Trend Mining with Semantic-Based Learning

Olga Streibel

Networked Information Systems, Free University Berlin,
Königin-Luise-Str.24-26 , 14195 Berlin, Germany
streibel@inf.fu-berlin.de
http://www.ag-nbi.de

Abstract. Mining trends by analyzing text streams could enhance the standard trend analysis based on numeric data. The use of qualitative information in the process of trend recognition, in addition to that of quantitative data, requires new analysis techniques. Since Semantic Web enables the appropriate and advantageous formalization of knowledge, we propose to include formalized expert knowledge in the process of trend recognition. In this preliminary work, we introduce our approach based on Semantic Web technologies combined with Data Mining methods for mining trends in a given domain.

Key words: trend mining, trend recognition, semantic technologies, pattern recognition, trend patterns, learning methods, trend pattern ontology

1 Introduction

"Stock market news has gone from hard to find (in the 1970s and early 1980s), then easy to find (in the late 1980s), then hard to get away from."[1]
A huge amount of textual information like business news is freely available on the Internet[2]. This abundance of information makes the access of new information far easier, as is also true of previously hidden knowledge. On the other hand, in order to retrieve required information and discover the potential knowledge, we need to utilize appropriate search and analysis techniques. Regarding business news and the stock market, a "human" specialist can deduce information and knowledge she needs for the prediction of market movements. However, this recognition and comprehension process is very complex and requires experience as well as the initial context knowledge.
In our work, we concentrate on the *trend mining* process based on numeric data and on textual information. Research projects like GIDA and TREMA have shown that there is a huge demand for the research on and development of

[1] Peter Lynch,2000 "One Up On Wall Street: How To Use What You Already Know To Make Money In The Market"

[2] i.e. http://news.bbc.co.uk/2/hi/business, http://www.tagesschau.de/wirtschaft/index.html, http://faz.net

useful trend mining methods that are able to include analyses of textual information in the process of trend recognition. In our work, we define repositories consisting of quantitative data and qualitative data as simple hybrid information systems. Regarding specific application fields, i.e. financial markets, the qualitative data is represented by financial news whereas the quantitative data means the numeric values of different trade instruments. Consequently, we aim to use text corpus consisting of financial news in German language[3] and correlate this corpus with the trading values of a chosen financial instrument. In particular, we concentrate on the analysis of the business news filtered over a period of 12 months due to the trend segments deduced from the market values of a trading instrument. The focus of our research is on developing a solution relevant to the trend mining problem in simple hybrid information systems which combines a Data Mining approach and adequate Semantic Web technologies. There are many other examples of simple hybrid information systems in application areas like weather forecasting, traffic analysis, customer opinion mining, etc. We will work on a solution that will be applicable in those different systems.
In the following, we outline briefly the idea of our novel approach for trend mining. Section 2 gives an insight into research relevant to our work. In section 3, we specify our definition of a "trend" and outline the issues of our research. Describing briefly the different methods from Computer Linguistic which can be partially applied to the trend mining difficulty, we introduce *Extreme Tagging System* (ETS) in 3.2. We close the section with a short paragraph about learning methods that we aim to apply in the future.

2 Related Work

The research project GIDA[4][6][1] and its follower, TREMA[5], concentrated on the fusion of multimodal market data in order to mine trends on financial markets (GIDA, TREMA) and in market research (TREMA). These projects provide us with our research direction. Since we aim to focus only on a fraction of the whole trend mining process, in particular, on the search for the trend indicating language patterns in news, we are not going to concern ourselfs with the conception of a complex trend mining framework as the project TREMA does. Similar to TREMA, we are using the Semantic Web technologies in order to support the textual trend recognition. The difference lies in our idea of applying an ETS, as described in section 3.2, instead of applying classic and hierarchical ontologies.
In [3] the concept of velocity density estimation is discussed for the trend mining in supermarket customers' data.This work "provides the user generic tool to understand, visualize and diagnose the summary changes in data characteristics". The aspects of dynamics and evolving data included in this research, could also

[3] The corpus is available due to the cooperation with the German company, neofonie GmbH
[4] Description online: www.computing.surrey.ac.uk/ai/gida
[5] Project website: www.trema-projekt.de

be important for our work. The authors of [16] introduce a simple and interesting knowledge-based approach for the kidney function monitoring in medical diagnosis systems. In particular, the trends appear in the form of trend reports which are counted on the numeric data and explained using a knowledge-base. The use of a semantic knowledge-base will also be a part of our work. We are going to use the knowledge base not only to explain the emerging trends but also to learn from them. Trends based on keyword search statistics are well visualized by the Google-Trends [24] feature. Here, the trend mining of searches actually shows anomalies appearing in the historic patterns of Google search on the Web. Search for certain text patterns in the text corpus is also a part of our work. The difference is that we aim to search for trend indicating keywords that have been learned from historic data using semantic, not only statistic methods. Another interesting tool is the BlogPulse [25] that identifies topics and subjects that people are talking about in their blogs. BlogPulse shows the complex trend concept. A trend is a phenomenon that consists of trend setters (blogs' authors), detected topics, "buzz" words, etc. In our work, we are assuming a simplified, data and text oriented, trend definition that can be treated as a fraction of the complex trend mining process.
As last, the work described in [10] could be very useful for us. In particular, the definitions of *theme*, *theme life cycle*, and *theme snapshot* could be important for our approach.

