Mining Trends in Texts on the Web

Olga Streibel
2. year PhD, 1.Supervisor: Prof. Dr.-Ing. Robert Tolksdorf

Networked Information Systems, Free University Berlin,
Königin-Luise-Str.24-26 , 14195 Berlin, Germany,
streibel@inf.fu-berlin.de

Abstract. From online news and blog articles, a human can often deduce information and knowledge needed for the prediction of market movements or sociological trends. However, this recognition and comprehension process is very complex and requires experience as well as some context knowledge about the domain in which trends are to detect. In order to support human experts in trend analysis, I propose an automatic trend mining method based on knowledge integrating learning approach.

Key words: trend mining, machine learning, knowledge acquisition, knowledge integration, semantic learning, tagging, folksonomy

1 Problem statement

"Many people have been led to believe that trends are about intuition. This is because the majority of the people who work with trends find it difficult to explain why something will happen the way they say it will. The explanation often boils down to "because I think so." Some people do seem to be able to predict what will happen based on their own intuition. Unfortunately, there are too many cases in which people's intuition has obviously been mistaken (...)" [26]
Detecting trends from the sociologists' point of view is an analytical method for observing changes in people's behavior over time with regard to "six attitudes towards trends" defined as *trendsetters, trend followers, early mainstreamers, mainstreamers, late mainstreamers and conservatives* (s. "The Diamond-Shaped Trend Model" in [26]). Consequently, trends are certain patterns of people's behavior and lifestyle that evolved over a focused time interval and the word *trend* refers to a process of change.
Detecting trends from the statistical point of view is based on trend analysis of time-series data regarding two goals of analysis: "*modeling time series* (i.e. to gain insight into the mechanisms or underlying forces that generate the time series) and *forecasting time series* (i.e., to predict the future values of the time-series variables)" (p. 490, [10]). In this terms, *trend* refers to the general direction in which a time-series graph, based on numeric data, is moving over a focused interval of time.
Detecting trends from text collections refers to the detection of emerging topics in texts. In terms of textual data mining a *trend* in texts is defined as "a

topic area that is growing in interest and utility over time" [13] whereas topic in terms of Topic Detection and Tracking (TDT)[3] research is "defined to be a set of news stories that are strongly related by some seminal real world event". All of these points of view on trend detection show the different dimensions of trend analysis research. However, they have one thing in common: observing patterns of changes that are based on certain variables (i.e., people, numbers, words) and lead to a general change- the emerging trend- in the system which is depending on these variables.
As already defined in my trend ontology approach[23], this research uses *trend mining* as a general term describing trend detection, trend recognition and trending analysis. It can refer either to the detection of emerging topic areas from text analysis or to the detection of trends based on numeric data analysis as in the case of stock values. However, this work focuses only on textual data available on the Web, i.e. online news and blogs, and on learning this data under inclusion of related background knowledge in order to capture and explain trends. In general, I refer to the "emerging topic areas" (see also Section 4) while using the term *trend* in texts whereas the objective of mining trend is "to provide an alert that new developments are happening in a specific area of interest in an automated way"[13].
Interesting approaches have been developed in the field of trend mining on texts (s. following Section) but they are still lacking the integration of expert knowledge in the process of trend recognition. Such knowledge is crucial for the proper trend mining and the lack of methods that integrate expert knowledge is a research gap. This thesis aims at closing this gap. It deals with the trend detection task as with a complex learning task based on learning and recognizing of complex relations and dependencies in given domain regarding the time dimension. I focus on the learning method able to integrate expert knowledge in order to automatically recognize trends in text collections.
Considering that *"In general, trending analysis of textual data can be performed in any domain that involves written records of human endeavors whether scientific or artistic in nature."*[20] trend mining based on texts is useful for many application domains, i.e. medical diagnosis, opinion mining, market monitoring, stock market analysis, etc., and, regarding the increasing information availability on the Internet with its need for intelligent data analysis, it is becoming more and more important research topic in the recent years. Besides contribution to the Trend Mining research, this thesis can have important impact for Machine Learning, and also for the Semantic Web.

