VIORELA MILEA

News Analytics for Financial Decision Support

News Analytics for Financial Decision Support

News Analytics for Financial Decision Support

Kwantitatieve nieuwsanalyse voor financiële besluitvorming

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof.dr. H.G. Schmidt

and according to the decision of the Doctorate Board

The public defense shall be held on

Thursday, 7 February 2013, at 15:30 hours

by

Viorel Milea
born in Bucharest, Romania.

Doctoral Committee

Promotor:	Prof.dr.ir. Uzay Kaymak
Co-promotor:	Dr.ir. Flavius Frasincar
Other members:	Prof.dr.ir. Rommert Dekker
	Prof.dr.ir. Geert-Jan Houben
	Prof.dr.hab. Witold Abramowicz

Erasmus Research Institute of Management - ERIM
The joint research institute of the Rotterdam School of Management (RSM)
and the Erasmus School of Economics (ESE) at the Erasmus University Rotterdam
Internet: http://www.erim.eur.nl

ERIM Electronic Series Portal: http://hdl.handle.net/1765/1

ERIM PhD Series in Research in Management, 275
ERIM reference number: EPS-2013-275-LIS
ISBN 978-90-5892-321-9

Design: B&T Ontwerp en advies `www.b-en-t.nl`

This publication (cover and interior) is printed by haveka.nl on recycled paper, Revive®. The ink used is produced from renewable resources and alcohol free fountain solution. Certifications for the paper and the printing production process: Recycle, EU Flower, FSC, ISO14001.
More info: `http://www.haveka.nl/greening`

"We must die as egos and be born again in the swarm,
not separate and self-hypnotized,
but individual and related."
– Henry Miller

SIKS Dissertation Series No. 2013-01
The research reported in this thesis has been carried out in cooperation with SIKS, the Dutch Research School for Information and Knowledge Systems.
http://www.siks.nl/

Preface

My interest in research started while I was writing my bachelor thesis, under the supervision of dr. Jan van den Berg. Our discussions, ranging from my thesis to the Feynman lectures, helped me further along an academic career path, and resulted in my application for a PhD position at the department. Jan, thank you for pointing me in a direction I have never regretted choosing.

To a large extent, a PhD is a guided journey. I was guided by prof.dr.ir. Uzay Kaymak and dr.ir. Flavius Frasincar, my promotor and co-promotor. I am thankful to both for their time and trust in me during the past years.

Uzay, your door was always open and our discussions fueled many thoughts and ideas, some published. You pleasantly pointed out, regularly, that the 'Ph' in PhD really stands for Philosophy. But also that science is much more than a fun thought experiment.

Flavius, I have benefited a lot from your help. Your high standard of quality and patience inspire me. The lunches where Semantic Web and world history come together in single sentences make me unrealistically overrate the catering, in retrospect.

I thank the PhD committee for being here today, and for the feedback provided on my thesis. Prof.dr.ir. Rommert Dekker, you were kind enough to read and evaluate my thesis proposal at the beginning of my PhD, and I hope you find the end result close to what was imagined back then. I look forward to our cooperation during the coming years. Prof.dr.ir. Geert-Jan Houben, I am thankful for the feedback that you gave during my PhD, and for the discussions on the tOWL language, and the semantics of timeslices. Prof.dr.hab. Witold Abramowic, thank you for your help during the tOWL project and the hospitality in Poznan. Prof.dr. Euripides Petrakis, thank you for your ideas on space and time, and for receiving us so warmly in Chania. Prof.dr. Patrick Groenen, thank you for finding the time for good advices as ERIM Director of Doctoral Education.

I was offered the opportunity to get to know and work with researchers from different universities during the TOWL project. I have worked closely with Kelly Zervanou and Tomaso di Noia, whom I thank for the intense discussions and ideas we exchanged during the project. Also, it has been a pleasure cooperating with dr. Tomasz Kaczmarek from Poznan University of Technology, Bram Stalknecht and dr. Mark Vreijling from SemLab, prof.dr. Eugenio di Sciascio from Politecnico di Bari, and Rachel Yager from Citigroup.

During the past years, I have had the opportunity of interacting with many academic staff members as well as PhD candidates active in different departments of the Erasmus School of Economics. The academic environment provided here is unique, and I consider myself lucky to complete my PhD at this Faculty. I am especially delighted with being part of the Econometric Institute and receiving the opportunity to further develop myself here while bringing my own contribution to research and teaching.

I am grateful to the Erasmus Research Institute of Management for supporting my research and participation in education, and for providing a good academic, as well as social environment. I was lucky to collaborate with various researchers during my PhD, with some of our work ending up in different publications. I enjoyed a lot working with Frederik and Alexander Hogenboom, Kees van de Sluijs, Michael Mrissa, and Nurfadhlina Mohd Sharef. I hope we will have many opportunities to work together in the future.

Along the years, many students have contributed to my research, by being enthusiastic enough to put some of our ideas into practice. I especially thank Remy Stibbe, Wijnand Nuij, Alex Micu, en Laurens Mast.

My PhD would not have been possible without the appropriate funding. This was made available by the European Union through the STRP FP6 project 26896: Time-determined ontology based information system for real time stock market analysis. Also, my research visit to the University of Bristol was made possible by the European Science Foundation through the COST action IC0702. Being part of the HERA research group has greatly helped in crystallizing my ideas on the tOWL language and in keeping up with the newest research in Semantic Web and Web Engineering.

Rui and Nalan thank you for the great time in the middle of busy days. Nufer, I appreciated our discussions very much, and getting used to a new room took time. Kyle, Tommi, Oke, Ruben, Eran, Sun, thank you for the fun that makes the PhD experience complete.

I am also indebted to Odilia and dr. Roder, whom I hope will appreciate proposition 9.

My paranimfs, Roxanne and Iulian, I thank for standing by me on this symbolic day.

Iulian, esti un prieten bun. Mi-e greu sa nu zambesc la amintirile ultimilor ani, si astept cu nerabdare urmatoarea dezbatere (in Rotterdam).

Dragi parinti, mi-ati fost mereu alaturi. Increderea si experienta voastra de viata m-au ajutat, in mare parte, sa ajung aici. Va multumesc.

Roxanne, mijn titel verbleekt bij jouw krullen. Ik hoop dat we nog vele mooie momenten samen zullen beleven.

Viorel Milea
Rotterdam, 2012

Contents

List of Tables

List of Figures

Chapter 1

Introduction

"What we call the beginning is often the end. And to make an end is to make a beginning. The end is where we start from." – T.S. Elliot

One third of all share trades in the United States were initiated through algorithmic trading in 2006. Recent estimates of this share for US equity trading reach as high as 77%, while for European markets this figure ranges between 30% and 50% (Foresight, 2010a). Machines are the new breed of trader.

Despite its widespread adoption and large-scale involvement across different exchanges, Automated Trading (AT), also known as High-Frequency Trading (HFT), amongst the most popular denominations, is yet to receive a proper definition. A description of this type of trading is given in a 2010 report of the Securities and Exchange Commission:

> *The term is relatively new and is not yet clearly defined. It typically is used to refer to professional traders acting in a proprietary capacity that engage in strategies that generate a large number of trades on a daily basis... Other characteristics often attributed to proprietary firms engaged in HFT are: (1) the use of extraordinarily high-speed and sophisticated computer programs for generating, routing, and executing orders; (2) use of co-location services and individual data feeds offered by exchanges and others to minimize network and other types of latencies; (3) very short time-frames for establishing and liquidating positions; (4) the submission of numerous orders that are cancelled shortly after submission; and (5) ending the trad-*

> *ing day in as close to a flat position as possible (that is, not carrying significant, unhedged positions over-night).* (SEC, 2010)

How far does the reach of such algorithms extend? On 6 May, 2010, the Dow Jones Industrial Average plunged close to 600 points over a period of time spanning less than five minutes, only to recover back to the level before the 'crash' twenty minutes later. This event is known as the 'Flash Crash', and is mostly attributed to (poorly designed) automated trading algorithms.

There is indeed evidence that, under certain assumptions, HFT can function as an accelerator to bubbles and crashes, and market dynamics in general (Sornette and von der Becke, 2011). However, academic studies also indicate that algorithmic trading narrows spreads, reduces adverse selection, and reduces trade-related price discovery, thus improving market liquidity and the informativeness of quotes (Hendershott et al., 2011; Hendershott and Riordan, 2011). In terms of risk, a recent study shows that HFT decreases volatility (Brogaard, 2012) although, under very unusual market conditions, HFT can actually contribute to increased volatility (Kirilenko et al., 2011). Finally, HFTs acting as market-makers significantly contribute to the success of new markets (Menkveld, 2011). Although it creates new ways for manipulating markets, algorithmic trading generally improves market efficiency and contributes significantly to lowering transaction costs (Foresight, 2010a).

The fast growth of the use of automated trading and the overall advancement of such technologies contribute to decreased profitability of such algorithms. This trend is set to continue in the same direction over the coming ten years (Foresight, 2010a). The competitive advantage in this market is most likely to come from developments in News Analytics (NA) – automated techniques for collecting, processing, and using the information contained in news messages (Mitra and Mitra, 2011). The value of using information from news in AT comes from enabling (re)action to new information before it is incorporated into market prices.

