Semantic-Based Learning Method for Trend Recognition in Simple Hybrid Information Systems

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. Determination and early detection of emerging trends can be retrieved from numeric data as well as from texts. Using texts for trend mining brings advances for the recognition process. The systematic integration of informaion descriptions and metadata schemes enable the additional semantic analysis of the available information. In this paper, we introduce the issue of trend recognition in information systems based on texts and numeric data. We present our idea of a novel semantic based learning approach which supports the recognition of temporal changing patterns, here called trends, in texts. Since our work is in the early stages, we provide an outline of the direction of our research, providing an overview of the main research issues.

Key words: information system, trend mining, trend recognition, learning method, pattern recognition, trend patterns, semantic trend scheme

1 Introduction

Information systems provide knowledge in a raw state. In order to discover knowledge in any given information system, we have to search through the stored data and analyse this. However, the form and the quality of the final information (the discovered knowledge) depends on the chosen methods of analysis as well as on the amount and the quality of the original retrieved information, ergo on the quality of the analysed data. The more meaningful the stored information, the more powerful is the knowledge we can retrieve from any given information system. Therefore, additional benefits for the knowledge discovery emerges with a rigorous method of analysis which includes both, quantitative and qualitative data.
Repositories consisting of texts and numeric data, each associated with a specific area of application, can be interpreted as information systems. Since the qualitative and quantitative data provide hybrid properties of this system, we refer in this work to a simple hybrid information system. This is simply a system providing information based on qualitative and quantitative data.

We can find many examples of simple hybrid information systems in different areas of application, i.e. in financial market analysis, customer opinion analysis, market research, weather forecasting, traffic analysis, aerial surveillance, etc. Tasks such as strategic planning, decision support, early emergency detection, and trend recognition are parts of those application areas and can be supported by intelligent data analysis. However, it seems to lack computational methods of trend analysis that provide for multimodal data, in particular, numeric data and texts.
The key objective of this research will be to develop a semantic based learning method for trend recognition in simple hybrid information systems. Research projects such as GIDA[1] and TREMA[2] have shown that there is a huge demand for research on and development of useful trend mining methods which are able to include analyses of textual information in the process of trend recognition. We use a specific but free available data set and text corpus as an example of a simple hybrid information system. With regard to the multimodal data, we will develop an adequate trend recognition method for the recognition of temporal changing patterns in textual information sources relevant to the given business field. The focus of this research will be on developing a solution relevant to the trend mining problem which combines a Data Mining approach and adequate Semantic Web technologies.
In the following sections, we describe our approach and give an insight into our idea. Section 2 contains the general methodology of our research. In section 3, we state the research issues of our work. Following to that, we compare the partially related work in section 4. In Section 5 we outline the future work.

2 General Approach

The aim of the research is the development of an intelligent trend recognition method that based on triangulated data is able to find trends in texts. We are considering several approaches. One of them is the intuitive approach. Choosing a specific business field, and regarding the corresponding simple hybrid information systems, we are comprehending the human's way of thinking and acting in the process of trend recognition. This can be simplified and defined in few general steps. One precondition here is that we assume that the 'human' is a specialist in their choosen business field and has gained experience in the trend recognition within this particular field. Considering the trend mining in the finance markets, the following main steps are accomplished by a finance market analyst:

- Correlation of numeric data with texts: the numeric data is analysed by computer-based methods and the trends are estimated mathematically. Texts (i.e. business news and the opinions of other analysts) serve here as the

[1] Generic Information-based Decision Assistent
[2] Trend Mining, Fusion and Analysis of Multimodal Data

explanation for the trends and are mostly analyzed on demand by a human specialist.
- Trend recognition and trend forecasting: based on experience, the human specialist can, in most cases, spot emerging trends.

