Trend detection in formal models of dynamic domains

Uta Hellinger*

Insitut AIFB, Universität Karlsruhe
uta.hellinger@aifb.uni-karlsruhe.de

Abstract: Today, traders in financial markets are confronted with the problem that information is distributed over diverse sources and that there is too much information available. In our work we develop methods and tools to help traders to overcome this information overload. The goal is to help traders develop a better understanding of the importance of a piece of information. We will present a modelling paradigm that combines static domain knowledge with change information and with the possibility to access former states of the knowledge base as well as the history of actions that have led to the current state.

1 Introduction

Today, traders in financial markets are confronted with the problem that too much information is made available from a growing number of sources. The goal of our work is to provide methods and tools that can help to overcome this information overload by integrating news from various sources, by filtering relevant news and by enabling the estimation of the news' impact.

Several authors have proposed to use machine learning methods to learn relations between news and market developments. The goal of these methods was to develop tools predicting the market's reaction to the publication of some information. We think that such a purely machine learning based approach is not suitable (as the results obtained using these methods show). We propose to produce a more comprehensible model of news and their relation to market developments. This will be achieved by using events and their parameters as features for learning relations. The two approaches yield very different feature sets as illustrated in figure 1. The benefit of our approach is that rules can be obtained that directly help traders in their decision making as they are not based on the words used in the text but on the actual content.

The development of this kind of analysis methods raises a set of research questions:

- What kind of information should be used as features? Which kind of events should be considered?

*Supported by the Deutsche Forschungsgemeinschaft (DFG) under the Multipla project and the Graduate School Information Management and Market Engineering.

Text: Cablevision buys Newsday from Tribune for $650M
Terms: Cablevision, buy, Newsday,. . .
Event-based: Acquisition, Buyer: Cablevision, Target: Newsday, former owner: Tribune,...

Figure 1: Text and features for existing and new methods

- How can additional information (e.g. previously published information) considered for prediction? This is relevant as information is only expected to influence the market when it is first published.
- How can a model representing the frequently changing information on events and expectations on future events be efficiently implemented?
- How can a regression model be built using structured information as input variables?

Although our work focuses on modelling financial markets, the idea of using features that directly rely on the content of text instead of its wording may be relevant in other application domains of text mining.

The rest of this proposal is structured as follows: in section 2, we will briefly introduce the process of our proposed learning method. In section 3 we present requirements for the domain model used for representing structured information extracted from texts, section 4 discusses related work, before dealing with the workplan in section 5 and concluding in section 6.

2 The learning process

The subsequent steps of the learning process are shown in figure 2. The first step in the analysis of news is their collection from various, heterogenuous sources like newswires, forums, blogs, or collaborative tools. In the feature extraction phase the text is transformed into a set of features. This phase consists of two steps in our approach. The first step is to extract the relevant content from the text and thus obtain a structured representation of the text's content. The second step will then be to derive feature weights from the extracted content, possibly also using already published information. The thus derived feature vectors can then be used for an estimation of the expected market response in terms of volatilities or returns.

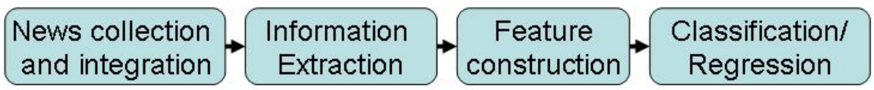

Figure 2: The learning process

3 The domain model

We have identified a set of specific requirements for the formalism employed for building the domain model. These will be presented in the following.

Information in the model may either change rarely (like the name of a company's CEO) or frequently (like a company's results - there may be a variety of rumours preceding the actual announcements of the actual figures). Therefore, it is crucial to store the changing information in an efficient way.

Information on the events that entail changes should be available later on, such that they can be used for the feature construction for later arriving messages. One example where this information is needed is the decision whether some information is already known or not (as the market should only respond to the first publication of a piece of information).

A message may not contain factual information, but other uncertain information like expectations for future developments, like for example a rumour. It should therefore be possible to distinguish between certain and uncertain information in the model.

A set of event types that are considered will be defined. These event types will be considered as change operators, narrowing down the space of changes that have to be expected.

This kind of model can easily be stated as a labeled transition system. A *labeled transition system* is a pair $(P, \rightarrow)$ where P describes the set of possible states, and $\rightarrow \subset P \times Act \times P$ such that for $p \in P$ and $a \in Act$ the class $q \in P | (p, a, q) \in \rightarrow$ is a set [BPS01]. The set P corresponds to the set of possible states of the domain, the set Act to the set of considered events and $\rightarrow$ defines the changes that are entailed by an event. Although this part of the translation is straightforward, we have to solve the problem of describing the states and events in a concise way and of defining the semantics of $\rightarrow$.