2 Main questions of the thesis

Two main research questions are important for this thesis: 1) How to change existing machine learning approaches for trend mining into knowledge integrating learning approaches with regard to the development of the Semantic Web? 2) How to acquire and formalize trend knowledge?
Main research projects in the field of trend mining are described in Topic Detec-

tion and Tracking (TDT) research[3] and in Emergent Trend Detection (ETD)[5]. Regarding relevant work for this thesis, in first I concentrate on the research done in the field of trend mining with a focus on the machine learning algorithms since they seem to be crucial in the automatic trend mining. One of the researches, EAnalyst system described in [15], proved that determination and early detection of emerging trends can be retrieved from numeric data as well as from texts. EAnalyst 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, redescribes time series data into "high-level features", called trends, and aligns each trend with time-stamped news stories. Such news stories serve as training set for learning the *language model* which determines the statistics of word usage patterns in the stories. This language model, learnt for every trend type, helps to monitor a stream of new incoming news stories. The model processes new news stories due to the learnt hypothesis. Authors define here the task of trend detection as a special case of the *Activity Monitoring* as introduced by [7]. This research allows for the general precondition in my thesis: it is possible to automatically recognize trends by analyzing texts. Different from EAnalyst, I do not elaborate on text stream monitoring but focus more on the recognition and comprehension process for trend mining.
Emergent Trend Detection (ETD) systems that concern with detection of trends presented in [13] have been characterized based on the following aspects: *input data and attributes*, *learning algorithms* and *visualization*, that are important for creating a trend analysis system. The most relevant comparison perspective for our work are the *learning algorithms*. According to the system description in [13] and regarding the prototypes [27][17][6], following learning algorithms have been proven useful for trend mining:

- combined "hypothesis testing"-based methods (Time Mines[24])
- single-pass clustering (New Event Detection[4])
- sequential pattern matching and shape query processing (Patent Miner[16][1])
- feed-forward, backpropagation NN, c4.5 and SVM (Hierarchical Distributed Dynamic Indexing[20], Wüthrich[27])
- k-NN classifier (Wüthrich[27])
- regression analysis (Wühtrich[27])

Besides, there are many research works related to trend mining, i.e, trend detection based on a fuzzy temporal profile model[8], modeling bursty streams using infinite-state automaton[12], finite mixture model for tracking dynamics of topic trends[18], and clustering approaches [14][3]
Concerning both, the trend mining based on texts and enhanced text analysis, there are many related projects on the Internet, scientific and commercial, as well as services that are to some extend relevant for this work: GoogleTrends [1], BlogPulse[2], OpenCalais[3] Two interesting research project GIDA (Generic

[1] `http://www.google.de/trends`
[2] `http://www.blogpulse.com/`
[3] `http://www.opencalais.com/`

Information-based Decision Assistant) [9][2] and its follower, TREMA (Trend Mining, Fusion and Analysis of multimodal Data) [19], that concentrated on the fusion of multimodal market data in order to mine trends in financial markets (GIDA, TREMA) and in market research (TREMA) are relevant for this thesis. Several projects that concern themselves with lightweight ontologies and extended vocabularies are relevant for the trend knowledge representation part of this thesis, in particular: ConceptNet[4] and OpenMind[5] of MIT, MoaT[6], WordNet[7], SentiWordNet[8], Wortschatz Uni Leipzig[9], DWDS[10], SKOS[11], SCOT[12]
Regarding relevant work outlined above and according to the two research questions, this research focuses on the development of a semantic learning approach for the automatic trend mining in texts on the Web. It also proposes the use of trend ontology and elaborates on the extreme tagging approach[25] for knowledge acquisition in the trend mining task. However, the main goal of this work is not to predict stock prices for the stock markets based on news analysis nor to create an artificial trader for market trading based on text analysis. This research is neither about a general trend analysis system and it is not studying the influence of Web news on emerging trends (it doesn't take into account the distinction into trend creator news, trend follower news and mainstream news). General assumptions for this thesis are: context is crucial for successful trend mining, collective associations like user tags from folksonomies enable the creation of context knowledge, statistical learning can be enhanced with background knowledge using knowledge representation approach from Semantic Web.

