 Preprocessing

In general, the preprocessing has to be performed in order to reduce the amount of text features for the further analysis. Methods applied in preprocessing always depend on the art of aimed analysis. In most cases it is useful to apply lexical

[9] http://www.comdirect.de/inf/index.html, online accessed 25-April-2012
[10] http://derivatecheck.de/, online accessed 25-April-2012
[11] http://www.handelsblatt.com/weblogs/, online accessed 25-April-2012
[12] http://www.godmode-trader.de/, online accessed 25-April-2012
[13] http://de.biz.yahoo.com/, 25.04.2012
[14] http://www.ftd.de/, online accessed 25-April-2012
[15] http://www.finanzen.net, online accessed 30-April-2012

analysis methods in which words in the text are lowcased (capital letters replaced by lowercase), numbers, punctuation and additional characters are excluded, and stop words are removed. Stop words are words that do not carry any content per se and therefore can be omitted in the further content analysis. As for the stop words, we apply the Apache Lucene library[16]: GermanAnalyzer[17] default stop word set. Additionally, we extend it by our own 300 stop words. An example of the additional stop words used for our test corpus is shown below.

```
aber, alle, allem, ..., allen, aller,
aus, ..., derselbe, derselben, ...,
eines, einig, ..., zu, zum, zwar
```

Next to the removing of stop words is the stemming or lemmatization of the text. Sometimes it is useful to stem the documents before further analysis. In some cases it is better to lemmatize the text. A difference between both methods is very well described in one sentence:
"If confronted with the token *saw*, stemming might return just s, whereas lemmatization would attempt to return either see or saw depending on whether the use of the token was as a verb or a noun."[18]

Stemming

Stemming is the process of reducing a given morphological word form so that the word is represented by its stem. A stem is the basic morphological form of a given word. The resulting stem of the word is depending on the particular definition of stems in the particular stemming function. As for the stemming of our documents we apply the GermanAnalyzer from Apache Lucene. An example to a stemmed word from the test documents is shown below.

```
word:            stem:
     getragen         getrag
     vergleichweise   vergleichwei
     übetroffen       ubertroff
```

Lemmatizing

Lemmatization is the lexicographical reduction of a word so that the given word is represented by its *lemma*. A lemma is the basic linguistic form of a given word. As in many cases of natural language processing, the available libraries are

[16] http://lucene.apache.org/core/, online accessed 10-Jan-2012
[17] http://lucene.apache.org/core/old_versioned_docs/versions/3_0_1/api/all/org/apache/lucene/analysis/de/GermanAnalyzer.html, online accessed 10-Jan-2012
[18] http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html online accessed 10-Jan-2012

easier to apply and more sophisticated for English than for the German language. Standford Lematizer[19] is a very good tool for lemmatizing texts in English. We experimented with another library: LanguageTool[20] for German. LanguageTool is meant to be used for orthographical text correction but it can be extended very easily into a lemmatizer for the German language. However, our procedure is not efficient enough to be applied for document analysis on the fly. Applying the lemmatizer to 200 documents takes a few minutes and therefore we focus mainly on stemming. An example of a lemmatized word from the test documents is shown below.

```
 word:            lemma:
      machte           machen

The complete result of LanguageTool for machte\:

machen/VER:1:SIN:KJ2:SFT
machen/VER:1:SIN:PRT:SFT
machen/VER:3:SIN:KJ2:SFT
machen/VER:3:SIN:PRT:SFT
```

12.3.2 Named entity recognition

When it comes to information extraction, we concentrated on named-entity recognition to extract text elements. Different types of entities can be recognized in the given text. We focus mainly on organizations, locations, and persons. For that we have used the Stanford Named Entity Tagger[21] to identify entities from the text body, abstract and title. As the corpus is in German, we used the German Named Entity Recognition[22] with the *Huge German Corpus-generalized classifier*[23]. The results are grouped in four categories: location, person, organization and miscellaneous. Below is an example of recognized entities in one of the test documents:

```
nerorganization = "Audi RS 4",SEAT,"Volkswagen AG",
                  "Volkswagen Konzerns",Jetta,"Passat Variant",
                  "Volkswagen Cabriolet-Coupés Eos",
                   WKN,IAA,Audi,DAX,DGAP,
                   "Volkswagen Konzern"
```

[19] https://github.com/larsmans/lucene-stanford-lemmatizer, gesichtet am 23.02.2012

[20] http://extensions.services.openoffice.org/project/languagetool, online accessed 23-Feb-2012

[21] http://nlp.stanford.edu/software/CRF-NER.shtml, 30.04.2012

[22] http://www.nlpado.de/~sebastian/software/ner_german.shtml, 30.04.2012

[23] http://www.ims.uni-stuttgart.de/~pado/hgc_175M_600.ser.gz, online accessed 20-Jan-2012

```
nerlocation =     Hannover,Antwerpen,Hamburg,Frankfurt,
                  London,Wolfsburg,Wien,Luxemburg,EUR,
                  USA,Berlin-Bremen,Stuttgart,Tokio,
                  Deutschland,Golf,China,Westeuropa,
                  "Wolfsburg Deutschland"
nerperson =       "Dow Jones"
```

As presented in the example above, the recognition of entities is quite good, at least in the case of locations. However, there are still some errors in the results of NER-method: many car names are recognized as organizations, and the Dow Jones, which is actually a stock index, is recognized as a person.

12.3.3 Part of speech tagging

In Chapter 7 we discussed the importance of verbs and adjectives in textual trend mining. In order to get an overview of the different POS-words, in particular nouns, verbs and adjectives, we used the Stanford Log-linear Part-Of-Speech Tagger[24] (POS) along with the STTS (Stuttgart-Tübingen-TagSet) tagset[25] to extract the POS-words from the text body. An example of a POS-tree of one of the test documents follows. Figure 12.3 shows a visualization of a POS-tree part

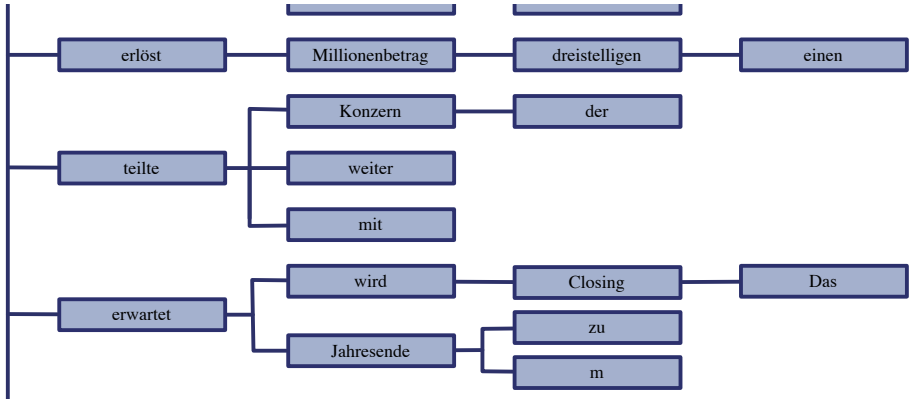


Figure 12.3: POS-tree. Source *author*.

from a test document. The sentence at the bottom of the figure: "Das Closing wird zum Jahresende erwartet" (EN: The closing is expected at the end of the year.) is splitted into the particular part-of-speech untities, a verb *erwartet*, the auxiliary verb *wird*, the substantive *Closing*, the article *Das*, and the preposition *zu* with the declined ending *m*, and the corresponding part of the adverbial phrase *Jahresende*.

[24] http://nlp.stanford.edu/software/tagger.shtml, online accessed 20-Feb-2012

[25] http://www.ims.uni-stuttgart.de/projekte/CQPDemos/Bundestag/help-tagset.html, online accessed 20-Jan-2012

12.3.4 Final Format

The resulting data is an almost spam free German text corpus ready to use, providing relevant additional information as described above. Listing 12.2 shows the resulting XML format and sample data. Some of the `metatags` can remain empty.

Listing 12.2: Sample XML file

```
<Document>
  <Meta>
    <ID>2</ID>
    <BaseURL>http://www.finanzen.net/</BaseURL>
    <URI>
      http://www.finanzen.net/nachricht/DAX_
                    am_Mittag_Positive_Entwicklung_bei_geringen_
                    Umsaetzen_Volkswagen_im_Fokus_360739
    </URI>
    <Keywords>Börse, Ad-hoc, Marktberichte, Pressemitteilungen, [...]</Keywords>
    <Copyright>SmartHouse Media GmbH</Copyright>
    <Author />
    <Email />
    <Title source="head">DAX am Mittag: Positive Entwicklung bei geringen Umsätzen, Volkswagen im Fokus | Nachrichten |</Title>
    <CreationDate>2008-03-20 18:45:23</CreationDate>
    <LastModifiedDate>2008-03-20 18:45:23</LastModifiedDate>
    <FirstVisitDate>2008-03-20 18:45:23</FirstVisitDate>
    <LastVisitDate>2008-03-20 18:45:23</LastVisitDate>
  </Meta>
  <Title>DAX am Mittag: Positive Entwicklung bei geringen Umsätzen, Volkswagen im Fokus | Nachrichten |</Title>
  <Abstract>DAX am Mittag: Positive Entwicklung bei geringen Umsätzen, Volkswagen im Fokus | Nachrichten | Aktienkurs | 750000 | | DE0007500001</Abstract>
  <Body>
  Die deutschen Standardwerte entwickeln sich zum Mittag hin beinahe ausnahmslos positiv und vergrößern ihre Gewinne aus den frühen Handelsstunden zusehends. [...]
  </Body>
  <Date>2005-12-27 00:21:36</Date>
  <NER>
    <Location>Europas</Location>
    <Person>Georg Kofler, Bernd Pischetsrieder, Axel Springer, [...]</Person>
    <Organisation>DAX, VW, Siemens, Hypo Real Estate, EADS, [...]</Organisation>
    <Miscellaneous>deutschen, kanadischen, Fußball-Bundesliga, TV-Rechte, [...]</Miscellaneous>
```

```
31    </NER>
32    <POS>
33      <Noun>Standardwerte, Mittag, Gewinne, Handelsstunden, [...]</Noun>
34      <Verb>entwickeln, vergrößern, gewinnt, markiert, [...]</Verb>
35      <Adjective>deutschen, ausnahmslos, positiv, frühen, [...]</Adjective>
36    </POS>
37    <StemmedBody>deutsch standardwert entwickeln positiv [...]</StemmedBody>
38  </Document>
```

12.3.5 Vizualization

In order to understand different test results and to have an overview of the corpus and its parts, we visualize the documents by using the timeline widget from Simile Widgets[26].

Timeline

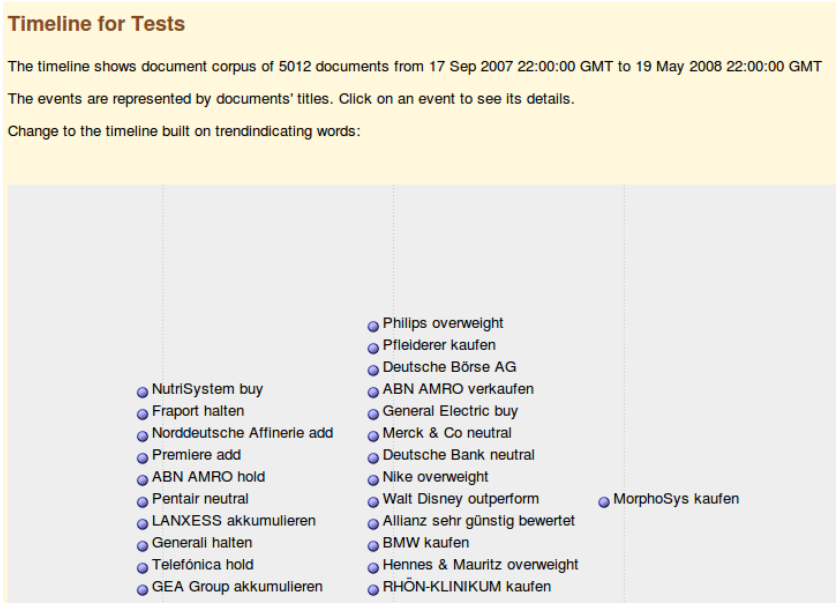


Figure 12.4: Simile timeline

DB GUI

As for the interpretation of different test results or the justification of errors, the GUI of the data bank was especially helpful in the beginning.

[26] http://www.simile-widgets.org/timeline/, online accessed 10-Jan-2012

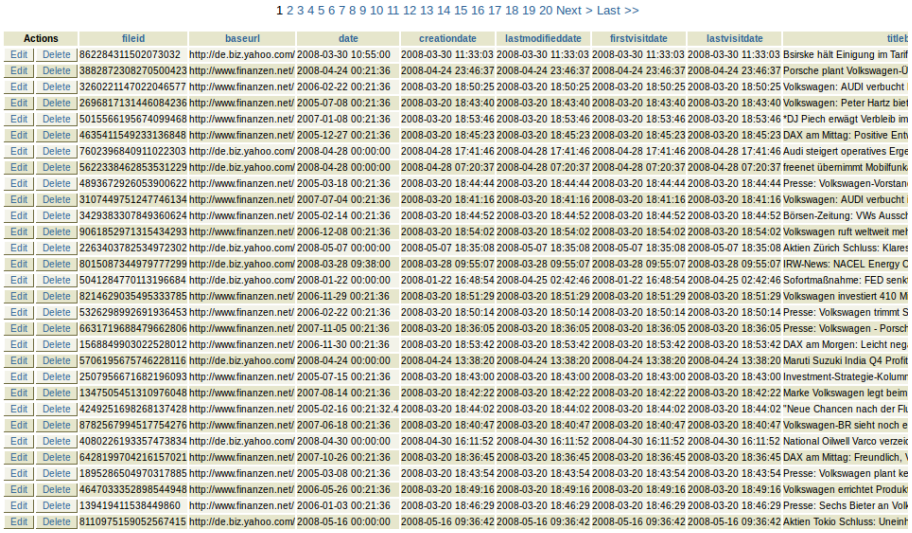


Figure 12.5: GUI for PostgreSQL

12.4 Resulting corpus

The resulting corpus contains general business news and DAX specific news in German. The average size of the articles is 2000 characters. 98% of the corpus is spam free. Documents are UTF8 encoded, timestamped, stored as XML files.In Table 12.1 the general statistic about the main part of our test corpus is presented. Table 12.2 shows the number of documents per month and week. The Figures

Year	Month	Number of documents
2007	9	1125
2007	10	3951
2007	11	3290
2007	12	4316
2008	1	2263
2008	2	94
2008	3	3203
2008	4	11647
2008	5	5663

Table 12.1: General statistics: number of files per month

12.6 and 12.7 visualize the distribution of files (the xml-based web documents, timestamped after the creation date) per month and per week respectively.