1.1 Electronic Markets

More than three centuries have passed since the crucial element of "technology" driving the financial markets was the horse. This drastically changed with the beginning of the twenti-

eth century and the use of telephones as the main means of communication. By the end of the twentieth century, financial markets critically depended on Information and Communication Technologies (ICT). Concurrently, in 1971, the National Association of Securities Dealers Automated Quotation (NASDAQ) became an electronic market where quotes and orders could directly be placed and retrieved by means of a computer (Black, 1971a,b; Foresight, 2010b). This new mechanism also allowed the automation of the order matching mechanism, overall leading to a decentralization of market access (Gomber et al., 2011).

The electronic financial markets of the 21st century, aided by new regulation, led to an increase in the availability and openness of exchanges, and subsequently to an increased number of market participants that can take advantage of lower transaction costs. The resulting spike in volume (see Figure 1.1) led to increased liquidity, and supported institutional traders in dividing large trades into smaller orders with less market impact. All this on a background of higher competition among exchanges (Angel et al., 2010).

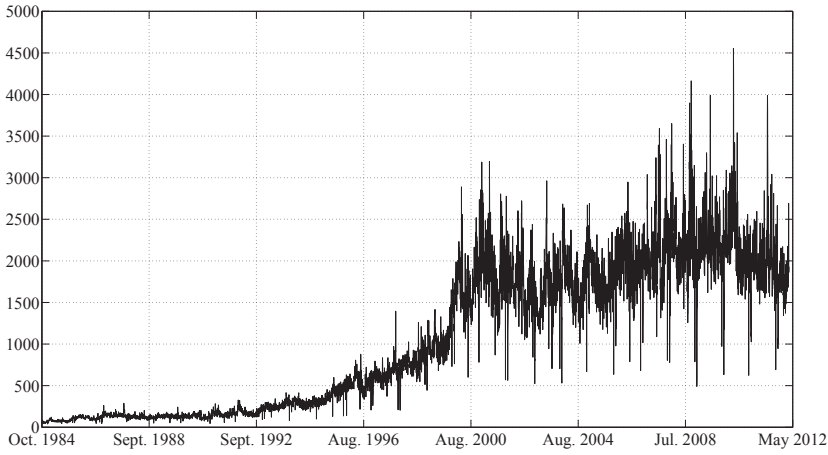

Figure 1.1: NASDAQ – Historical daily trading volume, in millions of shares
source: Yahoo! Finance

The large volume and wide availability of financial data typical for developed financial markets today do not only lead to increased transparency of these markets, but also to the devel-

opment and deployment of ever more complex models. Supported by the exponential rise in computing power available at decreasing relative prices, automated systems play an increasingly dominant role in financial trading. All these factors additionally contribute to a need for radically changing the business models of institutional facilitators in the financial sector (Dence et al., 2006). The contemporaneous electronic financial market might look only slightly changed to the superficial eye, but it is a completely new environment for all the market participants.

Most academics regard HFT as closely related to, yet significantly distinguishable from, Algorithmic Trading, with certain characteristics that distinguish it from the latter. Both forms of trading use direct market access and real-time data for completely automated trading decisions and order placing without human intervention. The focus of HFT however is rather on profiting from the buying/selling activity, thus accomplishing the role of the middleman, based on very low margins and very short holding periods throughout a day, cumulated over a large number of trades. Based on these characteristics, HFT requires trading in highly liquid instruments. Algorithmic Trading is a bit more general, focussing on longer holding periods and on splitting large orders for minimizing market impact of large trades (Gomber et al., 2011). Current algorithms used in markets rely increasingly on the use of news for high speed trading decisions. An overview of the emerging field of News Analytics, as well as the disciplines that come together under this umbrella term, is given next.

1.2 News Analytics

Information drives the markets. Available from an increasing number of sources, most of which are available on-line and to a wide audience, information is more than ever present in everyday life. Different news providers, as well as institutional entities, ranging from stock-listed companies to charities, use the Web as communication channel. While the problem of access to this information is diminishing at a fast pace, new challenges emerge. The most important such challenge consists of selecting just those pieces of information that are relevant to the user and aggregating them for decision-making.

A significant share of the information available today is presented as text. Generally speaking, these texts can contain either facts, opinions, or a combination of both. Those texts containing facts are a collection of (to a higher degree) objective statements about entities or events, while texts containing opinions are more subjective in nature and convey a sentiment on the subject being described (Liu, 2010). The sources of information themselves are diverse, but here we focus only on those sources deemed relevant for news analytics for financial decision support. We distinguish between main stream media, press releases of financial entities, (technical) reports of organizations and institutions involved in-, or overseeing, financial markets, message boards and social media, and finally, a source not yet considered in academic efforts in this area, dissident opinions.

Main stream media are the main provider of facts and opinions, and the high availability hereof is due mostly to the growth of the World Wide Web. Large, widely recognized news providers include Associated Press, Reuters, Dow Jones, Google–, and Yahoo! News, and many others, some of them also able to provide annotated news aimed directly at automated trading. Many other news sources populate the on-line landscape, local and international, general or aimed at specific niches in the market, and available in different languages.

Financial entities also increasingly use of the Web channel to broadcast information directly to stakeholders and market participants, in the form of press releases, periodical (financial) reports, and news messages regarding different aspects of the business. Companies also increasingly make use of social media for attracting and retaining customers, increasing brand loyalty, but also as communication channel to customers and stakeholders. For example, the official IBM twitter account, has more than 20,000 followers and more than 600 tweets. Boeing Airplanes has over 55,000 followers on Twitter, and has posted more than 3,000 tweets. An example tweet from Boeing announcing an order for the company's main product, posted on June 8th, 2012, is shown below:

Jakarta-based Lion Air announces a commitment for #Boeing Dreamliners for their new premium carrier Batik Air. `http://ow.ly/bsc2q` (source: Twitter.com)

Several approaches for the analysis of such information have been investigated in literature. At the highest level of abstraction, we distinguish between approaches that focus on content analysis, and sentiment analysis and opinion mining. News Analytics, an area concerned with the extraction and quantification of information contained in news is, for a large part, a combination of these two approaches.

Content analysis has a long history, which although disputable, can be regarded as having its roots in Rhetorical Analysis (McCroskey, 2001) and the early work of Aristotle (Aristotle, 1991). The migration, and concurrently the development of the field, towards a more quantitative approach was realized by Harold Laswell. His approach to the systematic analysis of communication is based on the question of "*who* says *what* to *whom* via what *channel* and with what *effect*?" (Neuendorf, 2002). This approach, with a large emphasis on the *effect* part of the question dominated his contributions to the field with most applications in political communication, starting with the seminal study of propaganda (Harold D. Lasswell, 1949), centred around the idea that "political power can be better understood in the degree that language is better understood". The applicability of content analysis, though on a different note, is shown by the predictability of the concentration of German troops from the music played on German radio stations (Roger D. Wimmer, 2010). With fast advances in the field during the Second World War, Content Analysis found successful applications in areas as varied as literature, psychiatry and psychographics, and marketing and advertising (Neuendorf, 2002). Worth mentioning is that content analysis cannot be the sole means by which the effect of content on its consumers is to be assessed, nor the single most accurate profiling tool. More recently, the advent of the Web brought Content Analysis to Finance and HFT.

The first attempt at content analysis in an economic context is presented in Klingemann et al. (1982). Here, the authors investigate the relationship between focus on wealth and wealth-related words in the speeches of the German Emperor and the state of the economy over the period 1870-1914. They find a strong relationship between the focus on wealth and the state of the German economy. More recent research, such as Tetlock (2007), relies on the General Inquirer dictionary (Stone et al., 1966) for explaining market prices and trading volumes. The author finds that a relationship exists between a daily Wall Street Journal column, 'Abreast of

the Market', and the market prices and trading volumes of that day for the stocks discussed in the column.

Rather than focussing on the content of communication in a broad sense, the analysis can specifically be aimed at sentiment and opinion classification, usually applied to textual information (Liu, 2010). Such approaches are gaining popularity in different areas of economics, and especially in the field of News Analytics. For example, (Sakaji et al., 2008) develop a method for the automated extraction of basis expressions that indicate economic trends. They are able to classify positive and negative expressions which hold predictive power over economic trends, without the help of a dictionary. (Zhang et al., 2009) focus on eight dimensions of sentiment: Joy, Sadness, Trust, Disgust, Fear, Anger, Surprise and Anticipation, and are able to provide visualizations of how these eight sentiments evolve over time on some concept, e.g., Iraq, based on news messages. The results are validated against ratings of human reviewers of the news messages, and the method performs satisfactorily in visualizing the evolution of these sentiments over time for the studied concepts.

(Esuli and Sebastiani, 2006) investigate term subjectivity and orientation in free text. The approach starts with two training sets consisting of Positive and Negative words, respectively. It extends these two sets with WordNet synonyms and antonyms of the words found in the sets. Next, a binary classifier is generated by a supervised learner that is able to categorize vectorized representations of terms and classify them as Positive or Negative. Extraction of fuzzy sentiment is done by (Andreevskaia and Bergler, 2006), where the authors are able to assign a fuzzy membership of Positive or Negative to a set of words using the Sentiment Tag Extraction Program (STEP). The extraction of investor sentiment from Web communication is done by (Das and Chen, 2007), based on a sentiment index of small investor opinions.

1.3 Goal and Perspective of this Dissertation

This dissertation addresses the problem of designing automated financial trading systems that are able to incorporate information from news in trading strategies. Our main focus is on news-based trading, rather than on building complete automated trading systems. More formally, the

problem statement that motivates this dissertation is:

How can fully automated financial trading systems be designed that are able to autonomously extract information from news messages, represent this information, and aggregate it into actionable, valuable knowledge?