3 General approach

This thesis is anchored in Information System research and Design Science paradigm[22][11] is the methodology that provides the scientific framework for my research. Two main research issues are in focus of my thesis: knowledge-integrating learning approach for trend mining based on Machine Learning and the representation of trend knowledge based on Semantic Web approach. Concentrating on them, I create my artefact (in terms of Design Science), test and evaluate my trend mining approach.
So far, first of all I did an extensive literature review comparing following general aspects of related projects on trend mining: trend definitions, general trend analysis approaches, applied machine learning methods and document corpora. Regarding this issues I elaborated on a general definition of trends in text (this

[4] http://conceptnet.media.mit.edu/
[5] http://commons.media.mit.edu/en/
[6] http://moat-project.org/
[7] http://wordnet.princeton.edu/
[8] http://sentiwordnet.isti.cnr.it/
[9] http://wortschatz.uni-leipzig.de/
[10] http://www.dwds.de/
[11] http://www.w3.org/2004/02/skos/
[12] http://scot-project.org/

gives the main setting for defining the learn problem in the next steps). Furthermore, I implemented a static storage, parsing and partially preprocessing of document corpus that consists of about 200000 business news in German language in the time interval 2006-2007. I also elaborated on the trend ontology approach [23] and on the knowledge acquisition approach using tag tagging[25]. In the next steps, I have to concentrate on the general description of the learning problem in case of mining trends in texts (what kind of feedback is available, what kind of features should be learnt, how to extract trend labels, what is the feature space and how good separable are different classes, how can the features be extended into semantic features, etc.). While defining the learning problem, I also have to consider the representation of the learning data and the representation of the background knowledge.
In general, this thesis elaborates on the idea of semantic learning which is the adoption of inductive learning approach from the Machine Learning with the knowledge representation approach from Semantic Web. The outcome of this thesis is a knowledge integrating method for mining trends in texts which aims at improving the quality of trend mining methods and brings the additional value to the existing methods- the trend explanation.

4 Proposed solution

At this stage of my work, the solution proposed starts with few important definitions: time window, time slice, burstiness, interestingness, utility and trend indication. Based on them, an exact description of what are trends in text is possible:
Definition 1: Time window
t_{window} is a time interval in which trends can occur. Furthermore, it can be described as an ordered set of subintervals.
t_{slice}[13]is a subinterval of time window. If its starting point lies at t_0 the end point has to lie at $t_k < t_n$

$$t_{window} = [t_0 ... t_n] \wedge t_{slice_k} = [t_0 ... t_k]$$

$$t_{window} := \{t_{slice_k}, ..., t_{slice_n}\} \wedge |t_{slice_k}| = |t_{slice_n}| \wedge k, n \in \mathbb{N} \wedge k < n \qquad (1)$$

Time slices have the same length.
Definition 2: Burstiness
In order to distinguish words in the documents of given time slice from the all documents in time window, TFIDF (term frequency inverse document frequency)[21] function is adapted. The function result for each word says how important is a given word in a given period of time. This is the function to discover the burstiness of words: if there is a word in a given time slice which appears only in the documents of this time slice and not in the whole window

[13] this is needed since only long-term trends are relevant for this thesis

(backwards) it could be the so called entry point of a trend.

$$burst(w)_{t_{window}} := TF_{(w,|D|_{t_{slice}})} * IDF_{(w,|D|_{t_{window}})} \tag{2}$$
$$IDF_{(w,|D|_{t_{window}})} := \log \frac{|D|_{t_{window}}}{DF(w)_{t_{window}}}$$

whereas $|D|$ is the total number of documents. If the word continues to appear in next time slices, and becomes interesting, the word can become trendindicating. Based on the time component as in Def. 1, trendindication is defined by interestingness and utility as follows:

Definition 3: Interestingness

Interestingness is defined by the frequency of word w in the time window. This can be expressed for a time slice by the sum of the term frequency of word w in all the documents of given time slice divided by the number of documents in this time slice (scaled by binary logarithm).

$$interest(w)_{t_{slice}} = f(w)_{t_{slice}} := \log \frac{\sum TF_{(w,D_{t_{slice}})}}{|D|_{t_{slice}}} \tag{3}$$