Year	Month	Week	Number of documents
2007	9	35	1
2007	9	36	16
2007	9	37	39
2007	9	38	316
2007	9	39	753
2007	10	40	704
2007	10	41	730
2007	10	42	999
2007	10	43	1063
2007	10	44	455
2007	11	44	141
2007	11	45	689
2007	11	46	796
2007	11	47	881
2007	11	48	783
2007	12	1	82
2007	12	48	8
2007	12	49	861
2007	12	50	1162
2007	12	51	1568
2007	12	52	635
2008	1	1	546
2008	1	2	522
2008	1	3	564
2008	1	4	614
2008	1	5	17
2008	2	5	5
2008	2	6	13
2008	2	7	18
2008	2	8	28
2008	2	9	30
2008	3	9	5
2008	3	10	60
2008	3	11	68
2008	3	12	722
2008	3	13	1809
2008	3	14	539
2008	4	14	2146
2008	4	15	2217
2008	4	16	2595
2008	4	17	2928
2008	4	18	1761
2008	5	18	871
2008	5	19	1934
2008	5	20	1920
2008	5	21	938

Table 12.2: Statistics for month - week- day number of files

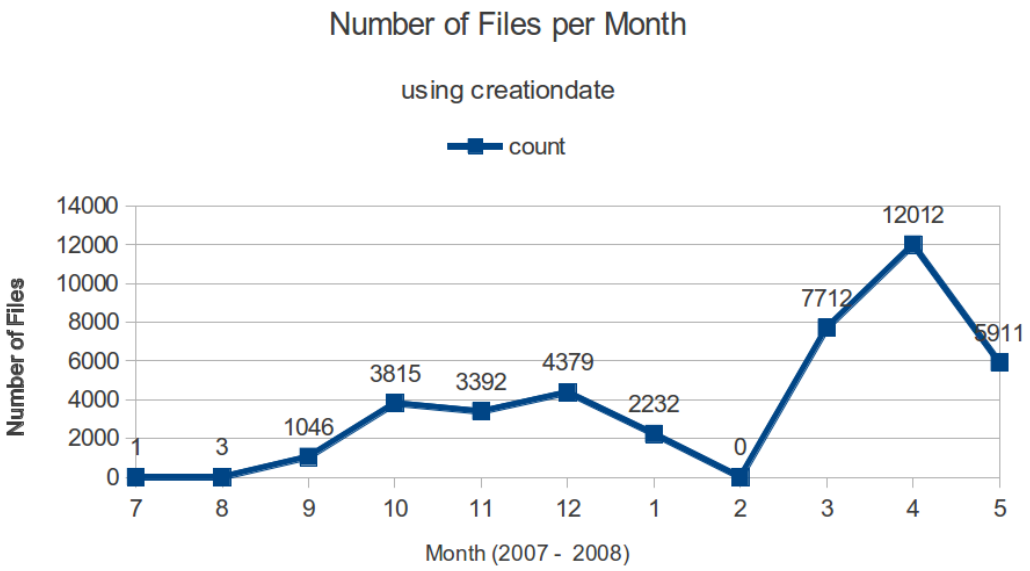


Figure 12.6: Distribution of corpus size (per month)

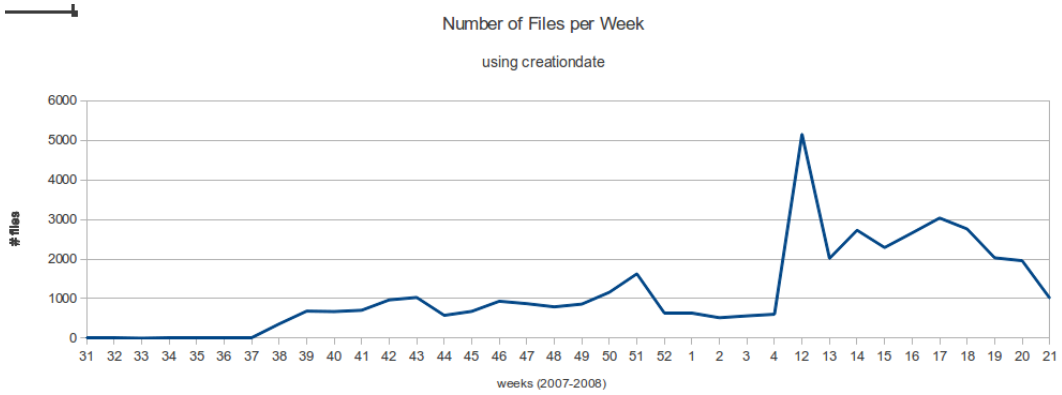
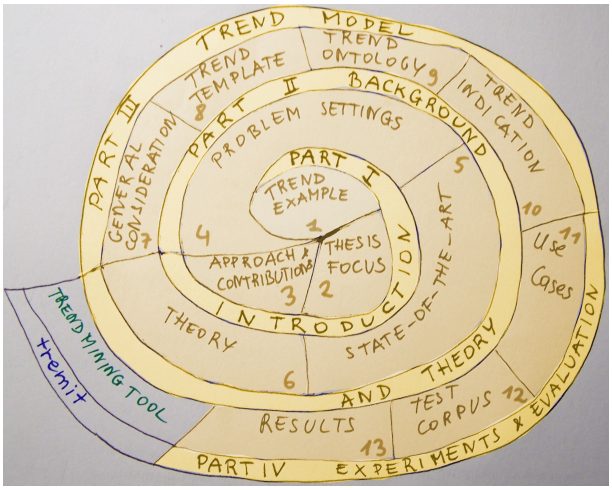


Figure 12.7: Distribution of corpus size (per week).



CHAPTER **13**

Results

"If there are facts that constitute the basis for science then those facts come in the form of experimental results" [Chalmers, 1999]

Chapter 13 contains the evaluation of research presented in this thesis. The description of the experiments and their results allows for a deeper understanding of advantages and limits in trend mining, in particular in the knowledge-based trend mining approach. Before presenting and discussing the actual results of this research, we start with the general introduction into the main evaluation frame, explaining why we choose to follow goal-free experimental evaluation, how this quantitative research fit into the overall scientific method, and which influence does the use case take on this evaluation. In the beginning of the Section 13.2, we define the overall frame of the experiments and set up goals for presenting results.

13.1 General introduction into evaluation

An exact and accurate evaluation frame seems to be a difficult problem if one does not dive into the theory and philosophy of science. Simply because a question about what does the scientific evaluation actually mean raises as first. Nevertheless, after diving into the philosophy of science, the creation of an satisfiable evaluation frame may appear even more difficult, however its obstacles become more understandable.

Linked to the question of scientific evaluation, it immediately appears that the existence of scientific method is necessery for the evaluation. And the first step of the method is to pose a *scientific question* for which one proposes a *falsifiable hypothesis* (p. 59–103, [Chalmers, 1999]) that in turn is the first step of scientific evaluation. The main scientific question of this thesis, stated in Chapter

2, sounds: *Can a knowledge-based trend model help us in understanding trends?* Our hypothesis is that it can and we propose the trend template as the basis for a knowledge-based trend model. The experiments with the implementations of the trend template should either prove or disprove this hypothesis.

[Dodig-Crnkovic, 2002] sketches the general scientific method showing the steps from the first: *posing the scientific question* to the last: *obtaining consistency for the formulated hypothesis* which is a tentative answer for the posed question. [Dodig-Crnkovic, 2002] and [Chalmers, 1999] investigate science, its methods and theory, discuss scientific method and scientific evaluation extensively and show ([Chalmers, 1999]) the meanders of the evaluation idea itself.

In Chapter3 we sketch the overall method applied while conducting this research. The 4th and 5th step of this method, called *experiment* and *data*, are the evaluation part of this research.

[Chalmers, 1999] devotes one chapter on the matter of experiment in which the author states: *"Not just facts, but relevant facts"* are leading to the science. *"(...) it should perhaps be somewhat obvious that if there are facts that constitute the basis for science, then those facts come in form of experimental results rather than any old observable facts."* Whereas *experiment* is a strong instrument in science, it does not always allow for a proof or disproof of scientific hypothesis [Chalmers, 1999]. As for the computer science *"There are plenty of computer science theories that haven't been tested. For instance, functional programming, object-oriented programming, and formal methods are all thought to improve programmer productivity, program quality, or both. It is surprising that none of these obviously important claims have ever been tested in a systematic way, even though they are all 30 years old (...)"*.[1]. In the context of any research in computer science there is another problem with the experiment about which [Chalmers, 1999] is writing– the general statement on *experiment* relates to the general view of *science*.

Regarding computer science it is worth mentioning that there are different opinions on classification of computer science as a science. As written by [Tichy, 1998] *"A major difference to traditional sciences is that information is neither energy nor matter."* However, [Tichy, 1998] states *"The fact that in the field of computer science the subject of inquiry is information rather than matter or energy makes no difference to the applicability of the traditional scientific method. In order to understand the nature of information processes, computer scientists must observe phenomena, formulate explanations and theories, and test them."*

[Dodig-Crnkovic, 2002] shows that computer science lies on the boarder between science and engineering (or technology). Accordingly, methods applied in science as well in engineering apply either to computer science. The scientific

[1] as for 2012, functional programming, object-oriented programming, and formal methods are now over 40 years old

method applied for this research as presented in Chapter 3 is a general method and through its generality can be adopted to natural science as well as to engineering. Evaluating through experimentation in computer science is, despite the general assumption that evaluation in computer science is somehow obvious, not always possible and very often not feasible. This problem of the scientific evaluation in computer science is widely discussed by [Tichy et al., 1995]. The authors conclude that there are several research projects that cannot claim in-depth scientific proof. However, the requirement for the *in-depth scientific proof* may be somehow out-dated or not always relevant regarding the research in computer science.

Computer science – being a mix from science and engineering, strongly depending on technology development and strongly influencing this development – is changing rapidly. The theory and the philosophy of science that lie at its ground do not seem to change. Therefore, while formulating the evaluation frame for a given research in computer science – beside the knowledge about the scientific evaluation within a scientific frame – it is important to focus on the following issues:

- a given research idea within its respective research field
- a given use case
- the available test data
- and the potential users (if applicable)

13.1.1 Possible evaluation directions

The research idea of this thesis concentrates in general on trend mining, which, located in the IR, DM, and TDM research (see Chapter 5), is itself identified through this thesis as a research field. The particular use case of the trend mining is connected to the test data, which is the German stock exchange news in our case. And we do not rely on a particular user group.

Before describing the setting for the experiments of this thesis, we should consider the fact that a deep going evaluation can have more dimensions. For example, it can concentrate on specific parts of a given model or algorithm. Within the trend mining this work proposes a trend template as the base for knowledge-based trend mining, from which two different implementations are shown – trend ontology and trend indication. The evaluation could take different directions by focusing on:

- theoretical correctness and consistency of the trend template
- the performance of the model (trend ontology and trend indication) on test data

- user satisfaction with the model results
- model's persistence in different use cases
- model's performance in different use cases
- model's ability for making its results interpretable

Taking one of the evaluation direction is helpful for setting the experiment's criteria. But, before we decide on a specific direction, we should consider another relevant evaluation issues – evaluation approaches and evaluation methods, explained in the following sections, 13.1.2 and 13.1.3. We describe these methods with regard to a general evaluation idea for the trend mining.

13.1.2 Possible evaluation approaches

There are different possible approaches for an evaluation. The most relevant for the trend mining are described as follows.

Goal-free evaluation

In some cases of evaluation it is useful to set up a set of criteria upon which different approaches can be tested against the test data and therefore can be compared by their results while focusing on discovering new aspects. This is called a goal-free evaluation. As described in [Scriven, 1991], "The value of a goal-free evaluation does not lie in picking up what everyone already *knows* but in noticing something that everyone else has overlooked, or in producing a novel overall perspective."

Quantitative evaluation

A quantitative evaluation approach is focused more on measuring of the achieved results. Applying the quantitative evaluation approach means to conduct the systematic measurements. The systematic measure and evaluation of the created model is conducted in order to test a given hypothesis.

A hypothesis – the testable, the null, and the fallacy

There is no scientific proof without a testable hypothesis. The main characteristic on which science in general is based, is the *falsifiability*. There can be no "real" science without falsifiable statements [Chalmers, 1999]. "There is an important characteristic of a scientific theory or hypothesis that differentiates it from, for example, a religious belief: a scientific theory must be 'falsifiable' "[Dodig-Crnkovic, 2002].

In statistical testing the alternative hypothesis is the assumed statement. This so called alternative hypothesis has to be proven against the null hypothesis. A null hypothesis is the hypothesis that actually has to be disproved by proving the alternative hypothesis. In our case the null hypothesis says that there will be no difference in the results of clustering, topic modeling, and trend template based algorithm related to our experiment frame. While testing a given hypothesis, one should be aware of so-called fallacy. "A formal fallacy is a wrong formal construction of an argument. An informal fallacy is a wrong inference or reasoning." [Dodig-Crnkovic, 2002]

13.1.3 Relevant evaluation methods

Regarding the research relevant for trend mining (see Chapter 5), different evaluation methods from the respective research directions can be considered as applicable for trend mining. In the following subsections, we show the main evaluation ways from TDT, IR, and the DM research.

TDT evaluation frame

TDT research proposes an evaluation cycle consisting of five steps: task definition, system design, system building, system testing, system refinement. Every TDT task is being evaluated as a detection task. As listed in Section 5.2.2 there are five tasks on which the TDT research builds. The most relevant for trend mining in general is the topic detection task.

A TDT system "is presented with input data and a hypothesis about data, and the system's task is to decide whether the hypothesis about this data is true". TDT research calls it a trial. "If the hypothesis is true, the trial is called a target; if not, the trial is called a non-target trial." Normalized detection cost function is applied as the function alongside with detection error trade-off (DET) curves for evaluating the TDT. " Since TDT evaluations use many topics, the global assessment of system performance is accomplished by averaging both the detection cost function and DET curves across topics." The single detection cost function is defined as follows:

$$C_{Det} = (C_{Miss} * P_{Miss} * P_{Target} + C_{Fa} * P_{Fa}(1 - P_{Target})) \tag{13.1}$$

whereas P_{Miss} and P_{Fa} are defined as follows:

$$P_{Miss} = \#MissedDetections/\#Targets$$

$$P_{Fa} = \#FalseAlarms/\#Non - Targets$$

More detailed description can be found in [Allan, 2002], p. 23. The TDT evaluation approach can be generally applied in trend mining if a given trend mining method focuses in particular more on a trend discovery than on a trend interpretation.

(Web) IR- evaluation techniques

Web information retrieval evaluation regards mainly the problem of web search. In order to evaluate web search, different techniques can be used. TREC [NIST, 2013] helps in defining the evaluation steps, some other relevant are described in [Göker and Davies, 2009] (p. 93–97).

The main and general basic formula used in the IR evaluation is the precision and recall method:

$$P = \frac{TP}{TP + FP} \tag{13.2}$$

$$R = \frac{TP}{TP + FN} \tag{13.3}$$

The evaluation of a trend mining approach (in case of a learning model) can be based on the evaluation of the model performance which can be conducted using cross-validation and measured in general by the recall and precision values. For the cross-validation, the document corpus is divided in i folders and the validation process is repeated i times whereas in every i-step of the validation the $\frac{1}{i}$ part of the document corpus is used as a test set while the rest $\frac{i-1}{i}$ stacks are used for building the learning model. If D is the set of documents, $|D|$ is the total number of documents in the set, the precision and recall value can be defined by:

$$recall = \frac{|D|_{trendindicating-and-retrieved}}{|D|_{trendindicating}} \tag{13.4}$$

$$precision = \frac{|D|_{trendindicating-and-retrieved}}{|D|_{retrieved}}$$

Evaluation in data mining

In general, DM offers different measures for different problems and the particular evaluation depends on the particular problem. Among others, the well known measures as the t-test (p. 370-371 in [Han and Kamber, 2006]), f-measure and ROC curves (p. 172 in [Witten and Eibe, 2005]). More measure methods can be found in the statistical data analysis, e.g. [Fahrmeir et al., 2007].

13.1.4 Basic metrics

The following basic metrics can be applied for determining thresholds of the different weighting functions. In particular, the arithmetic mean and the median can be applied for defining threshold values of outliers.

Arithmetic mean

Regarding the arithmetic mean given by:

$$\bar{x} = \frac{1}{n}(x_1 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{13.5}$$

we can extract the arithmetic mean values for the outliers in small corpus as follows:

Median

Given that a *median* is defined by:

$$x_{med} = \begin{cases} x_{(\frac{n+1}{2})} \text{ if } n \text{ is odd} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) \text{ if } n \text{ is even} \end{cases} \tag{13.6}$$

Absolute error

Also, for the numeric prediction, the relative absolute error measure can be applied:

$$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{|a_1 - \overline{a}| + \ldots + |a_n - \overline{a}|} \tag{13.7}$$

with:

$$\overline{a} = \frac{1}{n}\sum_{i} a_i$$

$p_1, p_2, \ldots p_n$ mean the predicted value for the test instances and $a_1, a_2, \ldots a_n$ the actual values.

The formulas above give only an insight into the possible measure ways that are applicable for trend mining. The final evaluation always depends on the final model.

13.2 Experimental evaluation

From Section 13.1, we learn that based on our scientific method, the experimental evaluation is relevant for our research, and that we should focus on the use case and the test data in our particular case. Our particular use case is the financial market, the German stock exchange (DAX). Our test data are web news about the DAX in German language. We identified two algorithms as mainly relevant for trend mining, the k-means clustering method and the LDA-based topic models. And we want to test them together with our knowledge-based approach on the test corpus. Section 13.1.1 lists the possibilities for the experiment's directions, from which we decide to focus on models' ability to simplify the results' interpretability.

Section 13.1.2 describes the different approaches from which we take the goal-free experimental evaluation. Below we describe our evaluation frame.