This PhD thesis thus contributes to the newly emerged, growing body of scientific work on the use of News Analytics in Finance. Regarded as the next significant development in Automated Trading, News Analytics extends trading algorithms to incorporate information extracted from textual messages, by translating it into actionable, valuable knowledge.

The use of News Analytics in Automated Trading relates to three main tasks: i) the extraction of the information contained in news, ii) the representation of the information contained in news, and iii) the aggregation of this information into actionable knowledge. These three tasks are extensively addressed in this dissertation, through an *interdisciplinary* approach. For the extraction of information from news we rely on approaches borrowed from *Computer Science* and *Linguistics*. The representation of the information contained in news is realized by using, and extending, the state-of-the-art in *Semantic Web* technology. We do this by bringing together insights from *Logics*, *Metaphysics*, and *Computational Semantics*. The aggregation of information is done by using techniques and results from *Computational Intelligence*, *Finance*, and *Content Analysis*.

1.4 Research Questions

This dissertation answers four research questions related to news analytics for financial decision support. In this section we present these research questions and we outline their relevance. Also, a link is made to the chapters in which each of the research questions is answered, accompanied by a brief outline of this chapter. Finally, we describe the methodology that is used for answering each of the research questions.

Question 1: How can we use information from structured text for the prediction of economic trends?

The first investigation on the use of news for automated trading is done by using structured text, in Chapter 2. The advantages of such an approach relate to the similarity of structure amongst documents, which provides a way to measure changes in content over time. In this chapter we investigate the design of a framework for the computational content analysis of European Central Bank monthly statements. These documents provide an overview of the current and the expected state of the economy. We validate the framework empirically in the context of a financial market.

The methodology we use to answer this question is largely based on the standard content analysis approach. As a first step we quantify the content of a document into a 'content fingerprint' that shows, for different content categories, the extent to which each category is represented in the analysed document. Our analysis of content is based on the General Inquirer dictionary. Once quantified, the fingerprints are mapped to values of the MSCI Euro index by means of a fuzzy model. The choice for such a model is largely motivated by its ability to capture non-linearities, a phenomenon that we cannot exclude from the relationship we investigate. We then proceed to extending our approach by looking at fuzzy grammar fragments – more complex constructs that capture relevant words from content categories that appear in conjunction with economic terms – rather than solely words belonging to different content categories with no regard for context. Again the mapping of the content fingerprints to values of the index is done by using a fuzzy model. The performance of our approach in predicting economic trends based on information from structured text is measured in terms of the accuracy of predicting upward or downward movements of the studied index.

Question 2: How can we include information from news in trading strategies?

A broader approach to incorporating news in stock trading strategies is presented in Chapter 3. Here, we consider news in the form of events that can be extracted from news messages. The framework introduced in this chapter is validated by using genetic programming for finding

trading rules. Next to event information, we consider technical trading indicators as input for the trading rules.

The methodology we use for answering this research question consists of a stepwise approach for including information from news into trading strategies. Rather than relying on content analysis, we perform an event study coupled to expert information. Based on his approach, major events, such as mergers and acquisitions, are identified and extracted from news messages, together with the relevant entities involved and the timestamp of the event. The expected impact of these events on the financial entities is predefined based on opinions of experts, and constitutes the input to the prediction model that we use.

Next to event information, we use technical trading indicators for the prediction of stock prices. Based on a large pool of such indicators, we derive the trading rules with potential of profitability based on genetic programming. Through this evolutionary approach, random rules are generated and consequently changed (crossover and mutation) based on their fitness (ability to generated profitable trading strategies on historical data). The rules obtained upon convergence of the algorithm are considered (close to) optimal. The performance of the (close to) optimal trading rules that are based only on technical trading indicators is recorded.

In the final step, the procedure described above is again applied to a pool of variables that now consists not only of technical trading indicators, but also contains the news variable quantified based on events as described above. The (close to) optimal trading rules generated by this methodology are again recorded and compared to the trading rules that have been obtained based only on technical trading indicators.

Finally, based on this comparison, we are able to verify whether the proposed methodology is indeed able to increase profitability of trading rules when news messages are used for trading based on events and expert opinion. If the performance of the optimal trading rules obtained when the pool of variables contains both the news variable, as well as technical trading indicators, is higher than the performance of the optimal trading rules based only on technical trading indicators, then the inclusion of information from news in trading rules is more profitable, and thus successful.

Question 3: How can we represent temporal knowledge in a reusable way while enabling consistency checks and inference over the represented information?

The first two chapters of the dissertation motivate the need for a more systematic representation of knowledge when news is considered as input for automated trading algorithms. Chapters 4 and 5 discuss and address the need for: i) consistent and reusable representations of (financial) knowledge, ii) ability to account for the ephemeral character of the relationships typical to the financial domain, and iii) the ability to aggregate this knowledge automatically, while inferring implicit information from the representation.

We start from the state-of-the-art in Semantic Web languages and focus mostly on the Web Ontology Language, the current standard for Web ontologies. Through a layered approach we introduce support for several temporal features. The first layer of the proposed extension relates to the expressiveness of the language when concrete measures are involved. The increasing literature on extensions of formal languages with concrete domains stands at the base of our approach and provides the critical expressiveness and good computational properties for introducing time in the formalism.

The second layer we introduce builds upon the previous layer by adding expressiveness related to time as measure. We add the infrastructure required for such a purpose by introducing time points, relationships between time points, intervals, and Allen's 13 interval relations into the language. This represents time in its most concrete from, but does not address the issue of change.

The third layer we introduce deals with the representation and identity of entities that change through time. By building our approach on the perdurantist view of the world, we are able to describe individuals as collections of temporal parts, that can have different values or instants for their properties at different points in time. Finally, we show that these three layers cover the aspects of time and change to a satisfactory extent by representing a complex financial process in the proposed language.

Some of the static concepts used in the OWL language that we extend change their meaning in a temporal context. We conclude the presentation of the language with a study of such

concepts and the properties of the temporal language that we propose.

Question 4: How can we build automated trading systems that take into account temporal knowledge?

The final chapter of the dissertation shows how trading systems can be built that take into account both static and temporal knowledge. The framework presented in this chapter uses the language introduced in the previous two chapters. In this way, more information can be represented, and thus used in news-based algorithmic trading. The framework that we introduce is validated empirically by applying it for the extraction, representation, and aggregation of market recommendations issued by different financial analysts.

From a methodological perspective, we extend the framework presented in Chapter 3 with components related to the representation and reasoning with temporal data. Additionally, since the information is extracted from free text, we include modules related to the processing thereof and its incorporation into a temporal ontology. The validation of the proposed framework is done by building an application that deals with knowledge collected from different sources and aggregates this information, at relevant points in time, in order to issue a recommended decision.

1.5 Academic Contribution of this Dissertation

This dissertation adds to the emerging field of News Analytics in Finance. The added-value of this dissertation, from an academic point of view, can be divided roughly into contributions to the *design and implementation of systems that use news for trading*, and more theoretical contributions related to the *technology that is required for such systems*.

The dissertation presents three systems that are able to use news for financial trading. The first system we present, in Chapter 2 of the dissertation, deals with structured text. A framework for introducing information from news messages, in the form of events, is presented in Chapter 3. A more complex framework, able to deal with time and time-related information as extracted from (historical) news, is presented in the last chapter of this dissertation. The outcome of this final chapter is a general framework for time-enabled decision support systems.

Two chapters of this dissertation deal with developing technology that we deem necessary for News Analytics in Finance. These two chapters present a temporal web ontology language. We envision this language as a building block of all systems that deal with temporal knowledge. The language is able to represent abstract domain knowledge as well as more concrete (temporal) facts, such as the information contained in news messages.

Finally, despite its high relevance for business and the "next big thing in automated trading" status, News Analytics, especially in a financial context, is still in its academic infancy. Next to the concrete contributions of this dissertation, our work also delivers a proposal. Our discussion identifies the disciplines that can contribute to this emerging field, and the outcomes that can be expected from such an approach. , and focus points for future research.

1.6 Outline of this Dissertation

This dissertation consists of seven chapters, including an introduction and a conclusion. The five chapters that represent the main body of this thesis address one main theme: the incorporation of news into trading algorithms.

In Chapter 2 of the thesis we focus on the computational content analysis of European Central Bank statements, by presenting a framework that addresses this goal. Based on this framework, we provide two approaches that can be used in a practical context. Both approaches use the content of European Central Bank statements to predict upward and downward movement in the MSCI EURO index. General Inquirer is used for the quantification of the content of the statements. In the first approach, we rely on the frequency of adjectives in the text of the ECB statements in relation to the content categories they represent. The second approach uses fuzzy grammar fragments composed of economic terms and content categories. Our results indicate that the two proposed approaches perform better than a random classifier for predicting upward or downward movement of the MSCI EURO index.

In Chapter 3 we move towards a more general approach that allows for information from news to be included in trading strategies. The news that we consider consist of events extracted from news messages. The news messages are represented in free text without annotations. We test the

introduced framework by deriving trading strategies based on technical indicators and impacts of the extracted events. The strategies take the form of rules that combine technical trading indicators with a news variable, and are revealed through the use of genetic programming. We find that the news variable is often included in the optimal trading rules, next to the technical indicators.