3 Mining Trends

In order to analyze trends, we have to define what is a *trend*. Since we aim firstly to originate our trend recognition process in the numeric data, we will treat the given text stream in a similar way as we might a data stream. With regard to the trend analysis based on time series, the analysis process consists of four major *components* or *movements* for characterizing time-series data [8]. We refer to the *long-term movements* that can be visualized by a *trend curve*. Based on the trend curve generated over quantitative data, we identify *time segments* for those long-term movements that can have positive or negative trend values ("ups" and "downs" on the market). Correlating this segments to the news stream, we identify a priori three trend classes: positive, negative and neutral class and divide the news stream in the 3-category text corpus. Analyzing text corpus, we will search for specific, so called *trend-indicating* keywords and statements. Trend-indicating keywords from the financial market domain are i.e. *cut*, *concern*, *recession*, etc. These simple keywords are subject to what we call trend indicating *language patterns*.
When analyzing text corpus, we are concentrating on trend indicating language structure and on the characterization of this structure. Firstly, we propose to divide the identification of trend indicating language patterns (in the following also called simple *trend patterns*) in the non-semantic feature extraction and in semantic feature annotation (more in sections 3.1 and 3.2).

In the following, we briefly describe stages in our proposed approach for the trend recognition method.

3.1 Non-semantic trend patterns

Since we analyze a given text corpus that is divided in trend classes, the "simplest" method for identifying trend patterns is the counting of the most frequent keywords or the TFIDF-method[15]. Different methods from *text mining* can be successfully applied in order to identify keywords or simple statements from the text corpus. However, we assume that not every keyword or statement extracted from the given trend class in text corpus is the trend-indicating one. The interesting point is how to recognize whether given keywords or statements are trend-indicating or not.
In particular we rely on the observation that there are characteristic words used in different domains describing the customer's opinion and/or her sentiment[2][9][19]. Following from this, since most sentiment indicating words are *adjectives* whereas the *nouns* build the sentiment concepts, then a possible and very simple trend pattern in the text could consist of an adjective-noun word pair. Using WordNet[6] or a Part-Of-Speech analysis, we could identify these pairs in the text corpus. Regardless, we assume that the search for trend patterns requires more complex text analysis then the POS. We assume, that we should investigate taxonomic and non-taxonomic relations between identified keywords or simple statements. Additionally, we should consider the semantic orientation as described in [7] and [19].

3.2 Trend pattern ontology

The non-semantic trend feature extraction provides a basis for a trend pattern structure. This can be useful for both, analyzing trend patterns on the non-semantic level and creating a trend knowledge base that provides insight into the general characteristic of the trend patterns. A knowledge base can be realized as a classic ontology. We propose the application of an adapted Extreme Tagging System (ETS) as a knwoledge base for trend recognition. An ETS as introduced in [18], is an extension of collaborative tagging systems which allow for the collaborative construction of knowledge bases. An ETS offers a superset of the possibilities of collaborative tagging systems in that it allows us to collaboratively tag the tags themselves, as well as the relations between tags. ETS are not destined to exclusively produce hierarchical ontologies but strive to allow the expression and retrieval of multiple nuances of meaning, or semantic associations. Our propose in this research is to use these novel knowledge acquisition techniques, which are based on lightweight annotations in social environments, in order to generate a semantic description for the analyzed application field. We will apply an adapted ETS in order to gain expert knowledge of trend recognition in the business field. We expect that the use of an ETS will bring an easy

[6] http://wordnet.princeton.edu/

retrieval and extraction of the expert knowledge in the form of a RDF triple set. An initial set of tags (which should be tagged by experts in a given domain) will be generated from the selected trend features that are extracted in a non-semantic way from the text corpus (described in 3.1). Experts using the ETS will play the "association game" on the initial tag set. Created association sets will be automatically converted to RDF-data. Produced RDF triple set will be then used to generate a trend scheme. Furthermore, we will use the data from ETS as the input for another feature extraction from the texts.
Combining the non-semantic search for trend patterns with the association sets based on expert knowledge, we aim to create an appropriate semantic trend pattern scheme- a trend pattern ontology- that will be applied to a learning algorithm.