13.2.1 Evaluation frame

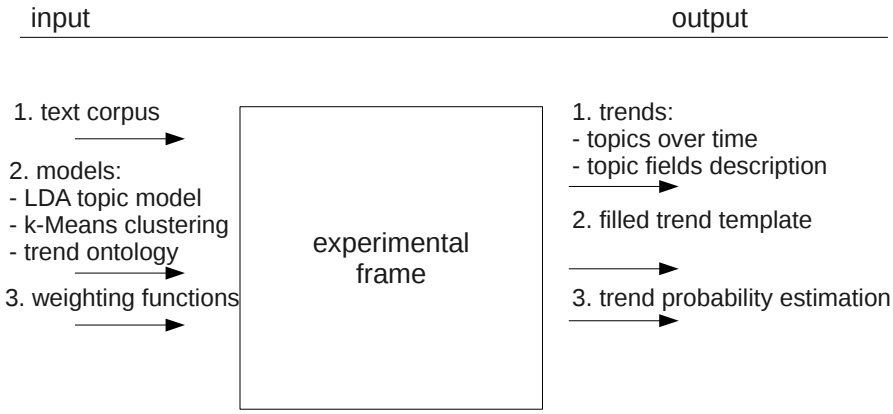


Figure 13.1: Experimental frame

In general, we aim at testing the different possibilities that the particular algorithms bring for trend mining. Based on the given algorithms, we propose to summarize the possible analysis:

1. NLP-based document analysis: using NLP (see Chapter 12), we can derive some simple information about the corpus.

2. probabilistic approach: applying LDA topic models, we can extract topics for the given time window and predict the upcoming topics

3. statistical analysis: applying the k-Means clustering we can derive topic clusters

4. combination of probabilistic and statistical approach: applying the k-Means method with the LDA topic models, we can derive upcoming topic clusters

5. knowledge-based ontology approach: applying trend ontology we can construct the topic fields for the topics from the test corpus

6. knowledge-based statistical approach: applying trend indication functions, we can extract the triggers for the document corpus

Obviously, much more combination and experiments are possible, however we focus in general on the results' interpretability while searching for the methods' features that contribute to the interpretability, and their features which contribute to the prediction. In Figure 13.1 we sketch the experimental frame that shows on the left side the input – algorithms and the test corpus, and on the right side the expected output – trends in forms of topics over time, topics interpretation, the filled trend template and the trend estimation. Some of the experiments conducted that build our experimental frame in the center of the Figure 13.1, are described in the following. Further experimentation will be the subject of future work.

13.3 Experiments conducted

In the following, we report chosen parts of our experiments conducted during this thesis. All of the experiments were performed using the tremit, our trend mining tool, which we briefly describe in A. The experiments presented here give an insight into the possible ways how to proceed with the trend mining relevant analysis of a given document corpus. They provide an interesting overview of the particular raw results that one gets from the state of the art trend mining algorithms by applying the methods to the (pre-processed) test corpus.

13.3.1 Corpus

For the experiments, we used different parts of the corpus, applying the algorithms mainly on two corpus parts which we call the small corpus, consisting of 5,012 web documents mainly reporting on particular DAX companies, and the big corpus, consisting of 35,549 web documents about the general developments on DAX. We conducted the experiments on different text parts of the documents, testing the methods on documents' title, its keywords, the abstract and the body content (see also the listing 12.2 in 12.3.4). In some cases, when the given document part was not in the corpus or many documents were described by the same document part (e.g. the same keywords), the number of documents tested varied (e.g. often it was possible only to test about 4,696 documents reporting on particular DAX companies).

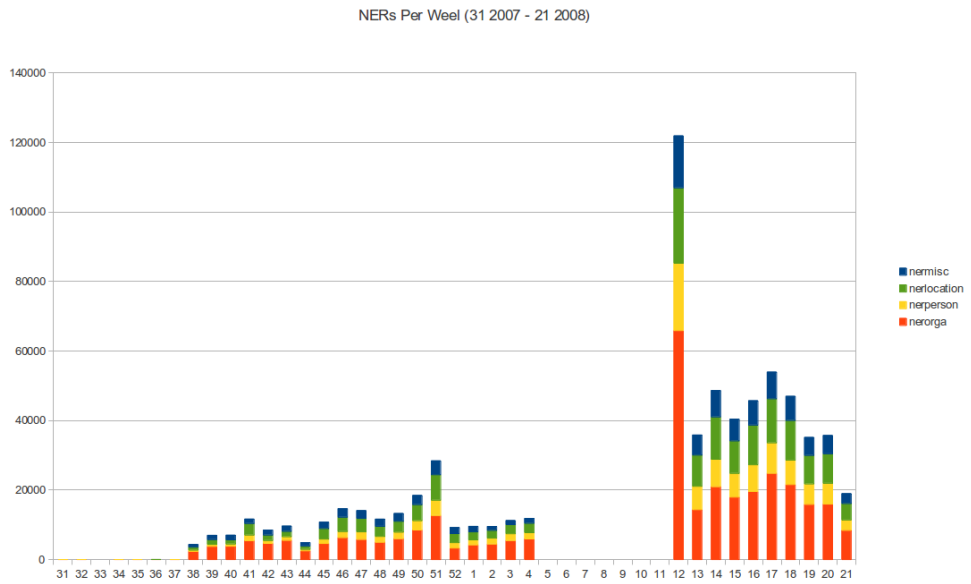


Figure 13.2: Number of different NERs per Week.

13.3.2 NLP on the corpus

From the different NLP techniques (see their description in Chapter 12) applied on the test corpus, we show in the Figure 13.2 to 13.6 the distribution of the recognized named entities in the time window: 31st week of 2007 to 21st week of 2008. Figure 13.2 visualizes the number of named entities per week. The red part of the graph relates to the recognized named entities (NERs) that are the names of different organizations mentioned in the news. We see that the number of the NERs relating to the organizations' names is bigger than the number of NERs describing persons (yellow), locations (green), and the miscellaneous NERs (blue). However, it varies over the weeks. Figure 13.3 presents the NERs in percentages, wherein the distribution is more clearly. The 'empty' NERs weeks refer to our missing articles in the corpus (see Figure 12.7 in Chapter 12). Figures 13.4 and 13.5 show the distribution of the NERs per day. What we can follow from all these figures is that the analyzed news are reporting mostly on organizations, followed by the locations and persons. Furthermore, the number of the particular NERs per week (organization, location, person NERs) varies more than the number of the same NERs per day, which looks as equal amount of information on organizations, locations, persons, and the miscellaneous NERs.

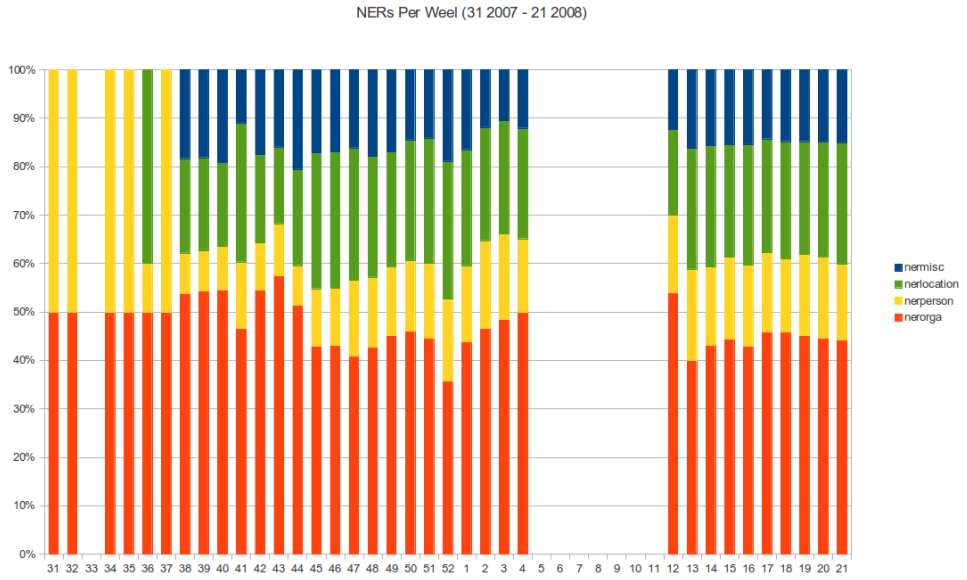


Figure 13.3: Percentage of NERs per week.

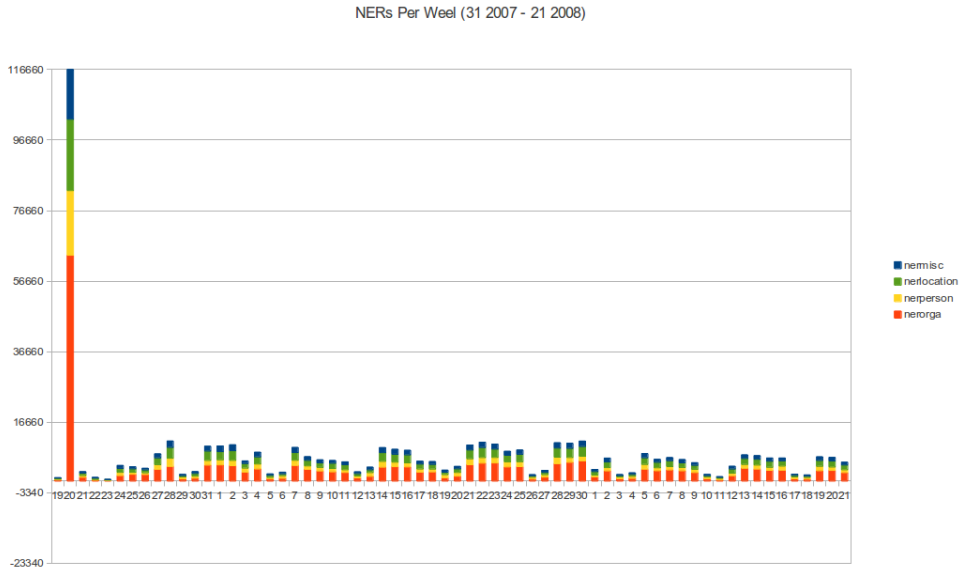


Figure 13.4: Number of NERs per day.

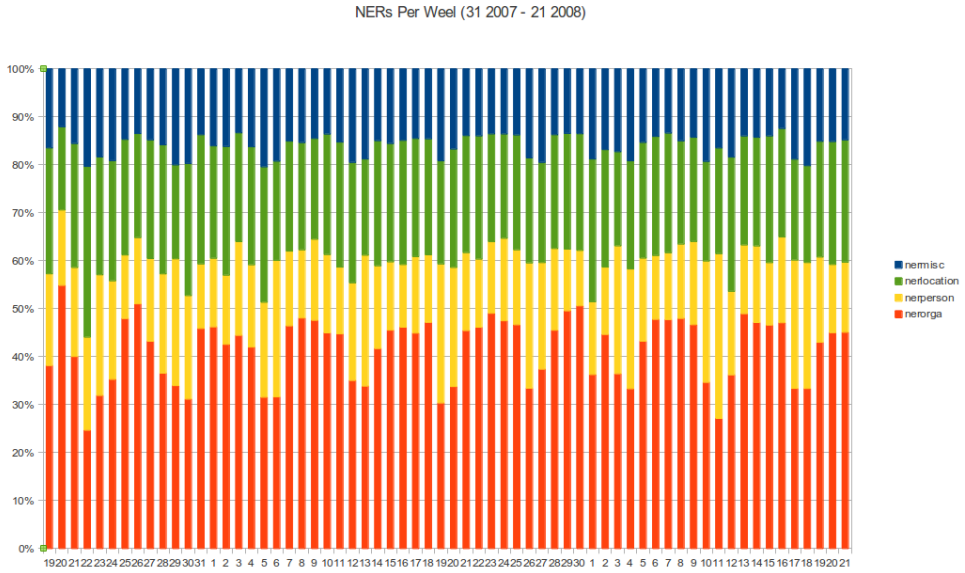


Figure 13.5: Percentage of NERs per day.

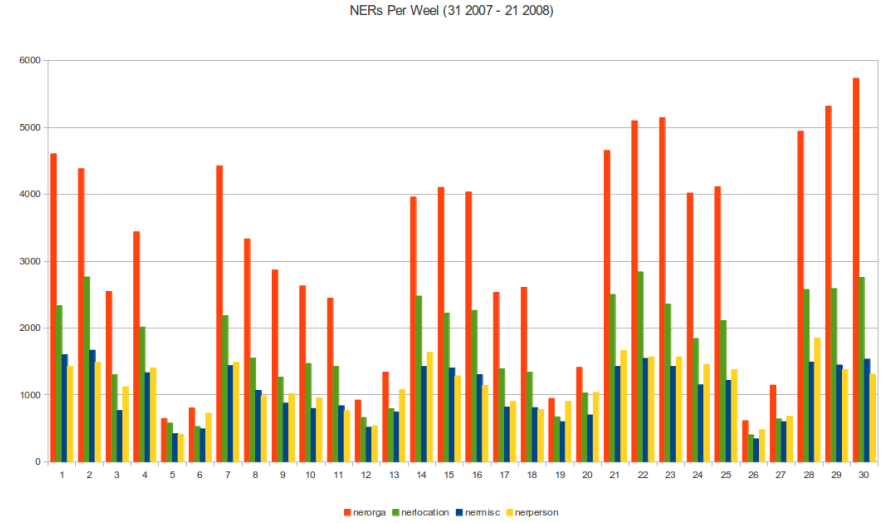


Figure 13.6: Number of NERs in April 2008.

13.3.3 Trend indication

The trend indication method as presented in Chapter 10 contains several steps based on the particular weighting functions. In the following we focus on the interestingness and the outliers values.

Interestingness

In Table 13.1, the summary of the interestingness values per month in the corpus is presented. Below, we summarize the most interesting words from the beginning of our time window (September 2007) and from the end (May 2008).

	Terms	Low.val.	High.val.	Diff.val.	Days
September	2217	4.51E-04	0.101534	72	8
October	2518	3.97E-4	0.140191	90	13
November	2368	4.22E-04	0.140625	86	9
December	2607	3.84E-04	0.133154	80	5
January	1992	5.02E-04	0.149022	72	3
February	1757	5.69E-04	0.06716	46	6
March	3507	2.85E-04	0.098945	93	9
April	2746	3.62E-04	0.122101	87	2
May	2395	4.8E-04	0.13737	80	2

Table 13.1: Interestingness values: summary

The most interesting terms from the test corpus in September 2007 according to the function output are presented below (21 highest scored terms):

```
organisch 4.512635379061372E-4 24 Sep 2007 22:00:00 GMT
filial 9.025270758122744E-4 18 Sep 2007 22:00:00 GMT
gebrach 0.0018050541516245488 23 Sep 2007 22:00:00 GMT
dresd 9.025270758122744E-4 23 Sep 2007 22:00:00 GMT
int 0.0013537906137184115 17 Sep 2007 22:00:00 GMT
abgeschloss 9.025270758122744E-4 23 Sep 2007 22:00:00 GMT
ins 9.025270758122744E-4 19 Sep 2007 22:00:00 GMT
jobgarantie 4.512635379061372E-4 23 Sep 2007 22:00:00 GMT
edelsteinvorkomm 4.512635379061372E-4 23 Sep 2007 22:00:00 GMT
veroffentlich 0.002256317689530686 18 Sep 2007 22:00:00 GMT
inn 0.0013537906137184115 23 Sep 2007 22:00:00 GMT
diamantbohrprogramm 4.512635379061372E-4 23 Sep 2007 22:00:00 GMT
verlier 9.025270758122744E-4 17 Sep 2007 22:00:00 GMT
san 9.025270758122744E-4 24 Sep 2007 22:00:00 GMT
bedeu 0.002707581227436823 17 Sep 2007 22:00:00 GMT
ing 0.004061371841155234 23 Sep 2007 22:00:00 GMT
visio 9.025270758122744E-4 18 Sep 2007 22:00:00 GMT
```

```
sal 0.0013537906137184115 29 Sep 2007 22:00:00 GMT
nord 0.002256317689530686 18 Sep 2007 22:00:00 GMT
ind 4.512635379061372E-4 18 Sep 2007 22:00:00 GMT
inc 0.007671480144404332 17 Sep 2007 22:00:00 GMT
```

The most interesting terms from the test corpus in May 2008 according to the function output are presented below (21 highest scored terms):

```
haltensw 4.175365344467641E-4 3 May 2008 22:00:00 GMT
einig 0.0016701461377870565 3 May 2008 22:00:00 GMT
insb 4.175365344467641E-4 30 Apr 2008 22:00:00 GMT
anlageurteil 4.175365344467641E-4 30 Apr 2008 22:00:00 GMT
bestatig 0.01837160751565762 30 Apr 2008 22:00:00 GMT
seak 4.175365344467641E-4 30 Apr 2008 22:00:00 GMT
konjunkturunabhangig 4.175365344467641E-4 30 Apr 2008 22:00:00 GMT
handel 0.0037578288100208767 30 Apr 2008 22:00:00 GMT
vorfeld 4.175365344467641E-4 3 May 2008 22:00:00 GMT
fertigstellung 4.175365344467641E-4 30 Apr 2008 22:00:00 GMT
wohl 4.175365344467641E-4 30 Apr 2008 22:00:00 GMT
uberschatt 4.175365344467641E-4 3 May 2008 22:00:00 GMT
kalliwoda 0.0025052192066805845 3 May 2008 22:00:00 GMT
bank 0.009603340292275574 30 Apr 2008 22:00:00 GMT
colorado 4.175365344467641E-4 3 May 2008 22:00:00 GMT
mess 4.175365344467641E-4 30 Apr 2008 22:00:00 GMT
research 0.01837160751565762 30 Apr 2008 22:00:00 GMT
guidanc 0.0029227557411273487 30 Apr 2008 22:00:00 GMT
cbs 0.0029227557411273487 30 Apr 2008 22:00:00 GMT
empfiehl 0.003340292275574113 3 May 2008 22:00:00 GMT
int 0.006680584551148226 30 Apr 2008 22:00:00 GMT
```

Since we experiment with the stemmed corpus, the output appears also stemmed. However, a back-stemming is possible.