In Chapter 4, we present a temporal extension of the very expressive fragment $\mathcal{SHIN}(\mathcal{D})$ of the OWL-DL language resulting in the tOWL language. Through a layered approach we introduce three extensions: i) Concrete Domains, which allows the representation of restrictions using concrete domain binary predicates, ii) Temporal Representation, which introduces timepoints, relations between timepoints, intervals, and Allen's 13 interval relations into the language, and iii) TimeSlices/Fluents, which implements a perdurantist view on individuals and allows for the representation of complex temporal aspects, such as process state transitions. The last part of this chapter shows the expressiveness of the newly introduced language by using an example from the financial domain.

Chapter 5 contributes to the further extension of the tOWL language. The design choices of the language pose new challenges from a temporal perspective. One such challenge is the representation of temporal cardinality. Another challenge consists of optimizing the temporal representations in order to reduce the number of axioms. One such optimization is temporal coalescing, which merges concepts that are associated with time intervals that either meet or share at least one instant with each other. In this chapter we formally introduce these concepts into the tOWL language and illustrate how they can be applied.

The final chapter brings together insights from Chapters 2 and 3, and the technology presented in Chapters 4 and 5. In Chapter 6 we present a time-based framework for decision support systems. The system uses the state-of-the-art tOWL language for the representation of temporal knowledge and enables temporal reasoning over the information that is represented in the knowledge base. Our approach is the first to use state-of-the-art Semantic Web technology for dealing with temporal data. We illustrate the applicability of our framework by building a market recommendations aggregation system. This system automatically collects market recommendations from online sources, and, based on the past performance of the analysts that

issued a recommendation, generates an aggregated recommendation in the form of a buy, hold, or sell advice. We demonstrate the flexibility of our proposed system by implementing multiple methods for the aggregation of market recommendations.

Chapter 2

Computational Content Analysis of European Central Bank Statements [1]

In this chapter we present a framework for the computational content analysis of European Central Bank (ECB) statements. Based on this framework, we provide two approaches that can be used in a practical context. Both approaches use the content of ECB statements to predict upward and downward movement in the MSCI EURO index. General Inquirer (GI) is used for the quantification of the content of the statements. In the first approach, we rely on the frequency of adjectives in the text of the ECB statements in relation to the content categories they represent. The second approach uses fuzzy grammar fragments composed of economic terms and content categories. Our results indicate that the two proposed approaches perform better than a random classifier for predicting upward or downward movement of the MSCI EURO index.

[1] An article based on this chapter has appeared in: Viorel Milea, Rui Jorge Almeida, Nurfadhlina Mohd Sharef, Uzay Kaymak, Flavius Frasincar. Computational Content Analysis of European Central Bank Statements. International Journal of Computer Information Systems and Industrial Management Applications, Volume 4, pages 628–640, MIR Labs, 2012.

2.1 Introduction

For a large part, corporate as well as government communication consists of free text. Although accessible to the human mind, such information fragments are difficult to process in an automated way. When faced with high volumes of such information, one would find it desirable to use machines for the processing, interpretation and aggregation of this knowledge. Ideally, such a process should lead to an advice in the form of a recommended decision that follows from the text that is being considered.

This issue is especially relevant for financial investors, who are often faced with high volumes of information and are under time pressure to incorporate this into their decision-making, in a time short enough to provide a competitive advantage over other market participants. The sources of information can be very diverse, ranging from formal means of communication to social media. Indeed, different studies show that European Central Bank (ECB) statements hold predictive power over financial markets (Milea et al., 2010a,b), and that the general mood on Twitter can be used in predicting upward or downward movement in the Dow Jones Industrial Average (DJIA) index (Bollen et al., 2011).

Most approaches to content and sentiment analysis in economic text are currently focused on different isolated problems rather than providing frameworks for this problem. In this chapter, we aim at providing a general framework towards the automated analysis of ECB statements, with the goal of aiding decision-makers with investment decisions. We focus the analysis on the content of fragments of text, based on the General Inquirer service (General Inquirer, 2011), which uses the Harvard-IV-4 and Lasswell content dictionaries.

We provide two approaches for our proposed framework, that can be used in a practical context. The first focuses on the frequency of content categories as encountered in text. The second one takes a more sophisticated approach in that, rather than focusing on word frequencies, it focuses on fragments of text containing both an economic term as well as a word denoting some content category. Again, the frequency of such fragments is measured in text.

The information source we choose consists of European Central Bank (ECB) communication. The statements we consider have appeared monthly starting at the end of 1998. In addition to discussing the levels of the key interest rates in the European Union, these statements provide

an overview of the economy in the past month, as well as an economic outlook for the period succeeding the issued statement. Given the importance of these statements to financial markets, the anticipation with which they are received by market participants, and the extensive information they contain, we deem these statements relevant for the price forming mechanism of European assets. Hence, we hypothesize that automated content extraction of the ECB statements can help predict a Europe-wide financial market index, such as the MSCI EURO. For modeling the levels of the index, we rely on a Fuzzy Inference System (FIS), mainly due to the interpretability of such models, which provides us with some insight into the relationship between the content of the ECB statements and the movement of the index.

We validate our framework by measuring the accuracy of the proposed approaches in terms of correctly predicting upward or downward movement of the index. We find that both approaches show a performance that exceeds the accuracy of a random classifier, thus validating our general framework fpr automated financial decision support based on economic text. While one approach gives a better accuracy on the test set, the other one helps at reducing the number of features used for prediction and gives more stable models.

The outline of the chapter is as follows. In Section 2.2 we present studies related to the extraction of content and sentiment from text. Section 2.3 presents our general framework for the content analysis of ECB statements. Two approaches based on our framework are described in Section 2.4. We present the fuzzy model that we use for the analysis in Section 2.5. The experimental setup and the results are presented in Section 2.6. Our conclusions and suggestions for further work are described in Section 2.7.

2.2 Content and Sentiment Analysis

The first attempt at content analysis in an economic context is presented in (Klingemann et al., 1982). Here, the authors investigate the relationship between a focus on wealth and wealth-related words in the speeches of the German Emperor and the state of the economy over the period 1870-1914. They find a strong relationship between the focus on wealth and the state of the German economy. More recent research, such as (Tetlock, 2007), relies on the GI dictionary

for explaining the market prices and the trading volumes. The author finds that a relationship exists between a daily Wall Street Journal column, 'Abreast of the Market', and the market prices and trading volumes of that day for the stocks discussed in the column. In (Sakaji et al., 2008) the authors develop a method for the automated extraction of basis expressions that indicate economic trends. They are able to classify positive and negative expressions which hold predictive power over economic trends, without the help of a dictionary.

Other research has focused on the extraction of sentiment from free text in an economic context. In (Zhang et al., 2009), the authors focus on eight dimensions of sentiment: joy, sadness, trust, disgust, fear, anger, surprise, and anticipation. They are able to provide visualizations of how these eight sentiments evolve over time for some concept, e.g., Iraq, based on news messages. The results are validated against ratings of human reviewers of the news messages. The method performs satisfactorily in visualizing the evolution of these sentiments.

In (Esuli and Sebastiani, 2006), the authors discuss a sentiment mining approach related to the extraction of term subjectivity and orientation from text. The approach starts with two training sets consisting of *positive* and *negative* words, respectively. It extends these two sets with WordNet synonyms and antonyms of the words in the sets. Then, a binary classifier is built by a supervised learner that is able to categorize vectorized representations of terms and classify them as *positive* or *negative*. In another approach in (Andreevskaia and Bergler, 2006) extraction of fuzzy sentiment is done, where the authors are able to assign a fuzzy membership of *positive* or *negative* to a set of words using the so-called Sentiment Tag Extraction Program (STEP).

The first approach presented in this chapter differs from the above approaches in that it relies on selected content categories from GI, and uses a fuzzy model for the prediction of movements in the MSCI EURO index. Rather than focusing on sentiment, we select a total of thirteen categories from GI and use the percentages of words that fall under those categories as document fingerprints for the individual ECB statements. By using a fuzzy model, we are able to investigate how each category impacts the index, and draw economic conclusions. Contrary to the approach in (Tetlock, 2007), we do not aggregate all content categories into one single

indicator, which would lead to losing the ability to question the impact of the different content categories on the explanandum.

In the second approach, we focus on the use of fuzzy grammars that are learnt from the text. Central to our work is the approach described in (Martin et al., 2008), where the evolution of fuzzy grammar fragments is studied for matching strings originating in free text. The basis of our approach are the methods described in (Sharef et al., 2009; Sharef and Shen, 2010) for learning and extracting such fuzzy grammar fragments from text.

2.3 Content Analysis Framework

In this section, we introduce the framework that we propose for the automated content analysis of ECB statements. An overview of the architecture that we propose is given in Figure 2.1. In the remaining part of this section, we discuss the different modules and the reasons for including them in the architecture, as well as different approaches for concretizing each of these modules.

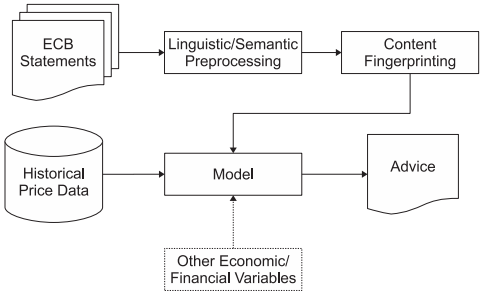

Figure 2.1: ECB statements analysis framework

Our framework consists of three main modules: the Linguisting/Semantic Preprocessing module, the Content Fingerprinting module, and the module responsible for creating the model based on historical data and the content of the text documents. We discuss each of these three modules in detail, together with the inputs they require and the output they generate.