3.3 Learning Trends

Regarding different possibilities of learning methods from machine learning [11][14][21], we firstly propose to use the supervised learning approach. Hence we work with strictly separable text classes- the texts with positive trend indicating patterns cannot belong to the neutral or negative trend category at the same time- standard classification seems to be an appropriate learning form for the trend recognition problem, particularly where the trend classes' ranges are well separable. With regard to the evaluation of the advantages achieved through applied semantics to the learning process, we propose to use firstly decision trees (i.e. C4.5) or decision rules [21] which both allow the vizualization of the learned model. Learning trends with decision trees means here learning trend indicating language patterns that are expressed in RDF-triples.
However, once the feature space has been created from the text corpus (as described in 3.1 and 3.2), we can use the features in order to validate the assumptions about the positive, negative and neutral trend indicating patterns. Therefore, we can use clustering as the alternative learning method for automatically assigning the trend classes' ranges. In our research we are considering also different alternative learning algorithms like rough sets, fuzzy case reasoning, neural networks or inductive learning approaches [14][21][13][8] in order to find the most appropriate one for the semantic-based trend recognition.

4 Future work

Given the directions for research outlined in section 3, we have chosen to continue our work on the theoretical and the practical solutions in order to create a prototype of here described semantic-based learning method for trend recognition in simple hybrid information systems.

5 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 would like to thank their supervisor, Prof.Robert Tolksdorf and the TREMA-project partners for the support of this work.

References

1. Ahmad, K.: Events and the Causes of Events, In Conference on Terminology and Knowledge Engineering 2002, online: http://www.computing.surrey.ac.uk/ai/TKE
2. Archak, K., Ghose, A., Ipeirotis, P. G.: Show me the Money! Deriving the Pricing Power of Product Features by Mining Consumer Reviews
3. Charu, C. Aggarwal: A framework for diagnosing changes in evolving data streams, SIGMOD 2003: Proceedings of the 2003 ACM SIGMOD international conference on Management of data, 575-586,(2003)
4. Hevner, A. R., March, S.T., Park, J., Ram, S.: Design Science in Information System Research,MIS Quarterly 2004
5. Esuli, A. and Sebastiani, F.: SentiWordNet: Publicly Available Lexical Resource for Opinion Mining
6. Gillam, L., Ahmad, K., Ahmad, S., Casey, M., Cheng, D., Taskaya, T., Oliveira, P.C.F. and Manomaisupat, P.: Economic News and Stock Market Correlation: A Study of the UK Market. In Conference on Terminology and Knowledge Engineering 2002, online: http://www.computing.surrey.ac.uk/ai/TKE
7. Hatzivassiloglou, V. and McKeown, K. R.: Predicting the semantic orientation of adjectives. In Proceedings of the 35th Annual Meeting of the ACL
8. Han, J., Kamber, M.: Data Mining Concepts and Techniques, 2.Ed. Morgan Kaufmann 2006
9. Hu, M., and Liu, B.: Mining and summarizing customers reviews. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004) (2004), pp. 168-177
10. Mei, Q., Liu, C., Su, H., and Zhai, C.: A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In Proceedings of the 15th International Conference on World Wide Web (Edinburgh, Scotland) WWW'06 ACM Press, New York, NY, 533-542.
11. Mitchell, T.M.: Machine Learning, Mc-Graw-Hill, 1997
12. Morinaga, S., Yamanishi, K..: Tracking Dynamics of Topic Trends, KDD'04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge Discovery and Data Mining, 811-816, ACM NY
13. Pal, S.K. and Mitra, P.: Pattern Recognition Algorithms for Data Mining, CRC Press LLC 2004
14. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, Prentice Hall, 2.Ed.2003
15. Salton, G., Buckley Ch.: Term Weighting Approaches in Automatic Text Retrieval, 1988 Information Processing and Management: an International Journal archive Volume 24 , Issue 5 (1988) Pages: 513 - 523

16. Schleutermann, S. and Heidl, B. and Finsterer, U.: Trenderkennung beim Nierenfunktionsmonitoring auf der Intensivstation, GMDS 139-142, 1996
17. Simon, H.A.: The Science of the Artificial, Ch.4: Remembering and Learning, MIT Press, Third Edition (1996)
18. Tanasescu, V., Streibel, O.: Extreme Tagging: Emergent Semantics Through the Tagging of Tags. In International Workshop on Emergent Semantics and Ontology Evolution, ISWC2007
19. Turney, P.D., and Littman, M.L.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems 21, 4 (2003), 315-346
20. Vejlgaard, H.: Anatomy of a Trend Mc-Graw-Hill, 1.Ed. 2007
21. Witten, I.h., Frank, E.: Data Mining Practical Machine Learning Tools and Techniques, 2.Ed.Morgan Kaufmann 2005
22. Witten, I.H., Gori, M., Numerico, T.: Web Dragons: Inside The Myths of Search Engine Technology, Morgan Kaufmann 2007
23. Wong, W.-K., Moore, A., Cooper, G., Wagner, M. What is Strange About Recent Events (WSARE) in Journal of Machine Learning Research 2005
24. `www.google.com/trends`
25. `www.blogpulse.com`
26. `www.projekt-trema.de`