Outliers

In Table 13.2, the summary of the outliers values per month in the corpus is presented. Below, we summarize the most interesting words from the beginning of our time window (September 2007) and from the end (May 2008).
The most outlying terms from the test corpus in September 2007 according to the function output are presented below (21 highest scored terms):

```
usd 72.22016172500157 17 Sep 2007 22:00:00 GMT
onvista 73.09600066456208 23 Sep 2007 22:00:00 GMT
nutrisyst 73.09600066456208 17 Sep 2007 22:00:00 GMT
```

	Terms	Low.val.	High.val.	Diff.val.	Days
September	226	1.385988	144.8752	526	8
October	2518	2.235465	222.7522	573	13
November	2358	2.838428	268.5888	556	9
December	2606	2.235465	198.8269	553	5
January	1991	2.080743	145.6746	525	3
February	1757	0.021026	194.2567	402	6
March	3507	2.078982	192.0817	630	9
April	2749	3.148583	221.3823	581	2
May	2384	2.713265	163.2804	535	2

Table 13.2: Outlier values: summary

```
sich 73.43245210287166 17 Sep 2007 22:00:00 GMT
allianx 73.82259990069717 18 Sep 2007 22:00:00 GMT
jahr 73.91827059992292 17 Sep 2007 22:00:00 GMT
commerzbank 80.333315608422 23 Sep 2007 22:00:00 GMT
bank 80.68549091335093 17 Sep 2007 22:00:00 GMT
den 84.34413901525676 17 Sep 2007 22:00:00 GMT
mrd 84.84496256129766 17 Sep 2007 22:00:00 GMT
mio 93.92502430125977 17 Sep 2007 22:00:00 GMT
abn 93.98361640900661 17 Sep 2007 22:00:00 GMT
amro 93.98361640900661 17 Sep 2007 22:00:00 GMT
eur 95.70093254178073 17 Sep 2007 22:00:00 GMT
tru 106.32145551209031 23 Sep 2007 22:00:00 GMT
gem 106.32145551209031 23 Sep 2007 22:00:00 GMT
north 106.32145551209031 23 Sep 2007 22:00:00 GMT
web 107.13498820102258 23 Sep 2007 22:00:00 GMT
gerry 107.13498820102258 23 Sep 2007 22:00:00 GMT
deutsch 109.40934625794432 17 Sep 2007 22:00:00 GMT
euro 144.87519080947405 17 Sep 2007 22:00:00 GMT
```

The most outlying terms from the test corpus in May 2008 according to the function output are presented below (21 highest scored terms):

```
funkwerk 79.74109163406773 3 May 2008 22:00:00 GMT
den 80.72939020031718 30 Apr 2008 22:00:00 GMT
euro 81.39055663453598 30 Apr 2008 22:00:00 GMT
chemie 84.67486078710405 30 Apr 2008 22:00:00 GMT
unvera 85.86711063094654 30 Apr 2008 22:00:00 GMT
york 86.19351058928687 30 Apr 2008 22:00:00 GMT
upda 86.83673208391096 30 Apr 2008 22:00:00 GMT
angehob 86.95852142700998 30 Apr 2008 22:00:00 GMT
```

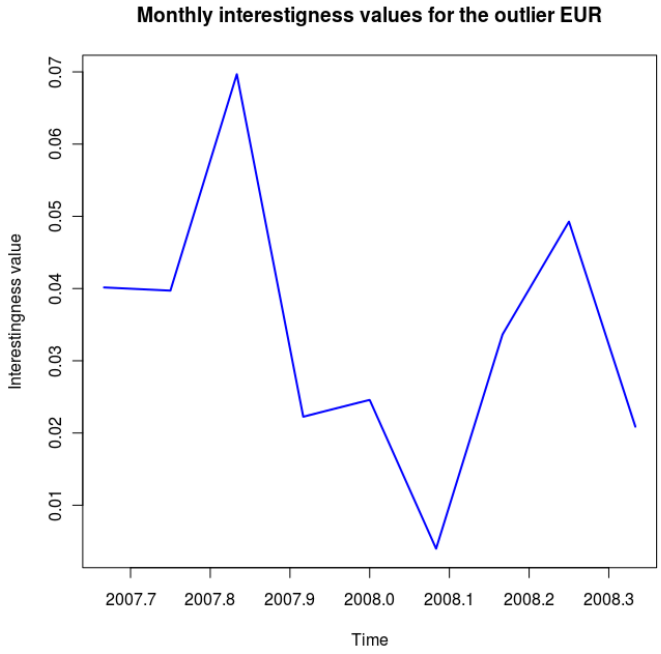


Figure 13.7: The most interesting values of term EUR.

```
palmol 93.03127357307902 30 Apr 2008 22:00:00 GMT
hab 95.53175471267494 30 Apr 2008 22:00:00 GMT
wack 97.06663000555179 30 Apr 2008 22:00:00 GMT
quartal 106.6943988580011 30 Apr 2008 22:00:00 GMT
citigroup 112.65798292066195 30 Apr 2008 22:00:00 GMT
eur 116.32613352061279 30 Apr 2008 22:00:00 GMT
mio 116.73538734585144 30 Apr 2008 22:00:00 GMT
erst 121.05107972402328 30 Apr 2008 22:00:00 GMT
freseniu 123.16867520451093 30 Apr 2008 22:00:00 GMT
gfk 124.99081956785967 30 Apr 2008 22:00:00 GMT
lehma 127.27165593304996 30 Apr 2008 22:00:00 GMT
broth 144.83483676980154 30 Apr 2008 22:00:00 GMT
usd 163.28036563913398 30 Apr 2008 22:00:00 GMT
```

For the two outliers, EUR and USD, we checked their interestingness values over the time window. Figure 13.7 presents the curve for the monthly interestingness value of EUR, and Figure 13.8 for the USD.

What we can follow from the brief presentation of the trend indication method's result is that applying the interestingness value, we can extract the terms that are potentially relevant for the trend. Applying the outliers weighting function it is possible to recognize the potential triggers for a trend. In particular, the terms: *USD*, and *Lehman Brothers* that appeared to be the most outlying in April and

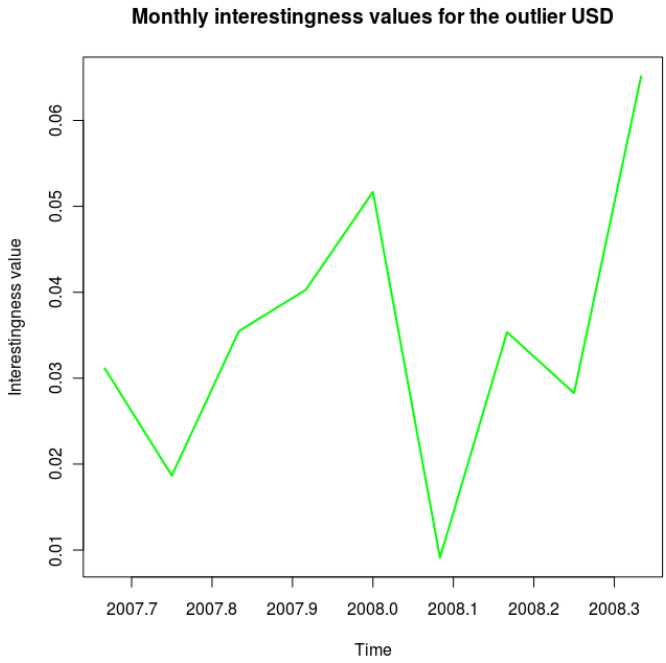


Figure 13.8: The most interesting values of term USD.

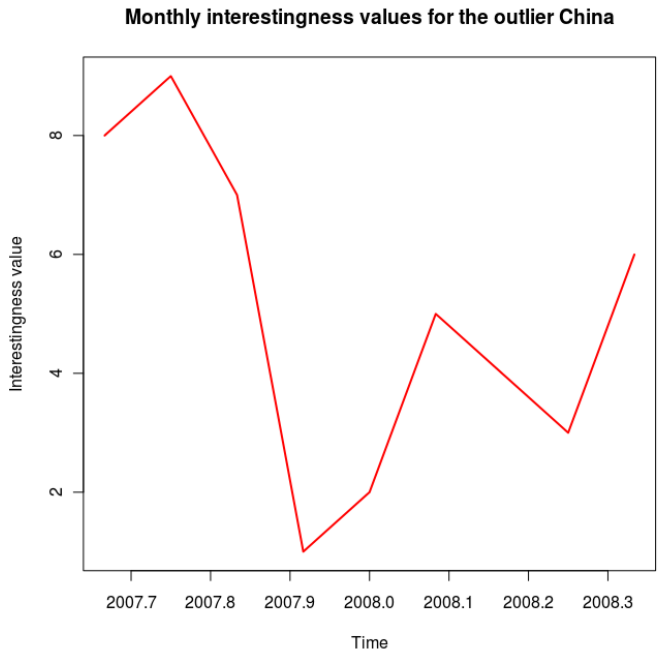


Figure 13.9: The most interesting values of term China.

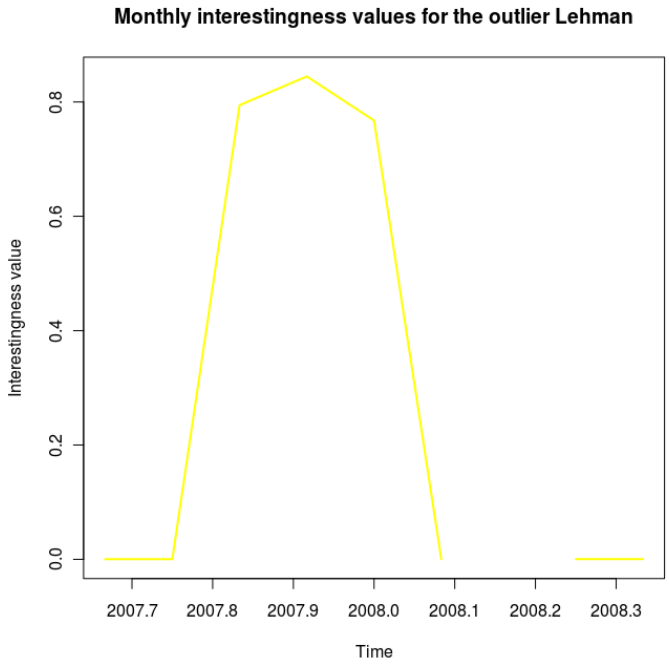


Figure 13.10: The most interesting values of term Lehman.

May 2008, were indeed significant in German business news few months before the Lehman Brothers insolvency (in September 2008) was announced. Figure 13.10 illustrates the interestingness for *Lehman*. Similar to the *USD*, *Lehman* appears to be on-topic in the last quarter 2007. Combining the respective most outlying terms with the most interesting once can bring some more insights into the particular emerging topics in the test corpus. Regarding the curves presented in Figures 13.7 and 13.8, we can follow that the analyzed news are more related to the EUR than to the USD (EUR's interestingness value lies above the USD's value) however, both terms are significantly important in the test corpus.

13.3.4 Trend ontology

Based on the trend ontology as described in Chapter 9, we tested the corpus particularly with regard to extracting information about the companies mentioned in it. In the following, we present the examples of the topic fields for three terms: Germany, Google, and Citigroup that were recognized as relevant topics according to our trend ontology.

```
trendonto:#Germany (9137) has Topic
trendonto:#Financial : 1142
trendonto:#buy : 1003
trendonto:#MachineBuildingIndustry : 650
```

```
trendonto:#Share : 606
trendonto:#StockPrice : 562
trendonto:#Up : 520
trendonto:#Industry : 510
trendonto:#Investment : 468
trendonto:#Supplier : 422
trendonto:#AutomobilIndustry : 414
```

The terms at the top of the examples are the main concepts that are recognized as topics. Depending on their structure, the sub-concepts derived based on the ontology, are additionally presented with their frequency values. Regarding the concept *Citigroup*, based on the simple ontology test, we can follow that this concept is a topic with the structure that contains the following terms (that are also ontology concepts themselves): *USA*, *service*, *bank*, *broker*, *buy*, *Germany*, *sell*. The values provided with the concepts describe their weight in the document corpus – the higher the value, the stronger the appearance in the corpus and the probability that the given concept is an emerging topic.

```
trendonto:#Google (154) has Topic
trendonto:#USA : 40
trendonto:#Service : 40
trendonto:#InformationTechnologies : 40
trendonto:#Stock : 16
trendonto:#Hold : 6
trendonto:#Share : 6
trendonto:#Up : 4
trendonto:#StrongUp : 2
```

Additional to the topic fields we focused on the buy and sell signals for the respective companies. These are the terms *up*, *hold*, *sell* which express the respective stock recommendation contained in the news.

```
trendonto:#Citigroup (344) has Topic
trendonto:#USA : 62
trendonto:#ServiceProvider : 54
trendonto:#Bank : 42
trendonto:#Broker : 42
trendonto:#Buy : 36
trendonto:#Germany : 14
trendonto:#Sell : 10
```

The term *China* was one of the outliers appearing throughout the time window. The curve based on its interestingness value is shown in Figure 13.9. It looks like developing in the opposite direction to the USD curve in the beginning of the

time window – the more web news are about China the less are talking about USD. From 2008 January it appears that China is the frequent term in the news about DAX and USD is off-topic. Below we present the topic structure for the terms China and USA.

```
trendonto:#China (152) has Topic structure:

trendonto:#ServiceProvider : 20
trendonto:#Up : 16
trendonto:#FederalBond : 12
trendonto:#MAN : 10
trendonto:#DowJones : 10
trendonto:#hold : 8
trendonto:#Share : 8
trendonto:#Hongkong : 6
trendonto:#Company : 6
trendonto:#AutomobileIndustry : 6
trendonto:#Industry : 4
trendonto:#EU : 4
trendonto:#Crisis : 2
trendonto:#Asia : 2
trendonto:#Strong_Down : 2

trendonto:#USA (339) has Topic structure:

trendonto:#Share : 58
trendonto:#Down : 28
trendonto:#Up : 18
trendonto:#Company : 16
trendonto:#Kurs : 16
trendonto:#FederalBond : 16
trendonto:#InformationTechnologies : 16
trendonto:#_Neutral : 10
trendonto:#Light : 10
trendonto:#Euro : 8
trendonto:#AutomobileIndustry : 8
trendonto:#Pharma : 8
trendonto:#MAN : 6
trendonto:#Bank : 6
trendonto:#EU : 6
trendonto:#Volkswagen : 5
trendonto:#Crisis : 4
```

```
trendonto:#Commerzbank : 4
```

For the evaluation purposes regarding usefulness and practicability the trend ontology has been filled with two independent corpora of stock market specific documents. They contain over 5,000 and 16,000 documents respectively (subsequently first and second corpus) in German language. The documents of the first corpus are share and market analyses by professional stock market analysts, the second a more general corpus consists of financial blog entries. Several basic questions have been identified as relevant for trends in general and specifically stock market trends.