Given a collection of ECB statements, some linguistic and/or semantic processing is required before analyzing the content of such documents. In the linguistic/semantic preprocessing

module one can envision transformations such as stemming, stopword removal, part-of-speech tagging, and/or more complex operations such as semantic analysis of the concepts presented in text, possibly based on a domain-ontology or economic thesaurus such as (STW Thesaurus for Economics, 2011). The output of this module consists of a text that is at least transformed in such a way that makes syntactic comparison across documents possible. If a semantic approach is considered, then the comparison across documents can move from the syntactic level to an analysis of how different concepts are incorporated in different documents, i.e., a comparison of the content is enabled at a deeper level.

The preprocessed documents can then be analyzed in terms of the content they present. While the previous step mainly concerns linguistic and semantic analysis, further processing of the preprocessed documents can be purely quantitative. Here, we envision the generation of content fingerprints that make a quantitative comparison between documents possible. The content fingerprints can be generated based on the frequency of (some) words and the content they denote. A more complex approach can incorporate an analysis of predefined concepts within the text, an ontology-based approach, or an analysis of recurring structures/patterns as encountered in the text documents, in terms of the link between the concepts and content that describe these patterns. The content analysis can be based on content dictionaries such as the General Inquirer (GI) (General Inquirer, 2011). The GI service provides over 180 content categories, each of them described by a set of words that fall under that category. Regardless of the approach being considered, the output of this step consists of a quantification of the content of the text document, a content fingerprint, which describes the document being considered in terms of the content it describes. This approach makes a comparison between documents possible, and, simultaneously, enables the mapping of the content fingerprints to some numerical variable that is influenced by the content of the economic texts, e.g., stock prices for some company's shares, levels of stock indexes, etc.

Provided that one has access to historical prices of an asset or the index being considered, and that these prices can be matched with the time when the economic text documents have been made public, the price variable can be modeled based on the content fingerprints of the documents being considered. In this step, one can also consider other economic/financial variables

that are relevant in the price formation of the object being considered. Different approaches can be envisioned in this step, such as Fuzzy Inference Systems (FIS) when interpretability of the model is desired, or Neural Networks (NN) for capturing the possible non-linearity of the relationship between the content fingerprints and the numerical variable being predicted without a focus on interpretability. Other approaches could also be considered.

The output of the modeling step is a forecast of the level of the asset value being considered, which can be translated to a recommendation with regard to some portfolio. For example, a price projection higher than the current price can be regarded as an advice of buying/increasing the weight of that asset within the portfolio, while a lower price projection can be interpreted as a (short) selling recommendation. In this chapter, we use the forecasts of our system for the evaluation of the approach.

2.4 Two Approaches to Computational Content Analysis

In this section we provide two approaches that show how the generalized framework that we present can be applied in a practical context. The text documents that we consider are the monthly statements of the European Central Bank (ECB). We consider these statements due to their comparable structure over the years, and the fact that they are issued regularly at predictable moments. In addition to the level of key interest rates, the ECB statements focus on the current state of the economy, and discuss likely economic developments for the short-, medium-, and long-term in the European Union (EU). Being received with much anticipation by the financial markets, we consider these statements highly relevant in the price formation of assets across the EU. Since the individual assets are affected to various extents by the considerations in ECB statements, we consider an aggregate index (MSCI EURO index) as the measure of performance for both approaches that we present in this chapter. This index is a measure of the equity market performance of developed European markets, and currently considers sixteen countries: Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and the United Kingdom. For the

quantification of the content of the ECB statements we use GI, while the model used is a FIS, chosen based on its ability to capture non-linear relationships and its interpretability.

The two approaches that we present quantify the content of ECB statements at different levels of complexity. The first approach that we consider, the Adjective Frequency Approach (AFA) presented in Section 2.4.1, looks at the frequency of adjectives within the text in relation to the content category they describe. The fingerprints thus generated are mapped onto levels of the MSCI EURO index. The second approach that we consider, the Fuzzy Grammar Fragments Approach (FGFA) presented in Section 2.4.2, looks at the frequency of grammar fragments composed of at least one economic term and a word belonging to a content category. Again, the fingerprints that we generate are mapped onto the levels of the MSCI EURO index.

2.4.1 The Adjective Frequency Approach (AFA)

In this approach, we require data from two different sources. On the one hand, we use ECB statements available from the ECB press website (ECB Press Conferences, 2011). On the other hand, we use the MSCI EURO index, available from the Thomson One Banker website (Thomson One Banker, 2011).

An ECB statement consists of different parts. The first part deals with the key ECB interest rates and their levels for the coming months. The following four parts deal with the economic and monetary analysis, as well as the fiscal policies and structural reforms. These first five parts are considered relevant for our purpose. Finally, approximately the second half of an ECB statement consists of questions and answers from the press directed towards the president of the ECB. For the current scope, we consider the Q&A part of an ECB statement relevant only indirectly, and only focus on the part describing the current and expected future state of the economy.

The relevant parts of the ECB statements for the selected period are extracted by using an HTML wrapper from the ECB press website. Upon successful extraction, each statement is annotated for parts of speech using the Stanford POS Tagger (Toutanova and Manning, 2000; Toutanova et al., 2003). Based on the part of speech annotation, we extract only the adjectives from the text. It should be noted that all ECB statements, at least in the part we consider

relevant for the current purpose, follow a similar structure. For this reason, we believe that the adjectives in the text could provide a good discrimination among the different statements.

For each ECB statement from the relevant period, the set of all adjectives contained in the text is fed to the General Inquirer web service. Based on this input, GI generates a document fingerprint consisting of the percentages of words from the document that fall under each category supported by GI. GI currently supports over 180 content categories, but for our current purpose we focus only on 13 of them, namely (General Inquirer, 2011):

- *positiv*, consisting of 1045 positive words, such as harmony, improve, and resolve;
- *negativ*, made up of 1160 negative words, such as adversity, grief, and quit;
- *strong*, consisting of 1902 words implying strength, such as apprehension, constrain, and fought;
- *weak*, containing 755 words implying weakness, such as defect, flee, and pitiful;
- *ovrst*, consisting of 696 words indicating overstatement, such as chronic, hopeless, and ridiculous;
- *undrst*, containing 319 referring to understatement, such as careful, hesitant, and light;
- *need*, made up of 76 words related to the expression of need or intent, such as famine, intent, and prefer;
- *goal*, consisting of 53 words referring to end-states towards which muscular or mental striving is directed, such as innovation, purposeful, and solution;
- *try*, containing 70 words indicating activities taken to reach a goal, such as compete, redeem, and seek;
- *means*, made up of 244 words denoting what is utilized in attaining goals, such as budget, debt, and necessity;
- *persist*, 64 words indicating endurance, such as always, invariable, and unfailing;

- *complet*, consisting of 81 words indicating the achievement of goals, such as enable, recover, and sustain;
- *fail*, which consists of 137 words that indicate that goals have not been achieved, such as bankrupt, forfeit, and ineffective.

By feeding the adjectives from each relevant ECB statement to GI, we obtain a matrix of percentages that indicate for each document, for each content category, the percentage of words in that document that fall under that category. Upon generating this matrix, we normalize it using min-max normalization across each content category.

Finally, we obtain the data on the MSCI EURO index from Thomson One Banker (T1B). We extract monthly, end-of-month data for the period January 1st 1999 until December 31st 2009. An overview of AFA is provided in Figure 2.2.

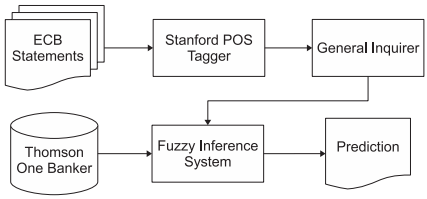

Figure 2.2: AFA steps

2.4.2 The Fuzzy Grammar Fragments Approach (FGFA)

The fuzzy grammar fragments approach focuses on the recognition and extraction of fragments from text. These fragments are defined and parsed based on a terminal grammar. In addition, the matching of text fragments to grammar fragments is achieved through a fuzzy parsing procedure.

For the purpose of extracting the fuzzy grammar fragments, we focus on a subset of 33 ECB statements. This is done in order to test the generalizability of the the approach and reduce the computational time. These statements are selected such that they are uniformly spread over the dataset: for each year from 1999 to 2009 we select 3 statements, from March, June, and

September, respectively. We use the same content categories from GI as in the AFA. These statements are then processed according to the flow in Figure 2.3.

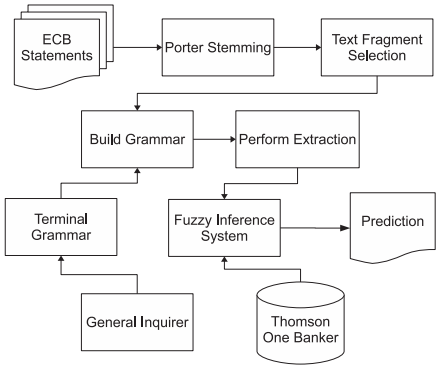

Figure 2.3: FGFA steps

Terminal grammar

For the purpose of information extraction, we begin by defining a terminal grammar around which the fuzzy grammar fragments are built. The complete terminal grammar used for the current purpose is presented in the Appendix. The terminal grammar is centered around `<EconomicTerm>` and `<ContentCategory>` as the current focus is on extracting combinations of the two from the text of the ECB statements.