For the trendindication it is important to know if the interestingness of a word is rising over time window. As given by formula 1 in Def.1, we define as follows for given time window:

$$interest(w)_{t_{window}} := \{f(w)_{t_{slice}k}, f(w)_{t_{slice}k+1}, \ldots, f(w)_{t_{slice}n}\} \tag{4}$$

expresses increasing interestingness if[14]:

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

Definition 4: Utility

Utility expresses how popular do users find a given word w in the given time window. I propose to retrieve it by analysing collaborative tagging systems (CTB), i.e. delicious, and estimating the popularity of given word as a tag in the same time window as for the trend estimation. The popularity can be simple described by the number of resources in CTB that in given time window have been tagged with the word w divided by the number of all resources tagged in this time window:

$$util(w)_{t_{window}} := \log \frac{|R|_{(tag=w)t_{window}}}{|R|_{(tag)t_{window}}} \tag{5}$$

Definition 5: Trend indication

$$trendind(w)t_{window} = \frac{burst(w)t_{slice_k} + interest(w)t_{slice} * util(w)t_{window}}{ratio(t_{window})} \tag{6}$$

[14] this thesis focuses only on upcoming trends and ignores falling trends

whereas:

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

is the number of time slices.
The definitions above allow for a general description of emerging topics in given time window: emerging topics are in the simplest case the intersection of the trend indicating words (set of all words that at some point in the time window start to have bursty behavior and appear frequently enough to be discovered and rare enough to be important in given time window) with the set of words used as tags in a CTB in this time window. Furthermore, the trend indication allows for automatic labeling of the document corpus and dividing it in trend indicating and neutral documents (regarding the time slices in which the documents appear).
However, this is the statistical part of the approach and it focuses only on simple words. At this stage of the thesis tests have to be done in order to prove it useful. Furthermore, I have to elaborate on the inclusion of the background knowledge into the labeling either by applying my trend ontology[23] approach or tag tagging approach[25] in order to extend the features into the "real" semantic concepts, which I call statements, and at the same time to reduce the dimension of the feature space.

As for learning approach I propose to adapt the Bayes learning[15]. The Bayes theorem could be in this case explain in very general way as:

$$P(T|S) = \frac{P(S|T)P(T)}{P(S)} \tag{7}$$

$P(T|S)$ is the a posteriori probability of T conditioned on S whereas T is the hypothesis and S a statement. In case of mining trends T says that there is an indication for a trend and $P(T|S)$ reflects the probability that the given statement S will indicate a trend (or that the given statement S is built on trendindicating concepts and therefore indicates a trend). $P(T)$ and $P(S)$ are the a priori probabilities: over T (any given statement causes trend) and over S (any statement from the training set is trend-indicating), $P(S|T)$ can be estimated from the given data.

At this stage of my work, I start the tests for trend feature extraction and continue to elaborate on my solution for integration of background knowledge as well as for proper definition of the learning method.

5 Evaluation

The evaluation of my approach is primary based on the evaluation of the model performance which can be conducted using crossvalidation and measured in general by the recall and precision values. For the crossvalidation, the document corpus is divided in i folders and the validation process is repeated i times whereas

[15] However, also decision trees (good for vizualization and comprehending of the model) and support vector machines (most reliable classification method) have to be considered

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/recall value are:

$$recall = \frac{|D|_{trendindicating-and-retrieved}}{|D|_{trendindicating}} \tag{8}$$

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

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

$$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{|a_1 - \overline{a}| + \ldots + |a_n - \overline{a}|} \tag{9}$$

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. The final evaluation depends on the final learning model and should also take into account the knowledge integration part (this could be done i.e. in case of decision trees by additional measure of changes in information gain values).

6 Future Work

Many research issues are relevant to this thesis. From the information retrieval point of view one of them is for example the research on graph-based representation model for documents and semantic indexing of the document collections. In this stage of the work it is too early to expand the remaining issues.

Acknowledgments This work has been partially supported by the InnoProfile-Corporate Semantic Web project funded by the German Federal Ministry of Education and Research (BMBF) and the BMBF Innovation Initiative for the New German Länder - Entrepreneurial Regions. The author wants to thank Prof. Robert Tolksdorf and Prof. Abraham Bernstein for their helpful comments on the content of this thesis.