13.3.5 Topic models

We tried out different settings for testing the topic models. In order to have more overview of the document corpus, we included also tests with several documents that were time-stamped with the dates earlier than the main test corpus, e.g. 2005. Figure 13.11 shows a timeline with a snippet of topics from 2006 and 2007 presented on it. The topics, represented by single terms, are grouped together. The red marked topics are the new topics in the corpus, and the blue marked topics are the topics that repeat over the corpus. The topics: *dax*, *deutlich*, *mittag*, *nachricht*, *neu*, *schluss*, *spitz*, *volkswagen* are the repeating topics in 2007 and topics: *allianz*, *nachricht*, *network*, *nokia*, *plan*, *tabelle*, *tochter*, *zahl* are the emerging once in 2007. The timeline snippet presented in Figure 13.11 is based on the tests with 4,696 documents from DAX, using the two-years time window and the time slice of one month. The number of topics[2] is 10 per year and 3 per month. We mainly analyzed the documents' titles in this experiment.

Figure 13.11: Timeline 2006 to 2007 – a snippet

[2]this is one of the parameter in the implementation

Figure 13.12 presents the prediction of the possible upcoming topics based on the results calculated by the topic modeling experiment presented above. We applied a similarity function to derive the potential upcoming topics. The Figure 13.12 shows the resulting similar topics, grouped together according to their similarity.

Mögliche ähnliche Topics

Topic 03 vom 2007:	dax; deutlich; mittag; nachricht; neu; schluss; spitz; volkswagen;
Topic 04 vom 2005:	allianz; **dax**; **deutlich**; melde; morgen; **nachricht**; schwach; **spitz**;
Topic 09 vom 2006:	behaupten; **dax**; freundlich; melde; **mittag**; **neu**; **spitz**; thyssenkrupp;
Topic 04 vom 2007:	bank; bayer; china; dresden; nachricht; sieh; update; zwei;
Topic 02 vom 2006:	**bank**; **bayer**; **china**; euro; hälse; merck; **nachricht**; **zwei**;
Topic 07 vom 2007:	chef; nachricht; press; roundup; siemens; steig; volkswagen; weit;
Topic 08 vom 2008:	auftrag; bank; bestätigen; **nachricht**; neu; **press**; **steig**; **volkswagen**;
Topic 09 vom 2007:	ceo; dgap; erhäl; löschen; mio.; nachricht; siemens; thyssenkrupp;
Topic 02 vom 2007:	analyse; **dgap**; eur; euro; **nachricht**; **siemens**; **thyssenkrupp**; vdo;
Topic 10 vom 2008:	dresden; **erhäl**; großauftrag; **mio.**; **nachricht**; **siemens**; stell; weit;
Topic 08 vom 2008:	auftrag; bank; bestätigen; nachricht; neu; press; steig; volkswagen;
Topic 07 vom 2007:	chef; **nachricht**; **press**; roundup; siemens; **steig**; **volkswagen**; weit;
Topic 10 vom 2008:	dresden; erhäl; großauftrag; mio.; nachricht; siemens; stell; weit;
Topic 02 vom 2005:	deutsch; **erhäl**; intern; **nachricht**; schließ; **stell**; web.de; **weit**;

Figure 13.12: Topics in 2006 and 2007 – a snippet

In Figure 13.13, a part of the timeline with topics from the corpus in 2005 and 2007 is presented. Important seems the content of the topics that starts to appear in November 2007. There are predictions of upcoming topics: *emfi*, *gold*, *heute*, *hongkong*, *jahr*, *london*, *peking*, *prozent*, *rohstoff*. However, the results of the topic model are particular topics listed in the groups that are difficult to interpret without any further information.

Figure 13.14 shows grouped similar topics from 2007 and 2008. This analysis was in particular based on the keywords-based description of the web documents. We can see, that the certain keywords: *aktie*, *aktiencheck*, *isin* are repeating and therefore make a reliable prediction even more difficult.

13.3.6 K-means clustering

The k-means clustering method has been tested on 35,804 documents analyzing the document's title and the document's body. Different numbers of cluster were chosen and the number of 100 clusters and 100 keywords per vector were helpful in achieving informative results. The following cluster summarization shows

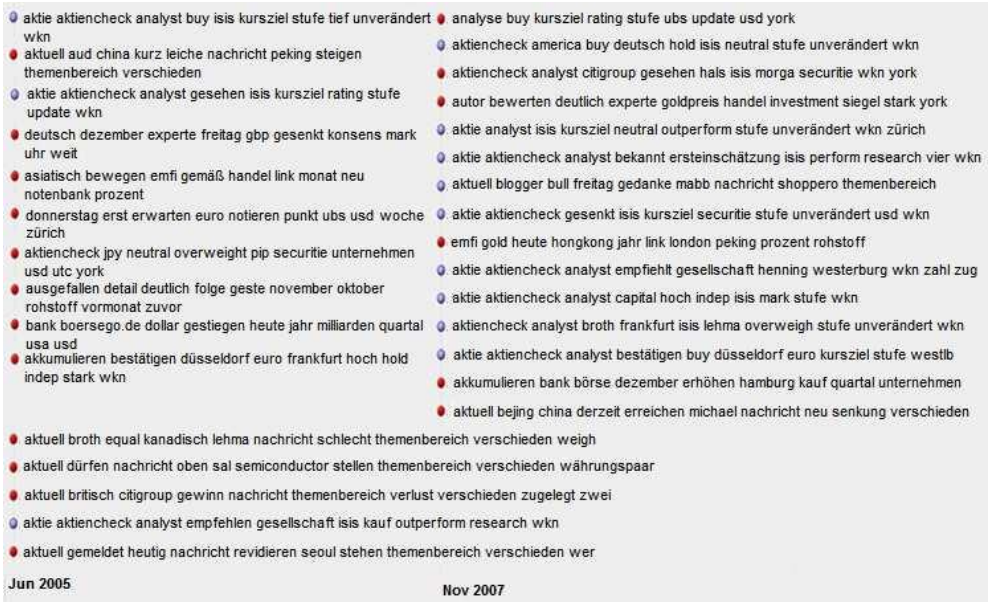


Figure 13.13: Timeline – a snippet with topics in November 2007

exemplary the results of document's description keywords and abstracts clustering. The results are interesting but less useful, since many keywords are not informative regarding the potential trends.

```
Year 2005, size: 66

0 archiv letzt podcast blog bull imagelooop
1 meiersonline.d startseit qype.com meiersworld.d fotos lieblingscafé
2 blog rss navigation eintrag itun frankfurt

Year 2006, size: 106

0 bull archiv letzt eintrag erweitert shoppero
1 blog meiersonline.d rss startseit weblog meiersworld.d
2 feed navigation lieblingscafé qype.com itun podcast

Year 2007, size: 12763

0 borsego.d punkt nachricht heut gegenub erwartet
1 aktull verschied themenbereich nachricht euro log
2 isin akti wkn usd analyst aktiencheck

Year 2008, size: 22869
```

Topic 02 vom 11.2007: aktie; aktiencheck; analyst; isis; rating; stufe; unverändert; update; wkn; york;
Topic 11 vom 09.2007: **aktie**; **aktiencheck**; **analyst**; **isis**; neutral; oil; research; schwäche; **wkn**; **york**;
Topic 02 vom 12.2007: **aktie**; **aktiencheck**; **analyst**; buy; **isis**; kursziel; **stufe**; tief; **unverändert**; **wkn**;
Topic 04 vom 12.2007: **aktie**; **aktiencheck**; **analyst**; gesehen; **isis**; kursziel; **rating**; **stufe**; **update**; **wkn**;
Topic 06 vom 01.2008: **aktie**; **analyst**; **isis**; kursziel; neutral; outperform; **stufe**; **unverändert**; **wkn**; zürich;
Topic 08 vom 01.2008: **aktie**; **aktiencheck**; gesenkt; **isis**; kursziel; securitie; **stufe**; **unverändert**; usd; **wkn**;
Topic 11 vom 01.2008: **aktie**; **aktiencheck**; **analyst**; capital; hoch; indep; **isis**; mark; **stufe**; **wkn**;
Topic 12 vom 01.2008: **aktiencheck**; **analyst**; broth; frankfurt; **isis**; lehma; overweigh; **stufe**; **unverändert**; **wkn**;
Topic 10 vom 03.2008: **aktie**; **aktiencheck**; **analyst**; **isis**; **rating**; **stufe**; **update**; usd; **wkn**; **york**;
Topic 12 vom 03.2008: **aktie**; **aktiencheck**; **analyst**; geschäftsjahr; **isis**; kauf; kursziel; research; **unverändert**; **wkn**;
Topic 02 vom 04.2008: **aktie**; **aktiencheck**; **analyst**; bestätigen; eur; frankfurt; **isis**; research; **stufe**; **wkn**;
Topic 06 vom 04.2008: **aktie**; **analyst**; buy; kursziel; **rating**; **stufe**; **unverändert**; **update**; usd; **wkn**;
Topic 08 vom 05.2008: **aktie**; **aktiencheck**; **analyst**; buy; **isis**; **rating**; **stufe**; **unverändert**; **wkn**; **york**;
Topic 11 vom 05.2008: **aktie**; **aktiencheck**; **analyst**; buy; kursziel; outperform; **stufe**; **update**; usd; **wkn**;
Topic 13 vom 05.2008: **aktie**; **aktiencheck**; **analyst**; hals; **isis**; kauf; rät; umsatz; **unverändert**; **wkn**;

Topic 02 vom 12.2007: aktie; aktiencheck; analyst; buy; isis; kursziel; stufe; tief; unverändert; wkn;
Topic 02 vom 11.2007: **aktie**; **aktiencheck**; **analyst**; **isis**; rating; **stufe**; **unverändert**; update; **wkn**; york;
Topic 04 vom 12.2007: **aktie**; **aktiencheck**; **analyst**; gesehen; **isis**; **kursziel**; rating; **stufe**; update; **wkn**;
Topic 03 vom 01.2008: **aktiencheck**; america; **buy**; deutsch; hold; **isis**; neutral; **stufe**; **unverändert**; **wkn**;
Topic 06 vom 01.2008: **aktie**; **analyst**; **isis**; **kursziel**; neutral; outperform; **stufe**; **unverändert**; **wkn**; zürich;
Topic 08 vom 01.2008: **aktie**; **aktiencheck**; gesenkt; **isis**; **kursziel**; securitie; **stufe**; **unverändert**; usd; **wkn**;

Figure 13.14: Topics September 2007 to May 2008 – a snippet

```
0 new deutsch erst dollar quartal unternehm
1 aktull verschied themenbereich nachricht bull com
2 aktiencheck isin wkn akti analyst unverandert
```

More informative cluster structure can be achieved by clustering the documents according to the document's specific description keywords. Below we present the selected clusters from the analysis on the general news corpus from September 2007 to May 2008.

```
Cluster 1

aktie börsenbrief daytrading devise empfehlen indiz kauf kostenlos

Cluster 52

rohstoff finanz bank buy verkauf kauf usd analyst bau

Cluster 61

deutschland halt devise bank buy neu dienstleist akkumulieren

Cluster 78
```

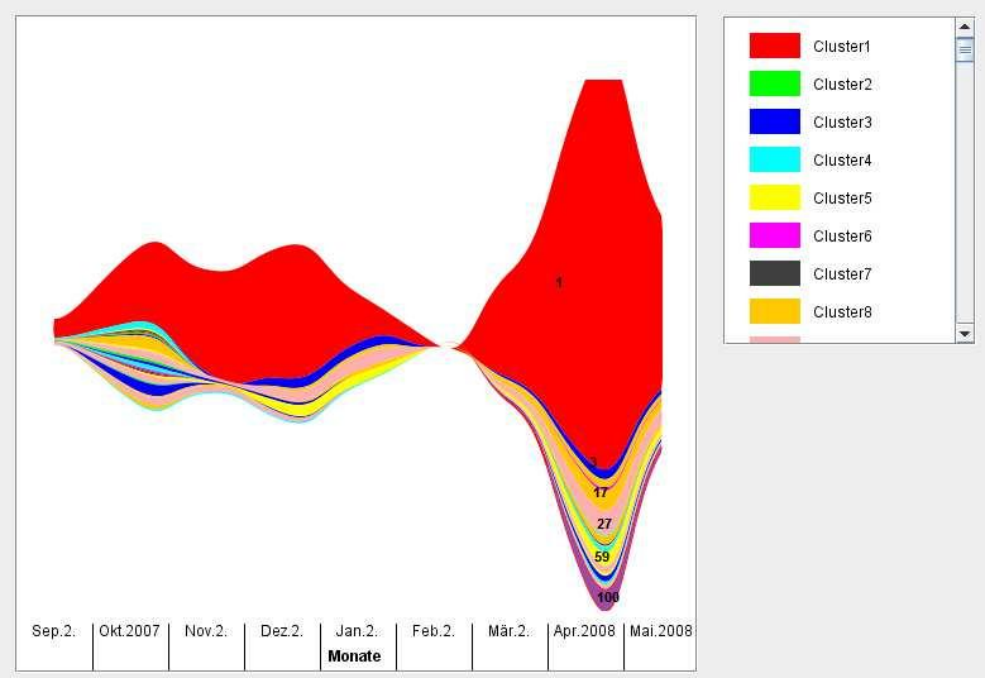


Figure 13.15: Clustering documents 2007 to 2008 – ThemeRiver visualization

Cluster:	Topwörter
1:	aktuell (0.45); boersego.de (0.04); deutsch (0.03); neu (0.03); heute (0.03); futur (0.02); bull (0.02); erst (0.02); prozent (0.02)
3:	aktiencheck (1.0); isin (1.0); new (1.0); york (1.0); wkn (1.0); aktie (0.99); analyst (0.98); stufe (0.88); usd (0.63); unveränder
17:	aktie (1.0); dollar (0.31); analyst (0.22); erst (0.15); stufe (0.09); frankfurt (0.06); boersego.de (0.05); heute (0.04); bank (0.03
27:	isin (1.0); aktiencheck (1.0); wkn (1.0); aktie (0.98); analyst (0.88); unverändert (0.28); hamburg (0.18); erst (0.16); stufe (0.1
59:	aktiencheck (1.0); isin (1.0); wkn (1.0); analyst (0.99); aktie (0.98); unverändert (0.83); stufe (0.78); usd (0.62); eur (0.3); hoc
100:	isin (1.0); wkn (1.0); aktiencheck (0.99); erst (0.28); freitag (0.09); analyst (0.06); deutlich (0.06); deutsch (0.06); notiert (0.0

Figure 13.16: Topics from clustering 2007-2008

```
usa software neutral  outperform buy overweight market internet
```

The different clusters can be interpreted as belonging to the general topics. However, it is difficult to understand the topics without further information about the particular terms identified as cluster components. In Figure 13.15 we illustrate the clustering results from another experiment by using the ThemeRiver [Havre et al., 2002] visualization tool. A visualization tool is very helpful for interpreting the potential trends.

13.4 Summary

From NER's distribution in the corpus, we can derive the information about the potential topics contained in the documents of the test corpus. This information

is only about whether the news are reporting more on organizations or locations or persons.

Applying the trend-indication weighting functions such as outliers or interestingness we can extract the information about the potential trend triggers. These are particular terms that, classified into the groups by their respective values, belong to a general emerging topic. Additionally, we can observe how they vary in the corpus over time regarding their interestingness values. More possibilities are given by applying a combination of the different weighting functions, in particular the additional information about the terms and their trend-indication values.

Applying the trend ontology to the test corpus we can derive the topic structures for the topics appearing in the given time window. The topic structure, consisting of fields of concepts that are mentioned in the news similar frequently, gives us information about the particular topics. Additionally, the concepts' values enhance the information about the topic structures.

Applying topic models to the test corpus we can extract topics that describe the documents and, based on the topics in the corpus, predict the upcoming topics. Applying similarity functions to the topic model's results, we can derive the information about which topic groups belong together and, based on it, enhance the prediction of the potential trends. The topics are described by words and does not contain any further information.

Applying the k-means clustering method to the test corpus, we can extract the information about the topics cluster over the time. With additional visualization tool, we can illustrate how the clusters are changing over time. The summarization of the two general features is given in the Table 13.3. Our methods offer a good interpretability and the possibility of prediction. The LDA topic models are good at prediction but they lack in interpretation possibility. The k-means method is interpretable, however the possibility for the interpretation is weak.