4 Related work

There is a number of systems that use text classification for the prediction of asset price changes based on text features. There are approaches to predict various target values and different approaches for determining the target variable (trend detection vs. fixed period after news release) and different classifiers are proposed (Naive Bayes, SVMs). An overview of these systems can be found in [MK06]. Our approach is different from this approach as we first extract relevant information on events from the news text and incorporate them in a formal model from which we can then determine the expected market response.

The finance literature provides us with methodology to study the influence of given event types on the market (event study, MacKinlay [Mac97] describe the methodology). The results of various event studies will help us determine which information should be incorporated in our formal model.

To the best of our knowledge, no approach exists that allows modelling of expectations and matching them against actual events. Halaschek-Wiener and Hendler [HWH07] have however proposed a news syndication framework that allows the definition of information needs as conjunctive ABox queries that can then be matched against actual news that are formalized as ABox assertions. Their approach might be adopted for this purpose.

For the modelling of temporal aspects in ontologies, the combination of temporal logics and description logics has been proposed by Artale et al [AKL$^+$07]. Their approach is to introduce the time operators of linear time logic in description logics and to allow temporal statements about concepts. This approach differs from our approach in that it seems neither possible nor desirable to describe all possible future changes within the ontology as their are many different ways in which the model will evolve from the state given at a certain point in time. An approach that models possible changes at a meta-level seems therefore more appropriate. Meta-level here means to define operators that can be applied to the system at a given time.

The combination of ontologies with π-calculus has been described by Agarwal [Aga07]. He proposes to model processes as series of states where the states are described by a knowledge base and the transitions between states are described using the π-calculus. The drawback of this approach in our application is that the π-calculus is especially adapted for describing communications between processes. Communication is however not an issue in our scenario.

A lot of work has been done on ontology evolution, which deals with handling changes in ontologies [Sto04]. Yet, the approaches we are aware of mainly concentrate on changing schema knowledge, while in our scenario we mainly have to deal with changing instance data.

5 Status and future work

So far, we have acquired the data which is necessary in order to evaluate our approach. We have acquired all news published by Reuters in 2003 and intraday trades and quotes for 240 major exchanges in the same period. We have implemented a set of aggregation functions that allow the selection of news, and the aggregation of trades and quotes into a set of indicators of the financial markets like returns or volatility in a certain time period.

We are currently working on the identification of events that should be considered in our approach. We use the OpenCalais web service[1] to extract structured information from the news. So far we limit ourselves to the set of events that is extracted by OpenCalais and try to determine which of them are promising to study as they entail significant market responses.

We are also working on the development of the domain model. We have identified requirements for the formalism that will be used to build the model and are currently working on ways to incorporate uncertain information as well as matching of pieces of information in the model.

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

The next steps will then deal with the learning algorithms: we will have to incorporate the structures that are extracted from text as numeric values such that the application of standard machine learning methods is possible.

The final step will be to evaluate the results of applying our formalism to model financial markets and of determining patterns in market responses to events.

6 Conclusion

The goal of our work is the development of news analysis methods that helps traders find relevant information and react appropriately to it. In this context, our work will make the following contributions:

- We develop a framework that allows the analysis of news and extraction of the most important information. This will enable the assessment of the prospective market impact of newly published information.
- We propose text mining methods that use content-based features, thus enabling the construction of more comprehensible models.
- We develop a modelling formalism for modelling uncertain and frequently changing information in ontologies.
- We develop methods to detect trends and frequent patterns in the dynamic knowledge base and in the event series such that we can predict future developments.

References

[Aga07] S. Agarwal. *Formal Description of Web Services for Expressive Matchmaking*. PhD thesis, Universität Karlsruhe, 2007.

[AKL+07] A. Artale, R. Kontchakov, C. Lutz, F. Wolter, and M. Zakharyaschev. Temporalising Tractable Description Logics. In *Proceedings of the Fourteenth International Symposium on Temporal Representation and Reasoning*. IEEE Computer Society Press, 2007.

[BPS01] J.A. Bergstra, A. Ponse, and S.A. Smolka. *Handbook of Process Algebra*. 2001.

[HWH07] C. Halaschek-Wiener and J. Hendler. Toward expressive syndication on the web. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 727–736, New York, NY, USA, 2007. ACM.

[Mac97] A. Craig MacKinlay. Event Studies in Economics and Finance. *Journal of Economic Literature*, 35(1):pp. 13–39, March 1997.

[MK06] M. A. Mittermayer and G. Knolmayer. Text Mining Systems for Market Response to News: A Survey. Working paper, 2006.

[Sto04] L. Stojanovic. *Methods and Tools for Ontology Evolution*. PhD thesis, Universität Karlsruhe, 2004.