Porter stemming

In order to be able to identify text fragments that are identical, one must be able to abstract beyond dissimilarities between the words and the dissimilarities that may relate to things like the tense of verb, plural vs. singular, etc. For this reason, both the terminal grammar as well as the text of the ECB statements are reduced to a root form through the Porter stemming algorithm (Porter, 1980).

Text fragment selection

The topic of interest in the current case consists of the words contained in <Economic Term>. For the purpose of building a grammar for ECB statements, text fragments consisting of 5 words preceding an economic term and 5 words succeeding an economic term are automatically selected from the text of the statements. In order to preserve the meaning of the selected text fragments, we focus only on words that are included in the same sentence. Thus, if an economic term is the first word in a sentence, no predecessors will be selected, and if an economic term is the second word in a sentence, only one word (the word preceding the economic term) will be selected as predecessor. The same applies in the case of successors. It should be noted that the text fragments that are automatically extracted have a length of maximum five, i.e., predecessors and successors are never considered together.

Building the grammar

Once all text fragments related to an economic term have been extracted, the process can proceed towards building a grammar for the ECB statements. For this purpose, all selected text fragments are transformed into grammar fragments, based on the terminal grammar presented in the Appendix. An example is presented next, where given a selected text fragment T1 and the terminal grammar as presented in the Appendix, T1 would be translated into a grammar fragment F1 as follows, with <aw> denoting any word that is not included in the terminal grammar but is present in the text fragment:

```
T1:   earli upward pressur price

F1:   <aw><PositivCat><StrongCat><EconomicTerm>
```

Once all text fragments have been translated to grammar fragments, we proceed to building the ECB statements grammar as described in (Sharef et al., 2009).

The focus of this research is on combinations of words from <Economic Term> and words from <ContentCategory>. For this reason, we only focus on fuzzy grammar fragments that

contain at least one <EconomicTerm> and at least one <ContentCategory>, regardless of the number of <aw>. For example, the fuzzy grammar fragment F1 would be selected, while F2 and F3 would be removed from the grammar:

```
F1:   <aw><PositivCat><StrongCat><EconomicTerm>
F2:   <EconomicTerm><aw><EconomicTerm>
F3:   <aw><PositivCat><aw><StrongCat>
```

Furthermore, in order to simplify the fuzzy grammar fragments we obtain, all trailing and preceding <aw> are removed from the fragments. For example, fuzzy grammar fragment F1 would become the fragment eF1:

```
 F1:   <aw><PositivCat><StrongCat><EconomicTerm>
eF1:   <PositivCat><StrongCat><EconomicTerm>
```

Finally, we group all the resulting fuzzy grammar fragments according to the <ContentCategory> they describe. When a fragment contains more than one <Content Category>, we classify this fragment under each of the <ContentCategory> elements it contains. For example, fragment F4 would be classified as strong, and fragment eF1 as both strong and positive:

```
 F4:   <StrongCat><EconomicTerm>
eF1:   <PositivCat><StrongCat><EconomicTerm>
```

This classification is important when we use the fuzzy inference system for predicting the MSCI EURO index. There, we rely on the frequencies of each group of grammar fragments, i.e., positive, negative, strong, etc., for the prediction of upward or downward movement in the index.

A final note that should be made regarding the building of the ECB grammar is that some words may fall under multiple categories, such as for the example the word *growth*, that falls both under `<EconomicTerm>`, as well as `<StrongCat>`. For this reason, we impose the following preference ordering over the grammar presented in the Appendix.

1. `<EconomicTerm>`
2. `<PositivCat>`
3. `<NegativCat>`
4. `<StrongCat>`
5. `<WeakCat>`
6. `<OvrstCat>`
7. `<UndrstCat>`
8. `<MeansCat>`
9. `<CompletCat>`
10. `<FailCat>`
11. `<NeedCat>`
12. `<PersistCat>`
13. `<GoalCat>`

Following this ordering, the word *growth*, that falls both under `<EconomicTerm>` as well as `<StrongCat>`, will be considered under `<EconomicTerm>`.

The extraction

After having built the grammar for the ECB statements, we proceed to the extraction of strings that can be parsed by the ECB grammar as described in (Sharef and Shen, 2010). The extraction from our set of documents is focused around the groups of 13 content categories as described in the Appendix. We count the number of strings that can be parsed by the grammar fragments under each category, for each ECB statement.

The output of this step consists of a matrix of frequencies of strings parsed by fuzzy grammar fragments under each of the 13 GI content categories. These frequencies are reported for each ECB statement that is available.

After the extraction process, no fuzzy grammar fragments have been found for the following content categories in combination with an `<EconomicTerm>`:

- `<Need>`;
- `<Complet>`;
- `<Fail>`.

In addition, the content category `<Goal>` is only seldomly encountered in the documents. For this reason we remove this content category from the results list. This reduces the number of content categories available for experiments to 9, namely:

- `<Means>`;
- `<Negativ>`;
- `<Ovrst>`;
- `<Persist>`;
- `<Positiv>`;
- `<Strong>`;
- `<Try>`;

- <Undrst>;
- <Weak>.

2.5 The Fuzzy Model

In this section we outline the basics of the adopted fuzzy model for the prediction of the MSCI EURO index based on the content of ECB statements.

Several techniques can be used in fuzzy identification. One possibility is to use identification by product-space clustering to approximate a nonlinear problem by decomposing it into several (Babuška, 1998; Kaymak and Babuska, 1995) subproblems. The information regarding the distribution of data can be captured by the fuzzy clusters, which can be used to identify relations between various variables regarding the modeled system.

Let us consider an n-dimensional classification problem for which N patterns $\mathbf{x}_p = (x_p^1, \ldots, x_p^n)$, $p = 1, 2, \ldots, N$ are given from κ classes $C_1, C_2, \ldots, C_\kappa$. The task of a pattern classifier is to assign a given pattern $\mathbf{x}_p$ to one of the κ possible classes based on its feature values. Thus, a classification task can be represented as a mapping $\psi : X \subset \mathbb{R}^n \to \{0,1\}^\kappa$ where $\psi(\mathbf{x}) = \mathbf{c} = (c_1, ..., c_\kappa)$ such that $c_l = 1$ and $c_j = 0$ $(j = 1, \ldots, \kappa, j \neq l)$. When the classification problem is binary, regression models can also be used as classifiers. In this approach, the regression model computes a score , e.g. probability of belonging to a class c_l, for each pattern. By applying a threshold to the score values at a suitable cutoff value, the class that a data pattern belongs to can be determined.

Takagi and Sugeno (TS) (Takagi and Sugeno, 1985) fuzzy models are suitable for identification of nonlinear systems and regression models. A TS model with affine linear consequents can be interpreted in terms of changes of the model parameters with respect to the antecedent variables, as well as in terms of local linear models of the system.

One of the most simple forms of TS models contains rules with consequents in the affine linear form:

$$R^k: \text{ If } \mathbf{x} \text{ is } A^k \text{ then } y^k = (\mathbf{a}^k)^T \mathbf{x} + b^k, \tag{2.1}$$

where R^k is the k^{th} rule, A^k is the rule antecedent, $\vec{a}^k$ is a parameter vector and b^k is a scalar offset. The consequents of the affine TS model are hyperplanes in the product space of the inputs and the output.

To form the fuzzy system model from the data set with N data samples, given by $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^T$, $Y = [y_1, y_2, \ldots, y_N]^T$ where each data sample has a dimension of n ($N >> n$), the model structure is first determined. Afterwards, the parameters of the model are identified. The number of rules characterizes the structure of a fuzzy system. For the models used in this work, the number of rules will be the same as the number of clusters. Fuzzy clustering in the Cartesian product-space $X \times Y$ is applied to partition the training data. The partitions correspond to the characteristic regions where the systems' behavior is approximated by local linear models in the multidimensional space. Given the training data X_T and the number of clusters K, a suitable clustering algorithm is applied.

In this work, we use the fuzzy c-means (FCM) (Bezdek, 1981) algorithm. As result of the clustering process, we obtain a fuzzy partition matrix $U = [\mu_p^k]$. The fuzzy sets in the antecedent of the rules are identified by means of the matrix U that have dimensions $[N \times K]$. One dimensional fuzzy sets A_i^k, $i = 1, \ldots, n$ are obtained from the multidimensional fuzzy sets by projections onto the space of the input variables x^i. This is expressed by the point-wise projection operator of the form

$$\mu_{A_i^k}(x_p^i) = \mathrm{proj}_i(\mu_p^k), \tag{2.2}$$

after which the pointwise projections are approximated by Gaussian membership functions.

When computing the degree of fulfillment $\beta^k(\mathbf{x})$ of the k-th rule, the original cluster in the antecedent product space is reconstructed by applying the intersection operator in the cartesian product space of the antecedent variables: $\beta^k(x) = \mu_{A_1^k}(x^1) \wedge \mu_{A_2^k}(x^2) \wedge \ldots \wedge \mu_{A_p^k}(x^n)$. Other t-norms, such as the product, can be used instead of the minimum operator. The consequent parameters for each rule are obtained by means of least square estimation, which concludes the identification of the classification system.

After the generation of the fuzzy system, rule base simplification and model reduction could be used (Setnes et al., 1998), but we did not consider this step in our current study.

We proceed as follows to generate the class labels. With the exception of the first observation from the dataset, all output values are set to 1 if the predicted value for the index in period $t+1$ is higher than or equal to the predicted value of the index in period t, and to 0 if the predicted value of the index is lower in period $t+1$ compared to the same value in period t. The same procedure is applied to the actual values of the index.