	Interpretability	Predictivity	
Trend-indication	+✓	+✓	
Trend ontology	+✓	-✓	
LDA topic model		+✓	
k-means clustering	-✓		

Table 13.3: Summary of the algorithms' features

CHAPTER 14

Outlook

This chapter summarizes the thesis. It gives an overview of the research done, describes what the research was about and names the critical aspects of the research results presented in this thesis. It ends with a summary of the issues that are left to be the subject of future work.

14.1 Summary

This thesis is devoted to trend mining. It sought a precise definition for trend mining, and asked whether a knowledge-based approach for mining trends is useful for the problem of mining trends in texts. Based on the relevant work, the goal of this research was to define trend mining mainly focusing on texts, to summarize research relevant for trend mining, to offer the theoretical frame, to look for a general trend model and to propose a knowledge-based approach for trend mining. These goals have been fulfilled and at the end of this thesis two main aspects are summarized in 14.1.1 and 14.1.2. The end of one project is always a good opportunity for beginning another research project, particularly because there is a possibility to learn from the critical aspects of the most recent research. And based on the answers to the questions asked in the beginning, almost always new questions emerge at the end.

In 14.2 we show the limits and summarize the critical issues of this research and in 14.3 we discuss the possible future work on trend mining.

14.1.1 Trend mining as a research field

Trend mining is a research field in its initial stage, existing in several different approaches distributed among the information retrieval, data mining and temporal data mining research. Trend mining can evolve into a separate research field. The first half of this thesis was focused on the methods and approaches from the

related research, on the definitions of a trend and trend mining. It showed how approaches from information retrieval can be adopted for mining trends in texts. The dimensions of time and knowledge are important for the analysis of many kinds of texts and data with a temporal aspect (and whenever a trend in the data progresses over time is relevant). Trend mining, in this thesis based on text and data, can be understood in a broader context, including multimedia and video data. Obviously, trend mining in texts requires different methods than trend mining in the real valued data or in the video analysis. However, the methods are based on similar assumptions about trends. Developing more research on trend mining could bring more interesting insights into the possibilities that trend mining methods bring for different use cases. In particular, we see two possible main directions for how to sustain the trend mining research – one is multidisciplinary, and the other is the engineering direction. As for the multidisciplinary direction, it would be interesting to combine more insights from sociology and financial data analysis as a basis for trend mining algorithms. To continue the research in the engineering direction would mean to focus on practical solutions for the different particular use cases, and different particular data sets.

In Section 4.2.3 we proposed a definition of a trend in the context of this research. When extending the scope of the trend mining, we should consider also an extended understanding of what a trend is. In general, to mine a trend means to observe patterns of changes that are based on certain variables (e.g. people, numbers, words, images) and which lead to a general change – the emerging trend – in the system which is depending on these variables.

14.1.2 Knowledge-based approaches to trend mining

Basically, our approach assumes that mining trends with knowledge incorporates the expectation that we are looking for the trend's trigger and the trigger's context, and that we can relate them both, concluding the trend's amplitude within a given time window, all of this will help us in understanding the trend (in texts). The proposed trend template shows a formalization of this idea, and the trend ontology and trend indication methods show its possible implementation. From the experimental part of this research we see what advantages this approach brings for the interpretation of the trend mining results. On the other hand, we can clearly see that the probabilistic topic modeling can be successfully applied for finding out different topics emerging in different time periods over a text corpus, thus for finding the trends. A combination of the knowledge-based and probability-based approaches could be very useful for trend mining. It would be interesting to find out how much a priori knowledge is good and useful for the task of mining trends while experimenting with different use cases.

14.2 Critical aspects

In general, we identify the following critical aspects of our research:

1. One critical aspect is the problem of the so called *"ex-post" model* versus the so called *"real-time" model*. An ex-post model for mining trends relies on a data set from the past, that is processed offline. This model does not contain any time restriction, can be performed offline at any time and serves as the ex-post analysis of a trend. A knowledge-based approach is, if based on the ontology, always somehow related to an ex-post model in which we learn from the historical data – an ontology has to be created offline and, once created, it represents the "old" knowledge. However, it is possible to adjust this approach for a real-time analysis (see also our discussion in Section 7.2).

2. Another critical factor is the amount of data that can be mined for trends. We tested our approach on a rather *small* data set, consisting of at most 40,000 web documents. Most of the test sets used for the trend mining tests are rather small (to our best knowledge, the largest test sets are less than 50,000 web documents). Going from the small into the *"big data"* would increase the requirements on the algorithms' performance and complexity. In our research we didn't put a high expectation on decreasing the complexity of our algorithms. It would be interesting to develop more implementations of the trend template and to apply them to available big data sets while focusing on the best possible complexity classes of the implemented algorithms (if there are any public big data sets available for tests).

3. Directly connected to the data set size, the problem of *amplified* trends as well as the emergence of *parallel trends* arises, which is another critical issue of trend mining. In particular, when we mine trends in test sets from different sources or if the test sets are based on social network data, most probably the phenomenon of trend amplification and parallel trends will occur. On the one hand, it is then possible to mine several different trends in one data set. On the other hand, some less important trends may be amplified through the users or sources just repeating particular news. Considering the aspects of amplified trends, or parallel trends could enhance the knowledge-based trend mining approach.

4. The distinction between the parallel trends and the identification of the amplified trends is possible by taking *users' feedback* into the evaluation of trend mining methods. The evaluation of trend mining research has been described in [Kontostathis et al., 2003] as critical. In this thesis, we offer an experimental evaluation, creating the use case based goal free experimental

setting. However, user feedback on trend mining results would definitely enhance the evaluation process.

The last issue that should be mentioned here, *sentiments* vs. *trends*, is not really critical, but it could become so if there is no clarity about the difference between *sentiment mining* and trend mining. Often, the expectation for trend mining is the same as on for sentiment mining. This is problematic. Sentiment mining is about mining the emotional value (e.g. positive or negative customer feedback) of statements. Trend mining is about mining the emerging topics from texts. Sentiment mining could be used in order to understand trends, but mining sentiments is not equal to mining trends. The use case of mining trends in market research that we describe in 11.1.2 is suitable for both, sentiment mining and trend mining.

14.3 Future work

We close this thesis with several different ideas for extending our research. In the following we summarize the main future work issues:

1. It would be interesting to work on a combination of methods from mathematical trend analysis and from common techniques of the chart analysis for financial markets data into the text-based trend mining. These could be the regression methods or different trend test methods, i.e. Mann-Kendall trend test, seasonal Kendall test, Spearman's rho test [Yue et al., 2002] Holt-Winter's method [Chatfield, 1978]. These methods, developed explicitly for the real valued time series, could be applied in trend mining.

2. In general, when combining more methods from applied statistics, it would be useful to extend the research on trend mining focusing more on the time series analysis – one possibility of trend analysis in web documents is to create time series from texts and to apply the methods for correlation of the time series, i.e. Box-Pierce test and similar methods [Makridakis et al., 1998] measuring skewness, kurtosis, self-similarity (as discussed in Section 4.1). Regarding the text based time series, the most challenging problem is the selection of web document's attributes on which the time series can be constructed.

3. Different extensions to the idea of the trend template are possible, such as by combining it with a decision tree algorithm or probability based methods. It would be helpful to develop more trend template implementations and to test them.

4. Our test tool presented in Appendix A is still under development and we are currently adjusting it in order to enable more flexibility in testing different

test corpora. It would be valuable to focus on other use cases and trend mining scenarios in order to develop reliable validation techniques for the different tests.

5. It would be helpful to extend the review of the relevant publications, tools and algorithms in trend mining that are currently under development. This would allow for a better understanding of the trend in trend mining research.

Looking at the critical aspects of our research as listed in Section 14.2, every single critical issue can be incorporated in the future work issue.

[0] *List of URLs referred to in this thesis and shortened with service TinyURL.com (accessed on 23-Jul-2013):*

1 https://www.facebook.com/Rockhampton.CQ.Floods as http://tinyurl.com/on2k3lj, p. 27

2 http://www.daviddlewis.com/resources/testcollections/reuters21578/ as http://tinyurl.com/len8xc2, p. 36

3 http://www.godmode-trader.de/artikel/us-dollar-vorsicht-euro,690942 as http://tinyurl.com/mnx83ea, p.108

4 http://www.nytimes.com/2008/03/17/business/17fed.html?_r=2& as http://tinyurl.com/k3xb3lm, p. 108

5 http://video.mit.edu/watch/discovering-the-scientific-method-mit-engineering-k-12-video-pilot-8073/ as http://tinyurl.com/nmwhwlt, p. 157

Bibliography

[Agrawal et al., 1995] Agrawal, R., Wimmers, E. L., and Zait, M. (1995). Querying shapes of histories. In *VLDB '95 Proceedings of the 21th International Conference on Very Large Data Bases*, pages 502–514. Morgan Kaufmann Publishers Inc. [cited at p. 30, 36, 38]

[Allan, 2002] Allan, J., editor (2002). *Topic Detection and Tracking. Event-based Information Organization.* Kluwer academic publishers. [cited at p. 23, 24, 35, 41, 125]

[Allan et al., 1998] Allan, J., Papka, R., and Lavrenko, V. (1998). On-line new event detection and tracking. In *SIGIR'98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37–45. ACM. [cited at p. 30, 35, 37, 38]

[Alnemr, 2012] Alnemr, R. (2012). *Doctoral thesis: Reputation Object Representation Model for Enabling Reputation Interoperability.* Potsdam University. [cited at p. , 58, 59]

[Alnemr and Meinel, 2011] Alnemr, R. and Meinel, C. (2011). From reputation models and systems to reputation ontologies. In *Proceedings of the 5th IFIPTM, Springer IFIP, Copenhagen, Denmark*, pages 98–116. Springer. [cited at p. 91]

[Antoniou and van Harmelen, 2003] Antoniou, G. and van Harmelen, F. (2003). Web Ontology Language: OWL. In *Handbook on Ontologies in Information Systems*, pages 76–92. Springer-Verlag. [cited at p. 58]

[Berkun, 2009] Berkun, S. (2009). *The Myths of Innovation.* O'Reilly Media. [cited at p. 9]

[Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American Magazine*, pages 29–37. [cited at p. 58]

[Blei, 2011] Blei, D. M. (2011). Introduction to probabilistic topic models. http://www.cs.princeton.edu/~blei/papers/Blei2012.pdf. [Online; accessed 27-May-2013]. [cited at p. 52, 53, 54]

[Blei et al., 2003a] Blei, D. M., Jordan, M. I., and Ng, A. Y. (2003a). Hierarchical Bayesian Models for Applications in Information Retrieval. 7:25–44. [cited at p. 36, 55]

[Blei et al., 2003b] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003b). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022. [cited at p. , 34, 53, 56]

[Bolelli et al., 2009] Bolelli, L., Ertekin, S., and Giles, C. L. (2009). Topic and Trend Detection in Text Collections Using Latent Dirichlet Allocation. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR2009, pages 776–780. Springer-Verlag. [cited at p. 34]

[Bolelli et al., 2007] Bolelli, L., Ertekin, S., Zhou, D., and Giles, C. L. (2007). A Clustering Method for Web Data with Multi-type Interrelated Components. In *Proceedings of the 16th international conference on World Wide Web*, WWW2007, pages 1121–1122. ACM. [cited at p. 34, 40]

[Braun et al., 2007] Braun, S., Schmidt, A. P., Walter, A., Nagypál, G., and Zacharias, V. (2007). Ontology Maturing: a Collaborative Web 2.0 Approach to Ontology Engineering. In *CKC*. CEUR-WS.org. [cited at p. 66]

[Brickley and Miller, 2010] Brickley, D. and Miller, L. (2010). Foaf. `http://xmlns.com/foaf/spec/`. [Online; accessed 27-May-2013]. [cited at p. 48]

[Chalmers, 1999] Chalmers, A. (1999). *What Is This Thing Called Science?* University of Queensland Press. [cited at p. 10, 121, 122, 124]

[Chatfield, 1978] Chatfield, C. (1978). The Holt-Winters Forecasting Procedure. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 27(3):264–279. [cited at p. 150]

[Cyger, 2011] Cyger, M. (2011). Online service, free for use on Twitter in 2010. `http://hashtags.org/`. [Online; accessed 01-February-2013]. [cited at p. 4]

[Dai and Davison, 2010] Dai, N. and Davison, B. D. (2010). Mining anchor text trends for retrieval. In *Proceedings of the 32nd European Conference on Information Retrieval (ECIR)*, pages 127–139. Springer. [cited at p. 41]

[Darling, 2011] Darling, W. M. (2011). A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling. Technical report, University of Guelph. [cited at p. 54, 55]

[Dimitrova et al., 2008] Dimitrova, V., Denaux, R., Hart, G., Dolbear, C., Holt, I., and Cohn, A. G. (2008). Involving Domain Experts in Authoring OWL Ontologies. In *Proceedings of the 7th International Conference on The Semantic Web*, ISWC '08, pages 1–16, Berlin, Heidelberg. Springer-Verlag. [cited at p. 66]

[Dodig-Crnkovic, 2002] Dodig-Crnkovic, G. (2002). Scientific Methods in Computer Science. In *Conference for the Promotion of Research in IT at New Universities and at University Colleges in Sweden*. [cited at p. 10, 122, 124, 125]

[Economics, 2011] Economics (2011). Trading Economics. http://www.tradingeconomics.com/germany/stock-market. [Online; accessed 27-June-2013]. [cited at p. , 105]

[Engel et al., 2010] Engel, D., Whitney, P., and Cramer, N. (2010). Events and trends in text streams. In Berry, M. W. and Kogan, J., editors, *Text Mining Applications and Theory*. Wiley. [cited at p. 25, 26, 44, 48, 49]

[Erkens et al., 2012] Erkens, D. H., Hung, M., and Matos, P. P. (2012). Corporate Governance in the 2007-2008 Financial Crisis: Evidence from Financial Institutions Worldwide. *Journal of Corporate Finance*, 18:389–411. [cited at p. 108]

[Facebook, 2013] Facebook (2013). "Facebook's mission is to make the world more open and connected.". https://de-de.facebook.com/. [Online; accessed 01-February-2013]. [cited at p. 4]

[Fahrmeir et al., 2007] Fahrmeir, L., Künstler, R., Pigeot, I., and Tutz, G. (2007). *Statistik. Der Weg zur Datenanalyse*. Springer. [cited at p. 126]

[Fawcett and Provost, 1999] Fawcett, T. and Provost, F. (1999). Activity monitoring: Noticing interesting changes in behavior. In *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, pages 53–62. ACM. [cited at p. 33]

[Genesereth and Nilsson, 1987] Genesereth, M. R. and Nilsson, N. J. (1987). *Logical Foundations of Artificial Intelligence*. Morgan Kaufmann Publishers Inc. [cited at p. 56]

[Godmode, 2011] Godmode (2011). Godmode Trader. http://www.godmode-trader.de/. [Online; accessed 27-June-2013]. [cited at p. , 108, 109]

[Göker and Davies, 2009] Göker, A. and Davies, J. (2009). *Information Retrieval: Searching in the 21st Century*. Wiley. [cited at p. , 44, 46, 47, 48, 126]

[Google, 2011] Google (2011). Online service, GoogleTrends. http://www.google.com/trends. [Online; accessed 01-February-2013]. [cited at p. , 39]

[Goorha and Ungar, 2010] Goorha, S. and Ungar, L. (2010). Discovery of significant emerging trends. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD2010, pages 57–64. ACM. [cited at p. 25, 26]

[Gruber, 2004] Gruber, T. (2004). Every Ontology is a Treaty. In *Interview for Semantic Web and Information Systems SIG of the Association for Information Systems*, volume 1. [cited at p. 56]

[Gruber, 1993] Gruber, T. R. (1993). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. In *In Formal Ontology In Conceptual Analysis And Knowledge Representation*. Kluwer Academic Publishers. [cited at p. 56]

[Guarino, 1998] Guarino, N. (1998). *Formal Ontology and Information Systems*. IOSPress. [cited at p. , 55, 56, 57]

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. http://www.cs.waikato.ac.nz/ml/weka/. [Online; accessed 01-February-2013]. [cited at p.]