Deliberating on this intuitive approach, we detected the correlation part, the recognition and learning part as well as a knowledge/experience component of the trend recognition process. Since human recognition and human learning process are different from the learning and recognition processes of a machine (i.e. aspects of recognition and association capability in semantic rich domains, intuition [15]), it is not possible to exclusively rely on our intuitive approach. In this case we want to treated as the complement to the classic design science paradigm.
The design science paradigm derived from [15], described in [4], is "fundamentally a problem solving paradigm. It seeks to create innovations that define the ideas (...) through which the analysis (...) and use of information systems can be effectively and efficiently accomplished"[4].

3 Mining trends in simple hybrid information system

Referring to trend analysis in Data Mining, the trend analysis process consists of four major components:[7](s. 490)

- Trend or long-term movements
- Cyclic movements or cyclic variation
- Seasonal movements or seasonal variations
- Irregular or random movements

Due to the work described in [11], where text based trend analysis is presented through the example of topic trends, texts streams are analyzed with regard to the following tasks in topic analysis:

- Topic Structure Identification: learning a topic structure in a text stream
- Topic Emergence Detection: detecting the emergence of a new topic
- Topic Characterization: identifying characteristics of topics

Since the components of trend analysis in the first definition give us an overview over the trend "arts", from the second definition we can derive the prototypic main stages in text-based trend analysis. In our research we are concentrating in particular on the long-term movements. This is what we call "trend-based trend detection"[3]. When analyzing text corpus, we are concentrating on trend indicating language structure and on the characterization of this structure. Starting from the numeric data, sequences of textual information from the choosen business field[4] will be related to a numeric time series in order to match the texts

[3] Analyzing irregular/random movements would mean the "event-based trend detection"

[4] We aim to use the *finance market* domain as an example of simple hybrid information system

to the trend (long-term movements). The identification of trend indicating language patterns will be divided in the non-semantic feature extraction and in semantic feature annotation (more in sections 4.2 and 4.3).
In the following, we briefly describe stages in our proposed approach for the trend recognition method.

3.1 Numeric data vs. text data

We handle numeric data as the base for the trend segments identification. In particular, we are concentrating on a chosen financial instrument and its market values in a specified time period. Using time series analysis on numeric data, we will identify the interesting trend segments. Depending on these trend movements, we will divide the text corpus in positive, negative or neutral texts.[5] These text sets will be used as the training data set with three training classes referring to three possible trend movements. We are not going to concentrate on different trend analysis techniques for numeric data here. However, the interesting point is the comparison of the existing trend analysis methods for numeric data in order to derive an approach of trend analysis in the texts. As the research described in [2] shows, it might be helpful to mix different approaches, i.e. a mixing econometrics and text mining methods is succsessful for mining consumer reviews. Similarly, we will also consider another possibilities- combining numeric based trend segmentation- of correlating numeric data with textual data in order to improve the trend analysis process.

3.2 Extracting trend patterns from texts

We assume, that different methods of Text Mining can be used for the extraction of trend patterns from texts. However, the most interesting point is the definition of a trend pattern in text. In our research, we define trend patterns as temporal changing language (syntactic and semantic based) patterns in texts. These are simple patterns based on so called *trend-indicating* keywords and statements. Trend-indicating keywords from the financial market domain are i.e. *cut*, *concern*, *recession*, etc. 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][8][17]. 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 can identify these pairs as trend features. However, in order to find an appropriate trend pattern structure, we are interested in analyzing the training set applying different text mining methods, i.e.using TFIDF-algorithm on a priori selected

[5] The text corpus is available in German language. The generation of the training set is kindly supported by the TREMA project and cooperation with a German company, neofonie GmbH

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

features (i.e. on bigramms, collocations, or POS-pairs)[19][20]. Since the training set is divided in three trend categories, we will search for the appropriate: positive, negative and neutral trend indicating language patterns. Therefore, the main part of trend pattern extraction from the texts will be the trend feature selection and trend feature extraction.