References

1. Rakesh Agrawal, Edward L. Wimmers, and Mohamed Zait. Querying shapes of histories. pages 502–514, 1995.
2. Khurshid Ahmad. Events and the causes of events. In Lee Gillam, editor, *Proceedings of the Workshop on Making Money in the Financial Services Industry, at the 6th International Conference on Terminology and Knowledge Engineering*, 2002.
3. James Allan. *Topic Detection and Tracking: Event-based Information Organization*. Kluwer Academic Publishers, 2002.
4. James Allan, Ron Papka, and Victor Lavrenko. 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, New York, NY, USA, 1998. ACM.
5. Michael Berry. *Survey of Text Mining: Clustering, Classification, and Retrieval*. Springer Science+Business Media, Inc, year = 2004.
6. Raymond K. Wong Desh Peramunetilleke. Currency exchange rate forecasting from news headlines. In *Proceedings 13th Australasian Database Conference*, pages 131–139, 2002.
7. Tom Fawcett and Foster Provost. Activity monitoring: Noticing interesting changes in behavior. In *In Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, pages 53–62, 1999.
8. Paulo Félix, Santiago Fraga, Roque Marín, and Senén Barro. Trend detection based on a fuzzy temporal profile model. *AI in Engineering*, 13(4):341–349, 1999.
9. L. Gillam, K. Ahmad, S. Ahmad, M. Casey, D. Cheng, T. Taskaya, P.C.F. Oliveira, and P Manomaisupat. Economic news and stock market correlation: A study of the uk market, 2002.
10. J. Han and M. Kamber. *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers Inc, 2006.
11. A. R. Hevner, S. T. March, J. Park, and S. Ram. Design science in information systems research. *MIS Quarterly*, 28(1):75–106, 2004.
12. Jon Kleinberg. Bursty and hierarchical structure in streams. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 91–101, 2002.
13. April Kontostathis, Leon Galitsky, William M. Pottenger, Soma Roy, and Daniel J. Phelps. *A Survey of Emerging Trend Detection in Textual Data Mining*. Springer-Verlag, 2003.
14. April Kontostathis, Lars E. Holzman, and William M. Pottenger. Use of term clusters for emerging trend detection. Technical report, 2004.
15. Victor Lavrenko, Matt Schmill, Dawn Lawrie, Paul Ogilvie, David Jensen, and James Allan. Mining of concurrent text and time series. In *In proceedings of the 6 th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining Workshop on Text Mining*, pages 37–44, 2000.
16. Brian Lent, Rakesh Agrawal, and Ramakrishnan Srikant. Discovering trends in text databases. pages 227–230. AAAI Press, 1997.
17. Marc-Andre Mittermayer and Gerhard F. Knolmayer. Newscats: A news categorization and trading system. *Data Mining, IEEE International Conference on*, 0:1002–1007, 2006.
18. Satoshi Morinaga and Kenji Yamanishi. 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, Seattle, Washington, USA, August 22-25, 2004*, pages 811–816, 2004.

19. Streibel Olga. Xml-clearinghouse report 17: Xml-technologies and semantic web for trend mining in business applications. Technical report, Freie Universitt Berlin, XML-Clearinghouse Project, 2007.
20. William M. Pottenger and Ting-Hao Yang. Detecting emerging concepts in textual data mining. pages 89–105, 2001.
21. Gerard Salton. *Automatic Text Processing*. Addison-Wesley, 1989.
22. Herbert A. Simon. *The sciences of the artificial (3rd ed.)*. MIT Press, Cambridge, MA, USA, 1996.
23. Olga Streibel and Malgorzata Mochol. Trend ontology for knowledge-based trend mining in textual information. In *7th International Conference on Internet Technology: New Generations*, pages 1285–1288, 2010.
24. Russel Swan and David Jensen. Timemines: Constructing timelines with statistical models of word usage. In *KDD-2000 Workshop on Text Mining*.
25. Vlad Tanasescu and Olga Streibel. Extreme tagging: Emergent semantics through the tagging of tags. In *ESOE*, pages 84–94, 2007.
26. Henrik Vejlgaard. *Anatomy of A Trend*. McGraw-Hill, 2008.
27. B. Wüthrich, D. Permunetilleke, S. Leung, V. Cho, J. Zhang, and W. Lam. 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, 1998.