[Han and Kamber, 2006] Han, J. and Kamber, M. (2006). *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers Inc. [cited at p. 21, 48, 49, 52, 126]

[Havre et al., 2002] Havre, S., Hetzler, E., Whitney, P., and Nowell, L. (2002). ThemeRiver: Visualizing Thematic Changes in Large Document Collections. *IEEE Transactions on Visualization and Computer Graphics*, 8:9–20. [cited at p. 145]

[Hofmann, 1999] Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57, New York, NY, USA. ACM. [cited at p. 36]

[Jelev, 2010] Jelev, J. (2010). Preprocessing of Documents for Emergent Trend Detection in Text Collections. Master's thesis, Freie Universität Berlin. [cited at p. , 23]

[Jones, 1972] Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21. [cited at p. 47]

[Jowett, 1949] Jowett, B. (1949). *Plato: Meno*. Prentice Hall, Boston. [cited at p. 72]

[Kawamae, 2011] Kawamae, N. (2011). Trend analysis model: trend consists of temporal words, topics, and timestamps. In *Proceedings of the fourth ACM International Conference on Web Search and Data Mining*, WSDM2011, pages 317–326, New York, NY, USA. ACM. [cited at p. 25]

[Kawamae and Higashinaka, 2010] Kawamae, N. and Higashinaka, R. (2010). Trend detection model. In *Proceedings of the 19th International Conference on World Wide Web*, WWW2010, pages 1129–1130, New York, NY, USA. ACM. [cited at p. 25, 34, 35]

[Kontostathis et al., 2003] Kontostathis, A., Galitsky, L., Pottenger, W. M., Roy, S., and Phelps, D. J. (2003). A Survey of Emerging Trend Detection in Textual

Data Mining. In *A Comprehensive Survey of Text Mining*. Springer-Verlag. [cited at p. 23, 24, 26, 30, 34, 36, 37, 38, 39, 41, 92, 149]

[Lavrenko et al., 2000] Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., and Allan, J. (2000). Mining of concurrent text and time series. In *Proceedings of the 6 th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Workshop on Text Mining*, pages 37–44. [cited at p. 30, 33]

[Lehrer, 1974] Lehrer, A. (1974). *Semantic Fields and Lexical Structure*. North-Holland Linguistic Series, 11. North-Holland. [cited at p. 67]

[Lent et al., 1997] Lent, B., Agrawal, R., and Srikant, R. (1997). Discovering trends in text databases. pages 227–230. AAAI Press. [cited at p. 30, 35, 36, 38]

[Lita Lundquist, 2000] Lita Lundquist, R. J. J. (2000). *Language, Text, and Knowledge. Mental Models of Expert Communication*. De Gruyter. [cited at p. 30]

[Maier, 2008] Maier, M. (2008). *Die ersten Tage der Zukunft. Wie wir mit dem Internet unser Denken verändern und die Welt retten können*. Pendo. [cited at p. 19]

[Makridakis et al., 1998] Makridakis, S. G., Wheelwright, S. C., and Hyndman, R. J. (1998). *Forecasting: Methods and Applications*. John Wiley Sons. [cited at p. 150]

[Mathioudakis and Koudas, 2010] Mathioudakis, M. and Koudas, N. (2010). TwitterMonitor: Trend Detection over the Twitter Stream. In *Proceedings of the 2010 International Conference on Management of Data*, SIGMOD2010, pages 1155–1158. ACM. [cited at p. 26]

[MIT, 2011] MIT (2011). Massachusetts institute of technology: Discovering the scientific method. http://tinyurl.com/nmwhwlt. [cited at p. 10]

[Mitsa, 2010] Mitsa, T., editor (2010). *Temporal Data Mining*. Chapman Hall/CRC Press. [cited at p. 21, 22, 32]

[Mittermayer and Knolmayer, 2006] Mittermayer, M.-A. and Knolmayer, G. F. (2006). Newscats: A news categorization and trading system. In *Proceedings of the Sixth International Conference on Data Mining*, pages 1002–1007, Washington, DC, USA. IEEE Computer Society. [cited at p. 37, 38]

[Morinaga and Yamanishi, 2004] Morinaga, S. and Yamanishi, K. (2004). Tracking dynamics of topic trends using a finite mixture model. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD'04, pages 811–816, New York, NY, USA. ACM. [cited at p. 25, 36, 40, 41]

[Naaman et al., 2011] Naaman, M., Becker, H., and Gravano, L. (2011). Hip and trendy: Characterizing emerging trends on Twitter. *J. Am. Soc. Inf. Sci. Technol.*, 62:902–918. [cited at p. 25]

[Neal and Hinton, 1996] Neal, R. M. and Hinton, G. E. (1996). A view of the em algorithm that justifies incremental, sparse, and other variants. In Jordan, M. I., editor, *Learning in graphical models*. Kluwer Academics Publishers. [cited at p. 36]

[NIST, 2013] NIST (2013). Text REtrieval Conference (TREC). http://trec.nist.gov/. [Online; accessed 01-May-2013]. [cited at p. 41, 126]

[Oxford, 2013] Oxford, U. P. (2013). Oxford dictionaries. http://oxforddictionaries.com/. [Online; accessed 23-May-2013]. [cited at p. 24]

[Peramunetilleke and Wong, 2002] Peramunetilleke, D. and Wong, R. K. (2002). Currency exchange rate forecasting from news headlines. *Aust. Comput. Sci. Commun.*, 24(2):131–139. [cited at p. 37, 38]

[Peter Norviq, 2003] Peter Norviq, S. R. (2003). *Artificial Intelligence: A Modern Approach*. Prentice Hall International. [cited at p. 56]

[Petrović et al., 2010] Petrović, S., Osborne, M., and Lavrenko, V. (2010). Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT2010, pages 181–189, Stroudsburg, PA, USA. Association for Computational Linguistics. [cited at p. 35, 40]

[PIA, 2013] PIA (2004-2013). Personal Information Assistant. http://www.pia-services.de. [Online; accessed 01-February-2013]. [cited at p.]

[Pottenger et al., 2001] Pottenger, W. M., bin Kim, Y., and Meling, D. D. (2001). HDDI: Hierarchical Distributed Dynamic Indexing. In *Data Mining for Scientific and Engineering Applications*. Kluwer Academic Publishers. [cited at p. 33]

[Pottenger and Yang, 2001] Pottenger, W. M. and Yang, T.-H. (2001). Detecting emerging concepts in textual data mining. pages 89–105. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA. [cited at p. 33, 38]

[Quinlan, 1993] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann. [cited at p.]

[RNN-Users, 2013] RNN-Users (2013). Facebook RNN group. https://www.facebook.com/RNN.World?ref=ts. [Online; accessed 01-February-2013]. [cited at p. 4]

[Salton et al., 1982] Salton, G., Fox, E., and Wu, H. (1982). *Extended Boolean Information Retrieval*. Department of Computer Science: Technical report. Cornell University, Department of Computer Science. [cited at p. 47, 92]

[Schumm, 1991] Schumm, S. A. (1991). *To interpret the Earth: Ten ways to be wrong*. Cambridge University Press. [cited at p. 10]

[Scriven, 1991] Scriven, M. (1991). Prose and Cons about Goal-Free Evaluation. *American Journal of Evaluation*, 12(1):55–62. [cited at p. 124]

[Shin, 2009] Shin, H. S. (2009). Reflections on Northern Rock: The Bank Run That Heralded the Global Financial Crisis. *Journal of Economic Perspectives*, 23(1):101–19. [cited at p. 108]

[Streibel et al., 2013a] Streibel, O., Ahrendt, P., Gessulat, S., Lahmann, D., and Michels, M. (2013a). Trend Mining: Seminar for master students @ Freie Universität Berlin. https://sites.google.com/site/seminartrendmining/. [Online; accessed 01-July-2013]. [cited at p. , 45]

[Streibel and Alnemr, 2011] Streibel, O. and Alnemr, R. (2011). Trend-based and reputation-versed personalized news network. In *Proceedings of the 3rd SMUC'11, international workshop on Search and Mining User-generated Contents*, pages 3–10. ACM. [cited at p. , 91, 100, 102]

[Streibel and Mochol, 2010] Streibel, O. and Mochol, M. (2010). Trend ontology for knowledge-based trend mining on textual information. In *Information Technology: New Generations (ITNG), 2010 Seventh International Conference on, IEEE Computer Society Proceedings*, pages 1285–1288. IEEE Computer Society. [cited at p. 63]

[Streibel et al., 2013b] Streibel, O., Wißler, L., Tolksdorf, R., and Montesi, D. (2013b). Trend template: mining trends with a semi-formal trend model. In *Proceedings of the 3rd UDM2013 Workshop on Ubiquitous Data Mining in conjunction with the 23rd. International Joint Conference on Artificial Intelligence (IJCAI 2013)*, pages 49–53. IJCAI AAAI. [cited at p. 75]

[Surowiecki, 2004] Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Knopf Doubleday Publishing Group. [cited at p. 19]

[Swan and Allan, 1999] Swan, R. and Allan, J. (1999). Extracting significant time varying features from text. In *CIKM'99: Proceedings of the eighth International Conference on Information and Knowledge Management*, pages 38–45. ACM. [cited at p. 30, 34, 37]

[Swan and Jensen, 2000] Swan, R. and Jensen, D. (2000). Timemines: Constructing timelines with statistical models of word usage. In *KDD-2000 Workshop on Text Mining*. [cited at p. 30, 34, 37, 38]

[Tichy, 1998] Tichy, W. F. (1998). Should computer scientists experiment more? *Computer*, 31(5):32–40. [cited at p. 122]

[Tichy et al., 1995] Tichy, W. F., Lukowicz, P., Prechelt, L., and Heinz, E. A. (1995). Experimental evaluation in computer science: A quantitative study. *Journal of Systems and Software*, 28(1):9–18. [cited at p. 123]

[Twitter, 2013] Twitter (2013). "A real-time information framework". https://de.twitter.com/. [Online; accessed 01-February-2013]. [cited at p. 4]

[Vejlgaard, 2008] Vejlgaard, H. (2008). *Anatomy of A Trend*. McGraw-Hill. [cited at p. , 20]

[W3C, 2004] W3C (2004). Resource Description Framework (RDF). http://www.w3.org/RDF/. [Online; accessed 01-May-2013]. [cited at p. 48]

[W3C, 2008] W3C (2008). SPARQL Query Language for RDF. http://www.w3.org/TR/rdf-sparql-query/. [Online; accessed 01-May-2013]. [cited at p. 48]

[Wang et al., 2005] Wang, X., Smith, K., and Hyndman, R. (2005). Dimension reduction for clustering time series using global characteristics. In Sunderam, V., van Albada, G., Sloot, P., and Dongarra, J., editors, *Computational Science - ICCS 2005*, volume 3516 of *Lecture Notes in Computer Science*, pages 11–14. Springer Berlin / Heidelberg. [cited at p. 32]

[Wang et al., 2006] Wang, X., Smith, K., and Hyndman, R. (2006). Characteristic-based clustering for time series data. *Journal Data Mining and Knowledge Discovery*, 13(3):335–364. [cited at p. 22]

[Watts et al., 1997] Watts, R. J., Porter, A. L., Cunningham, S., and Zhu, D. (1997). TOAS Intelligence Mining; Analysis of Natural Language Processing and Computational Linguistics. In *PKDD '97: Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 323–334. Springer-Verlag. [cited at p. 30, 34, 36]

[Wiki-Community, 2013] Wiki-Community (2013). Wikipedia: The free encyclopedia. http://www.wikipedia.org/. [Online; accessed 01-July-2013]. [cited at p. 4]

[Wikipedia-Users, 2011a] Wikipedia-Users (2011a). Day of revolt:. http://en.wikipedia.org/wiki/Timeline_of_the_2011_Egyptian_revolution_up_to_the_resignation_of_Mubarak#25_January_.E2.80.93_Day_of_Revolt. [Online; accessed 30-March-2011]. [cited at p. 4]

[Wikipedia-Users, 2011b] Wikipedia-Users (2011b). Egyptian revolution of 2011:. http://en.wikipedia.org/wiki/Egyptian_Revolution_of_2011. [Online; accessed 30-March-2011]. [cited at p. 4]

[Willinger et al., 1996] Willinger, W., Paxson, V., and Taqqu, M. S. (1996). Self-Similarity and Heavy Tails: Structural Modeling of Network Traffic. [cited at p. 22]

[Wißler and Streibel, 2012] Wißler, L. and Streibel, O. (2012). Ontologien im trend mining. Technical Report TR-B-12-07. [cited at p. , 82, 86]

[Witten et al., 2007] Witten, I., Gori, M., and Numerico, T. (2007). *Web Dragons: Inside the Myths of Search Engine Technology*. Morgan Kaufmann a series in Multimedia and Information Systems. Morgan Kaufmann Publ. Incorporated. [cited at p. 72]

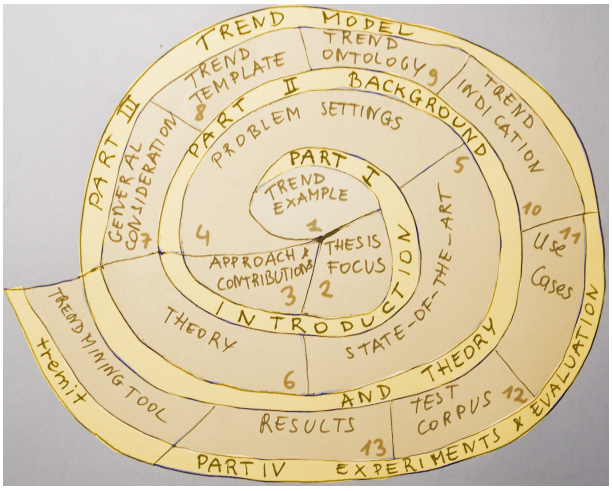
[Witten and Eibe, 2005] Witten, I. H. and Eibe, F. (2005). *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers Inc. [cited at p. , 26, 126]

[Wolf et al., 1985] Wolf, A., Swift, J. B., Swinney, H. L., and Vastano, J. A. (1985). Determining Lyapunov Exponents from a Time Series. *Physica*, pages 285–317. [cited at p. 22]

[Wüthrich et al., 1998] Wüthrich, B., Permunetilleke, D., Leung, S., Cho, V., Zhang, J., and Lam, W. (1998). Daily prediction of major stock indices from textual www data. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining - KDD-98*, pages 364–368. [cited at p. 37, 38]

[Yue et al., 2002] Yue, S., Pilon, P., and Cavadias, G. (2002). Power of the Mann-Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series. *Journal of Hydrology*, 259(1):254–271(18). [cited at p. 150]

[Zeng and Zhang, 2009] Zeng, J. and Zhang, S. (2009). Incorporating topic transition in topic detection and tracking algorithms. *Expert Systems with Applications*, 36(1):227 – 232. [cited at p. 36, 40]



APPENDIX A

tremit: the Trend Mining Tool

The idea of a trend mining tool emerged during this thesis. The main reason for creating our tool is the need for a tool with which new algorithms can be tested and in which state-of-the-art algorithms can be integrated easily. The trend mining tool – tremit – presented here is aimed to be a sandbox for every researcher, data scientist and developer interested in trend mining on web documents. It is under ongoing development and it implements the approaches presented in this thesis offering a simple GUI as well as easily extendible interfaces.

Tool description

This is the short version of the tremit description[1].

Goal

The primary goal of our trend mining tool is to have a flexible and extendible tool for mining trends primarily in a web document corpus. The secondary goal is to develop a general trend mining tool with several different built-in functionalities and tests with use cases on different data sets. Figure A.1 shows the GUI of tremit.

Functionality

The main functional requirements on the trend mining tool are listed as follows:

[1]The detailed description including explanation of the demo-software `https://sites.google.com/site/tremitool/` to appear

Figure A.1: tremit - GUI

1. Processing of data: this function allows for the different processing techniques on the web documents, including the common NLP analysis and simple text mining.

2. Calculation of a trend: under this function we summarize topic modeling and the documents' weighting functions that allow for calculating emerging topic areas.

3. Linking of knowledge: knowledge linking is the functionality that allows for the extending given parts of documents into graphs, provided the ontology concepts.

4. Generating of the trend model: this function allows for the generating of a model, such as a trend ontology or a clustering method, which can be applied on further data sets in order to calculate a trend.

Moreover, the following functionalities are included:

1. selecting of trend features

2. learning trend features

3. creating trend descriptions

4. extracting topics

5. deriving topic clusters

The tool contains the following help functions:

1. parsing

2. storage

3. calculation of best time window

4. visualization
5. read/write ontology
6. extract knowledge

Algorithms

The following algorithms are implemented in the tremit:

- topic models based on LDA (adjusted implementation)
- k-means clustering combined with topic models (adjusted implementation)
- trend ontology
- trend indication method

Data

As a test corpus, a chosen part of the corpus described in Chapter 12 is included in the tool.