3.3 Semantic trend scheme

While the non-semantic trend feature extraction provides a basis for non-semantic trend pattern structure, a semantic trend scheme should provide insight into the general characteristic of the trend patterns. However, we are not going to annotate the extracted text patterns directly. Instead, we propose the application of an adapted Extreme Tagging System (ETS) as a complement to the non-semantic features. An ETS as introduced in [16], 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. With the cooperation of the industry, 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 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 trend recognition) will be generated from the selected trend features that are extracted in a non-semantic way from the text corpus (described in 3.2). 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 that will be applied to a learning algorithm.

3.4 Learning Trend Patterns

Regarding different possibilities of learning methods from machine learning [10][13][19], 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 [19] which both allow the vizualization of the learned model.
However, once the feature space has been created from the text corpus (as described in 3.2 and 3.3), 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 [13][19][12][7] in order to find the most appropriate one for the semantic-based trend recognition.

4 Related Work

Up to this point we have not found any other research that focuses on semantic-based learning approach for correlated trend recognition using numeric data and texts. However, there are some related studies that have informed our work, and that we are discussing below.
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 be important for our work. The authors of [14] 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. Since the anomaly detection is a part of trend recognition, the rule-based algorithm for the early detection of disease outbreaks introduced in [21] is also useful for trend mining but not relevant for our research since we are not concentrating on event-based trend recognition. Trends based on keyword search statistics are well visualized by the Google-Trends [22] 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 [23] 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. Contrary to this, we are assuming a simplified, data and text oriented, trend definition (see description in 3.2).

The research project GIDA[7][6][1] and its follower, TREMA[8], 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. However, we are not going to concern ourselves 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 3.3, instead of applying classic ontologies.
As last, the work described in [9] could be very useful for us. In particular, the definitions of *theme*, *theme life cycle*, and *theme snapshot* could be important for our approach.

5 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. Our research will pose the following questions:

- How helpful is for our trend recognition approach SentiWordNet[9][5]?
- Is there an appropriate semantic trend scheme for textual data?
- Can an ETS bring the expected benefits to the semantic trend scheme that we are searching for?
- How independed from the given language are the semantic trend patterns?
- Which approach from Pattern Recognition and Data Mining is appropriated for the semantic based trend recognition?
- How strong is the semantic trend recognition depending on the given application domain?

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`

[7] Description online: www.computing.surrey.ac.uk/ai/gida
[8] Project website: www.trema-projekt.de
[9] http://sentiwordnet.isti.cnr.it/

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. Han, J., Kamber, M.: Data Mining Concepts and Techniques, 2.Ed. Morgan Kaufmann 2006
8. 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
9. 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.
10. Mitchell, T.M.: Machine Learning, Mc-Graw-Hill, 1997
11. 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
12. Pal, S.K. and Mitra, P.: Pattern Recognition Algorithms for Data Mining, CRC Press LLC 2004
13. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, Prentice Hall, 2.Ed.2003
14. Schleutermann, S. and Heidl, B. and Finsterer, U.: Trenderkennung beim Nierenfunktionsmonitoring auf der Intensivstation, GMDS 139-142, 1996
15. Simon, H.A.: The Science of the Artificial, Ch.4: Remembering and Learning, MIT Press, Third Edition (1996)
16. Tanasescu, V., Streibel, O.: Extreme Tagging: Emergent Semantics Through the Tagging of Tags. In International Workshop on Emergent Semantics and Ontology Evolution, ISWC2007
17. 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
18. Vejlgaard, H.: Anatomy of a Trend Mc-Graw-Hill, 1.Ed. 2007
19. Witten, I.h., Frank, E.: Data Mining Practical Machine Learning Tools and Techniques, 2.Ed.Morgan Kaufmann 2005
20. Witten, I.H., Gori, M., Numerico, T.: Web Dragons: Inside The Myths of Search Engine Technology, Morgan Kaufmann 2007
21. Wong, W.-K., Moore, A., Cooper, G., Wagner, M. What is Strange About Recent Events (WSARE) in Journal of Machine Learning Research 2005
22. `www.google.com/trends`
23. `www.blogpulse.com`
24. `www.projekt-trema.de`