2.6 Experiments and Results

In this section we outline the experiments that we perform and the obtained results. After first describing the experimental setup in Section 2.6.1, we present the results of the AFA approach in Section 2.6.2. The results obtained by using the FGFA approach are described in Section 2.6.3. The section ends with a discussion of our results in Section 2.6.4.

2.6.1 Experimental Setup

The dataset we used consisted of ECB statements and monthly closing values of the MSCI EURO index in the period January 1st, 1999 to December 31st, 2010. The index data is shown in Figure 2.4. We use 70% of the data for training the model and leave the remaining 30% for testing. For the training dataset, we generate a random permutation of indexes of the data points covering 70% of the complete dataset. In this way, every run of the system will be using different, randomly selected data. We do this in order to test the accuracy of the system regardless of economic cycles, as training the system on the first 70% of the data cannot account for the economic crisis from 2008 onwards. By using multiple runs on randomly selected data points we aim at reducing this effect. Furthermore, the model is then less likely to be influenced by any trend information that may be present in the data.

We run 100 experiments, and for each experiment the data are randomly drawn again from the dataset. For all 100 experiments, we maintain 70% of the dataset for training and 30% of the dataset for testing. Although different types of fuzzy systems have been tested, the best results have been obtained with a Takagi-Sugeno fuzzy system based on fuzzy c-means clustering

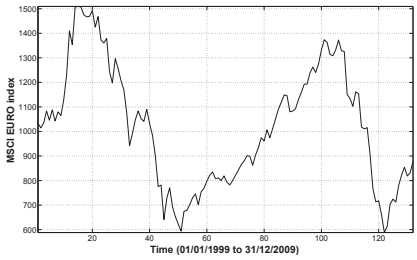

Figure 2.4: The MSCI EURO index

(Bezdek, 1981). We tried several numbers of clusters, and obtained the best results when using three clusters.

The fuzzy model is developed to predict the actual level of the MSCI EURO index in the month of the statement that is considered. In the final analysis, however, we are interested in the upwards or downwards movements of the index. Thus, the prediction of the index value by the FIS is used to determine whether the index will move up or down in the month of the respective statement. The accuracy of the fuzzy system is measured as the percentage of times that the system is able to correctly predict whether the index will move up or down. The formula for accuracy is presented in (2.3), where M^+ stands for the number of datapoints correctly predicted as upward movement, M^- stands for the number of datapoints correctly predicted as downward movement. D stands for the total number of datapoints.

$$ACC = (M^+/D + M^-/D) * 100\% \tag{2.3}$$

2.6.2 AFA Results

In Table 2.1 we present an overview of the results of 100 experiments on the data described in the previous paragraphs. For the 100 runs, for both the training set as well as the testing set, we provide an overview of the minimum, maximum, and the mean accuracy obtained. The standard deviation of the accuracy is shown between parentheses.

On the training set, the average accuracy ranges between 58.82% and 77.65%, while having a mean of 69.18% with a standard deviation of 4.01. A small standard deviation indicates

Table 2.1: AFA results of 100 experiments

	Min (%)	Max (%)	Mean (%)
Training	58.82	77.65	69.18 (4.01)
Testing	44.44	80.56	63.03 (7.88)

Table 2.2: AFA confusion matrix for 100 experiments

	True Up	True Down
Pred. Up	34.28%	16.72%
Pred. Down	20.25%	28.75%

consistent results. The average accuracy shows that in about 2/3 cases, the system is able to correctly identify an increase or decrease in the MSCI EURO index. The average accuracy goes down over the 100 experiments for the test set, but only slightly to 63.03%, indicating that some overfitting occurs. However, the standard deviation nearly doubles to 7.88, which can also be observed in the much wider range between the minimum and the maximum accuracy. Having a minimum accuracy as low as 44.44% on the test data might indicate that periods are present in the test set when the model does a very poor job at predicting the change in the index, such as when the movement of the index is solely determined by a crisis period.

In Table 2.2 we present the average confusion matrix for 100 fuzzy inference systems that we generate. The rows indicate the predicted movement direction of the index, while the columns indicate the true change in the index value.

As it can be seen from Table 2.2, a slight bias can be observed between true positives and true negatives. The system seems to be able to better predict upward movement rather than downward movement. In terms of misclassifications, the same can be stated about the false positives and the false negatives. In Table 2.2 we also show the standard deviation for all mean values between parentheses.

In Figure 2.5 we provide an overview of the FIS output surface for selected pairs of inputs. The values of the MSCI EURO index in this figure have been obtained by min-max normalization. Therefore, the values for the index range between 0 and 1. From this figure, one can notice

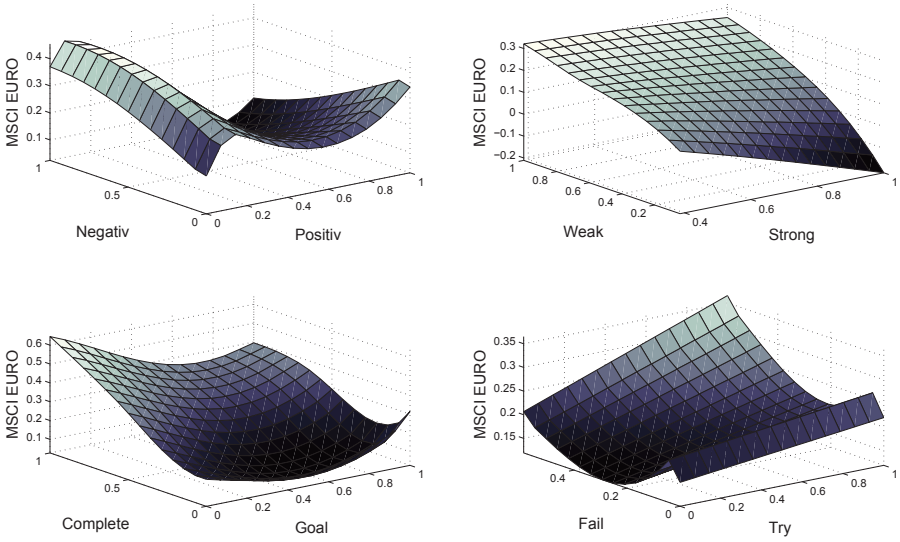

Figure 2.5: Fuzzy inference system output surface for selected pairs of inputs (AFA)

the high non-linearity of the relations, such as for example in the case of *positiv* vs. *negativ* selected inputs pair. The presence of nonlinear relations supports our choice for a fuzzy inference system to model the relationship between the content of ECB statements and the MSCI EURO index. It can also be observed that small parts of the results are sometimes counterintuitive, as in the case of the *positiv* - *negativ* plot. For example, one could observe that there is a positive correlation between *negativ* and the MSCI EURO index for very small values of the *negativ* variable. We consider this a spurious effect, and a direct result of the limited amount of data that is available for training, as well as testing, the model. For higher values of the *negativ* variable, the relation is as expected: higher values of this variable result in lower values for the index.

2.6.3 FGFA Results

In this section we report the results obtained from 100 experiments for FGFA. An overview hereof is provided in Table 2.3. For both the training and the testing set we report the minimum,

Table 2.3: FGFA results of 100 experiments

	Min (%)	Max (%)	Mean (%)
Training	52.94	70.59	61.65 (3.34)
Testing	47.22	77.78	61.36 (6.43)

Table 2.4: FGFA confusion matrix for 100 experiments

	True Up	True Down
Pred. Up	33.72%	17.64%
Pred. Down	21.00%	27.64%

maximum, and mean accuracy. Additionally, we report the standard deviation of the accuracy between parentheses.

In Table 2.4 we present the average confusion matrix for 100 fuzzy inference systems that we generate. The rows indicate the predicted movement direction of the index, while the columns indicate the true change in the index value. It can be seen that the confusion matrix obtained is similar to the confusion matrix obtained from AFA.

In Figure 2.6 we provide a few surface plots for pairs of selected inputs, for one of the fuzzy models generated by the system. All pairs of inputs are plotted against the output, which consists of the normalized levels of the MSCI EURO index. Again, we notice the non-linearity describing the relations between our content input variables and the values of the index. The results indicate that the *negativ* variable is inversely related to the values of the index, while the *positiv* category positively influences the index. The *ovrst* content variable also results in higher values for the index, when this category is present to a greater extent in the text of the ECB statements.