Language

The demo tool has been developed with a GUI and results' representation primarily in German (test web documents are in German). An English version is under development.

Architecture

The general architecture shows the main functionality of tremit. The functionality is covered by several components that are described in the following.

Components

Based on the general architecture of tremit, we describe the components of the tool corresponding to the respective functionality as presented above.

1. **Data processing:** Data processing consists of the following components:

- parser component
- db component

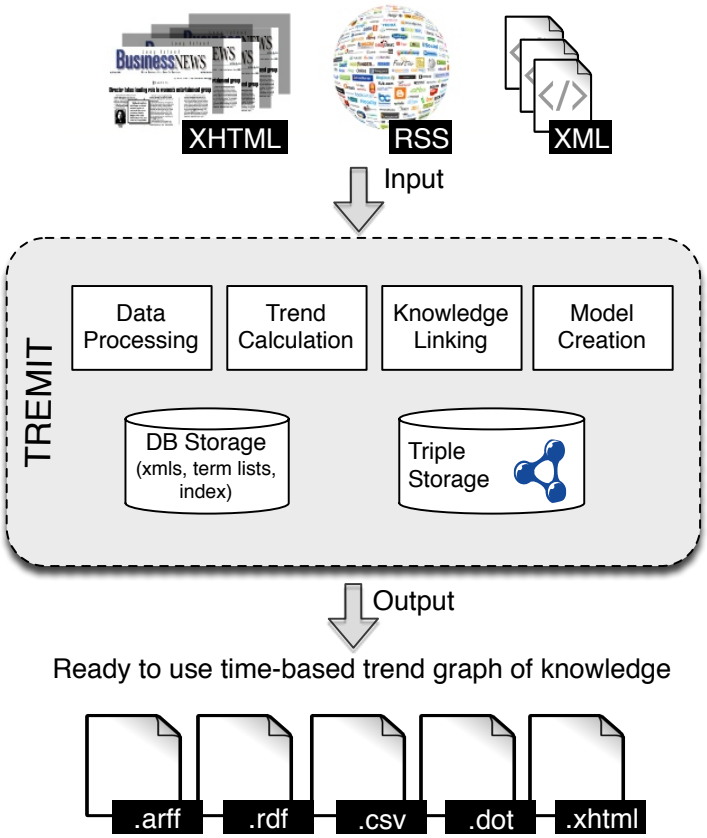


Figure A.2: Architecture

It is merged in the packages: *tremit.dataprocessing.parser* and *tremit.processing.db*. The main classes of data processing are visualized in Figure A.3.

2. Trend calculation: Trend calculation unites all functions needed for trend estimation and consists of the following components:

- indication calculator
- features extractor

Packages: *tremit.trendcalculation.indication* and *tremit.trendcalculation.features*. The main classes of trend calculation are visualized in Figure A.4.

3. Knowledge linking:

- trend ontology reader
- trend ontology writer

- trend ontology learner
- knowledge extractor
- knowledge connector
- help: ontology analyzer

Packages: *tremit.linking.knowledge* and *tremit.linking.ontology*. The main classes of trend calculation are visualized in Figure A.5.

4. Model generator:

- topic models converter
- k-means clustering converter
- help: topic models analyzer
- help: cluster analyzer

Packages: *tremit.modelgenerator.cluster tremit.modelgenerator.topic*. The main classes of trend calculation are visualized in Figure A.6.
Additionally a visualization package *tremit.viz* contains all extraction classes and methods for graph creation. Furthermore a helper package *tremit.analyzer.** contains the helpful components for analysis of ontology and analysis of generated models. The package *tremit.user.** provides interfaces for functionalities that are dependent on users.

Operation mode

Tremit is a a tool with three modes of operation:

- command line
- GUI
- Java API (later)

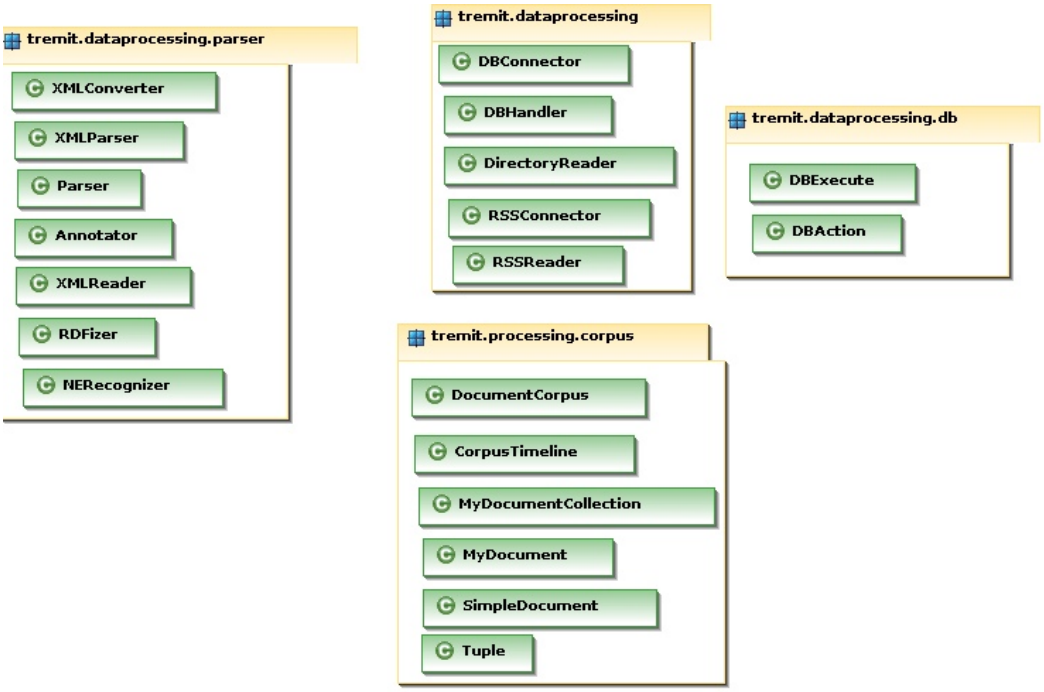


Figure A.3: Main classes in tremit.dataprocessing.*

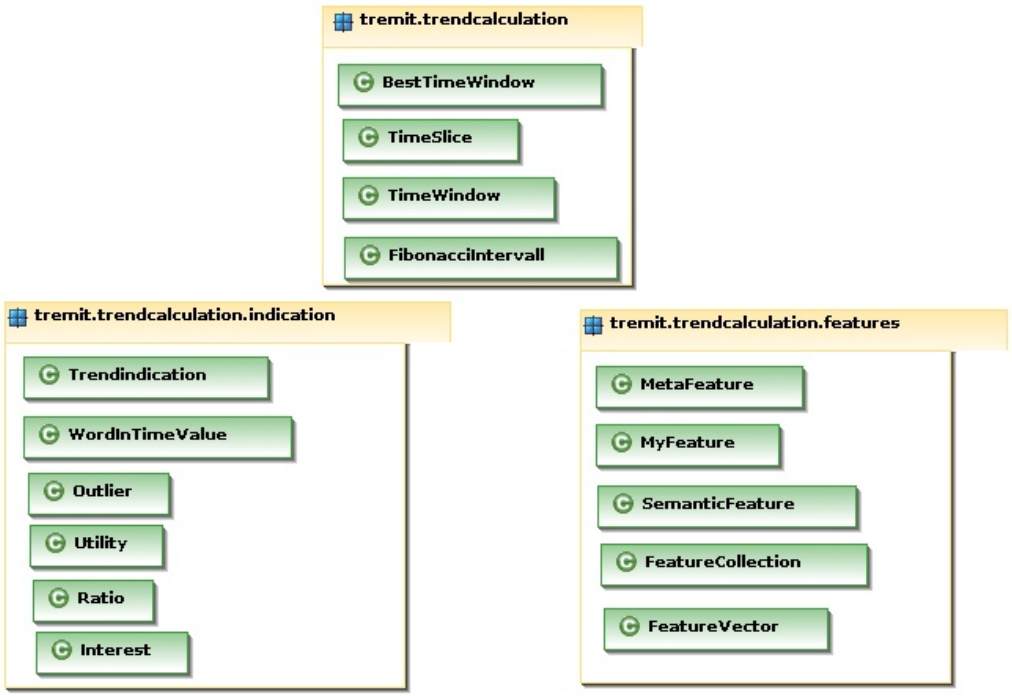


Figure A.4: Main classes in tremit.trendcalculation.*

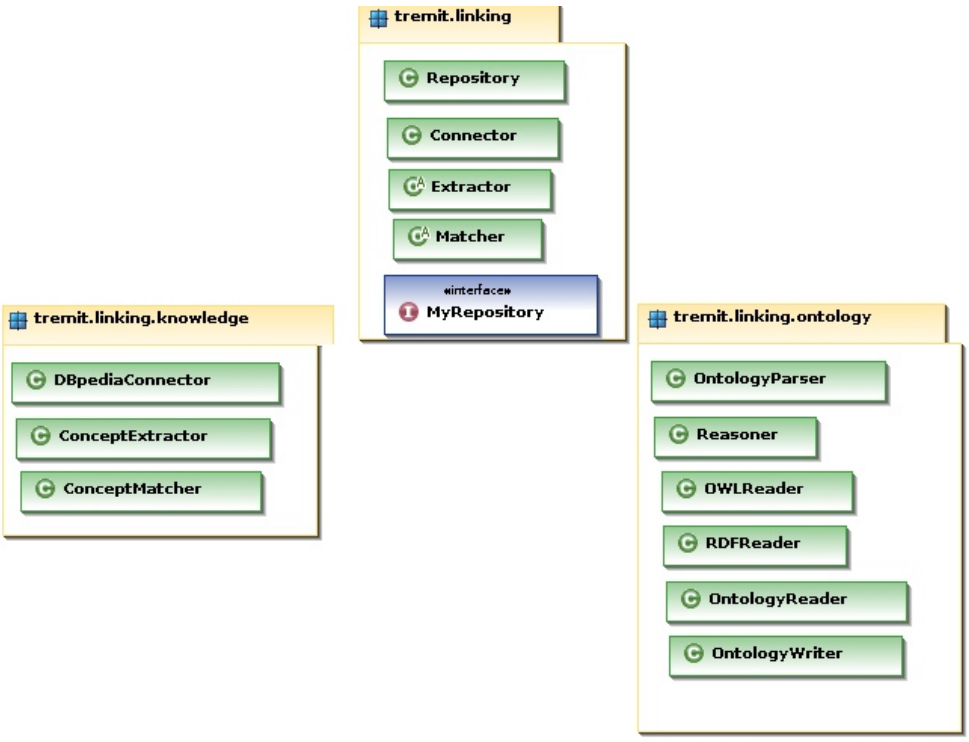


Figure A.5: Main classes in tremit.linking.*

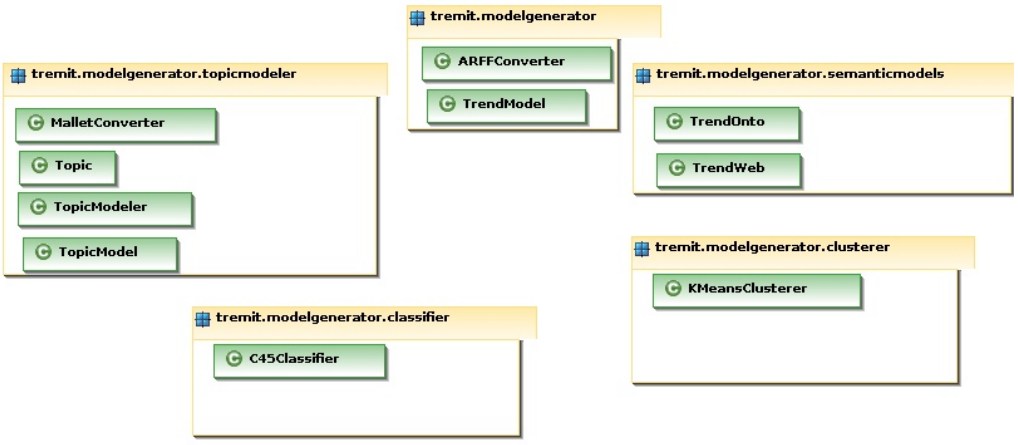


Figure A.6: Main classes in tremit.modelgenerator.*

APPENDIX B

Zusammenfassung und Kurzlebenslauf

(§7 Abs. 6 – Promotionsordnung des Fachbereichs Mathematik und Informatik der Freien Universität Berlin, Stand 2007)

Zusammenfassung

Ein *Trend* im Kontext des Information Retrievals (IR) ist *ein Themengebiet, das über einen Zeitraum an Nutzwert und Interesse gewinnt*, wie z. B. das allgemeine Thema *Finanzkrise* im Zeitraum 2008-2012 oder *Arabischer Frühling* im Zeitraum 2010-2011.
Es gibt Verfahren, verankert in Bereichen des Data Minings, Text Minings und des Maschinellen Lernens, die zur Lösung des Problems der Trenderkennung in Texten herangezogen werden. Zu den oft verwendeten gehören die probabilistischen Topic Models sowie verschiedene Clusteringverfahren.
Die Schwachstellen der existierenden Forschung über automatische Trenderkennung in Texten liegen in:

1. inkonsistenten Definitionen des *Trends*
2. fehlendem wissenschaftlichen Ansatz des *Trend Mining*
3. fehlendem Bezug zum expliziten Wissen und damit schlechter Interpretierbarkeit der Ergebnisse

Der wissenschaftliche Beitrag dieser Arbeit besteht in dem Vorschlag, die Forschung zur automatischen Trenderkennung aus der Sicht des *Trend Mining* zu betrachten, dessen Definition in dieser Arbeit vorgeschlagen wird.

Als Lösung für das Problem der schlechten Interpretierbarkeit der Ergebnisse von gängigen Trenderkennungsalgorithmen wird *trend template* vorgeschlagen, das ein wissensbasierter Ansatz für trend mining ist. Ausgehend von diesem trend template werden zwei Implementierungsrichtungen gezeigt: die Trendontologie und das Trend-Indication-Verfahren.

Die Trendontologie funktioniert nach dem Prinzip eines A-priori-Modells und ermöglicht die Entdeckung einer Trendstruktur in dem Webdokumentenkorpus. Tests mit diesem Verfahren auf dem Testkorpus zeigen, dass Trenderkennung mit einem A-priori-Modell unter Einbezug von explizitem Wissen, zu qualitativ besseren Ergebnissen, vor allem in Hinsicht auf die Interpretierbarkeit, führt.

Das Trend-Indication-Verfahren baut auf den zeitbasierten Gewichtungsfunktionen auf und konzentriert sich auf die Selektion der Trend Features aus den Webdokumenten. Mithilfe dieses Verfahren wird die Dimension der zu untersuchenden Daten im Hinblick auf die Trenderkennung sinnvoll reduziert und somit nur die zeitrelevante Information aus den Texten für weitere Analysen bereitgestellt. Die Tests mit diesem Verfahren zeigen, dass zeitrelevante Trendbegriffe alleine durch geeignete Gewichtungsfunktionen gut aufgedeckt werden.

Beide Methoden werden in dem *tremit* (TREnd MIning Tool), das für diese Arbeit entwickelte Testtool, implementiert und auf dem Testkorpus getestet. Der Testkorpus besteht aus 35.635 Wirtschaftsnachrichten und 4.696 DAX-Berichten des deutschsprachigen Webs aus dem Zeitraum September 2007 bis April 2008. Die Ergebnisse werden mit den Ergebnissen der gängigen Verfahren – LDA-basiertem Topic Model und k-means Clustering – auf dem gleichen Korpus verglichen und im Experimentierteil der Arbeit diskutiert und evaluiert.

Kurzlebenslauf

Der Lebenslauf ist in der Online-Version aus Datenschutzgründen nicht enthalten.

Erklärung

Hiermit erkläre ich, dass alle Hilfsmittel und Hilfen angegeben sind und versichere, auf dieser Grundlage die vorliegende Arbeit selbstständig verfasst zu haben. Diese Arbeit wurde bisher noch nicht in einem früheren Promotionsverfahren eingereicht.

(§7 Abs. 4 – Promotionsordnung des Fachbereichs Mathematik und Informatik der Freien Universität Berlin, Stand 2007)

Olga Streibel

Berlin, 23.07.2013