2.6.4 Discussion

The performance of a random classifier is expected to be roughly equal to 50% because the classes up and down movement of the MSCI EURO index are equally distributed in our dataset. For the selected period, we can conclude that both AFA, as well as FGFA, provide superior

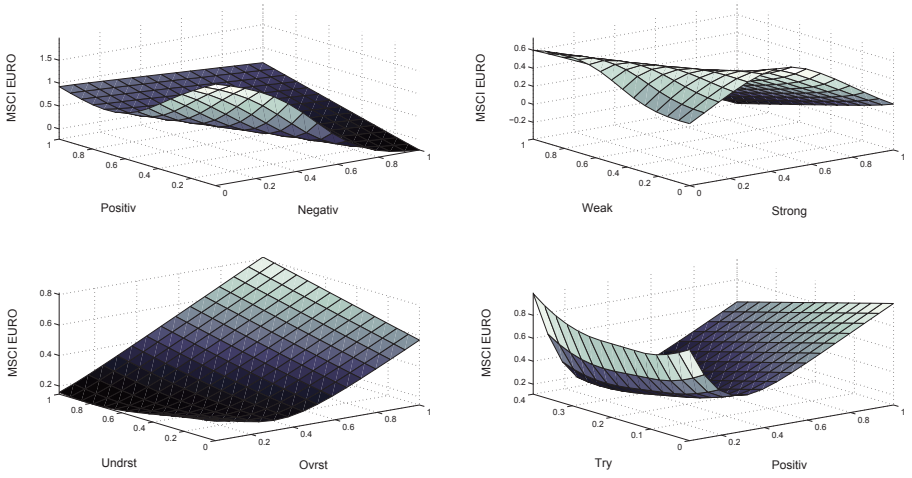

Figure 2.6: Fuzzy inference system output surface for selected pairs of inputs (FGFA)

performance, at least in terms of the mean accuracy of prediction, when compared to a random investment strategy. Hence, both approaches to computational content analysis of the ECB statements have predictive power over the MSCI EURO index. Hence, the general framework that we propose is useful for the computational content analysis of ECB statements. Such analysis can form the basis for the aggregation of multiple documents in a way that is more accessible to decision makers. In addition, such analysis can form the input to models that take such economic information into account and stand at the basis of (semi-) automated investment strategies that might be used, for example, in algorithmic trading.

Finally, we note that the relation between the content of ECB statements and the MSCI EURO index appears to be non-linear, both in the case of AFA, and in the case of FGFA. Although both approaches show this type of relation, further investigation is needed into the extent of the non-linearity before we can conclude that the contents of economic text is always non-linearly related to the variable being forecasted.

2.7 Conclusions and Future Work

In this chapter we present a general framework for the computational analysis of ECB statements. The application of this framework is illustrated by means of two concrete approaches. One is based on the frequency of adjectives in the text in relation to the content categories as outlined by GI. The other is focused on the frequency of fuzzy grammar fragments in relation to the economic terms and content categories they describe, again based on GI. The documents being considered are the monthly statements of the ECB, and they are used for the prediction of the upward or the downward movement in the MSCI EURO index. Our results indicate that, in both approaches, the movement of the index can be predicted with a higher accuracy than when a random classifier is used. We use these results to validate the ability of our proposed general framework to analyze the content ECB statements.

Note that our approach does not consider deep knowledge about the semantics of the text. It can be expected that the results will improve if the semantics of the text are taken into account explicitly. Ontology-based approaches based on state-of-the-art languages such as the Web Ontology Language (OWL) (Bechhofer et al., 2004) in static contexts or tOWL (Milea et al., 2012a,b) in time-varying contexts is an interesting direction for further research.

The next step should especially focus on relaxing the type of free text that is used for the analysis. While in the current work we use structured text due to its relatively comparable content over time, using news messages in a more general meaning can result in increased accuracy of the models.

Chapter 3

An Automated Framework for Incorporating News in Stock Trading Strategies

In this chapter we present a framework for introducing news into stock trading strategies. The news take the form of events extracted from news messages presented in free text without annotations. We test the introduced framework by deriving trading strategies based on technical indicators and impacts of the extracted events. The strategies take the form of rules that combine technical trading indicators with a news variable, and are revealed through the use of genetic programming. We find that the news variable is often included in the optimal trading rules, thus validating our proposed framework for automatically incorporating news in stock trading strategies.

3.1 Introduction

Financial markets are driven by information. One important source of information consists of news communicated by different media agencies through a variety of channels. With the increasing number of information sources, resulting in high volumes of news, manual processing of the knowledge being communicated becomes a highly difficult task. Additionally, given that this information is time-sensitive, especially in the context of financial markets, selecting and processing all the relevant information in a decision-making process, such as the decision whether to buy, hold, or sell an asset is an especially challenging task. This environment motivates a need for automation in the processing of information, to the extent that investment decisions where the news factor plays an important role can be based on an automatically generated recommendation that takes into account all news messages relevant to a certain asset.

Using information extracted from text in a financial context is an area enjoying increasing attention. In (Das and Chen, 2007) the authors extract investor sentiment from stock message boards. The prediction of bankruptcy of firms, as well as fraud, based on textual data from the Management Discussion and Analysis Sections (MD&A) of 10-K reports is investigated in (Cecchini et al., 2010). A popular Wall Street Journal column is used for investigating asset prices as well as trading volumes in (Tetlock, 2007). Financial news stories are used for the prediction of stock returns and firms' future cash flows in (Tetlock et al., 2008). Thus, the qualitative data may emerge from different sources, and can be used for the prediction of different financial aspects of firms' performance.

In this chapter we focus on information presented in textual format, taking the shape of news messages relating to some aspect of the financial market, with a particular focus on companies listed under the FTSE350 stock index. The question that we answer is how the information communicated through textual news messages can be automatically incorporated into trading strategies. This requires a two-step approach consisting of: i) extracting the relevant events, as well as the involved entities, from the text of the news messages, and ii) associating an impact with each of the extracted events.

Upon extracting the events and associating these with a predefined impact, trading rules based on news can be derived. In this chapter we only consider technical trading indicators

as part of these trading rules, but the approach can be easily extended to incorporate other indicators, such as, for example, indicators initiating from fundamental analysis. Technical trading has previously been used with the goal of financial forecasting (Leigh et al., 2002; Mehta and Bhattacharyya, 2004), thus motivating our choice for this approach. The trading strategies that are built are expressed in the form of trees, where the leaves are connected by the logical operators *and* and *or*. These trading strategies generate a buy or sell signal for the assets they are applied to, and are determined through genetic programming where a pool of possible trading strategies is tested on historical stock data.

We hypothesize that, if the proposed framework is correct, news will be included in the trading strategies generated through genetic programming. Additionally, the trading strategies that we derive in this way should generate positive returns. The first hypothesis comes from the idea that, when providing a genetic program with a pool of variables without the restriction that all these variables should be included in a trading strategy, only the variables that are maximizing the returns will be selected. Trading strategies including a news variable will thus indicate that the content of the news messages has been quantified in such a way that enables generation of profit beyond the ability of trading rules based solely on technical analysis. The second hypothesis states that, next to generating trading strategies based on news, the resulting rules should also be able to obtain a positive return.

The chapter is structured as follows. In Section 3.2 we present previous work related to the relationship between news and the stock market, and the type of events that are proven to influence stock prices. Section 3.3 provides an initial, quantitative investigation of the relationship between news messages and the stock market, as apparent from the used dataset. In Section 3.4 we discuss the technical indicators that we use for deriving stock trading strategies. We introduce our framework for automated trading based on news in Section 3.5 and discuss the results of validating the framework in Section 3.6. We conclude in Section 3.7 where we also give directions for future work.

3.2 Previous Work on the Relationship Between News and the Stock Market

In this section we provide an overview of work related to the relationship between news and the stock market. We focus on 3 key aspects: i) there is evidence that a relationship exists between news announcements and financial markets, ii) the impact of events on financial markets can be quantified, and a list of relevant events can be identified, and iii) the relationship between information in the form of news and financial markets is not a trivial one. One aspect left aside in this overview relates to mining the text of news messages for assessing market response. For a recent survey of the different methods used for this purpose, we refer the reader to (Mittermayer and Knolmayer, 2006).

In (Mitchell and Mulherin, 1994) the relation between the number of news announcements and trading activity is investigated. The research is focused on whether the amount of information that is publicly reported affects the trading activity and the price movements in the stock market. Here, information is defined and quantified by the number of daily announcements released by the Dow Jones & Company newswire. All news messages are assigned equal importance, regardless of the type of event being described. The following statement is tested: a greater number of news announcements induces greater trading volume and price variability. The results indicate a statistically significant positive correlation between the number of daily news announcements and trading activity. If the number of announcements increases with 100%, the trading activity will increase with 38%. The relation becomes stronger only if those news messages are selected that, besides being published through the wire service, are published in a newspaper the next morning.

The relation between news announcements and monthly returns is also investigated in (Chan, 2003). Several stocks are selected with at least one news story in a certain month. The news messages are divided into "news winners" (price increased after announcement) and "news losers" (price decreased after announcement). The abnormal returns are measured for 36 months after the month when the news was published. The results are compared to a group "no news"

containing those companies which had no news in a certain month. The authors conclude that stocks exhibit abnormal returns after public news.

The effect of analyst recommendations, with a focus on buy advices, is studied in (Kim et al., 1997). First, the authors test whether the advice issued especially for clients (before the opening of the stock market) contains information and then perform the same test but following the official release of the advice (to the main public). The authors find a strong relation between an initial coverage with a buy recommendation and a reaction in the stock market.

The relation between earnings announcements and trading volume around the announcement date is investigated in (Ewalds et al., 2000). The research is focused on stocks from the AEX exchange from 1994 to 1999. In all selected days around the earnings announcement a significant positive increase in trading volume is found. The increase in trading activity is the largest at the announcement date. The robustness of the relation is checked by dividing the companies into two categories, namely small and large. Both categories have a significant relation with trading activity, but the relation is much stronger regarding small companies compared to large companies. A possible explanation is that there is less information available about small companies. Another relation was found between the date of the announcement and trading volume. The longer a company waits with revealing the earnings, the smaller the change in trading volume. A possible explanation is that the expectations are more accurate in that case, i.e., analysts have more time and information (earnings from competitors) to accurately predict earnings.

Up until now, different meanings have been assigned to the word news when studying the relation to the stock market. The used news sources are arbitrary, they contain different news messages, and are not complete. Although a relation is apparent, it is necessary to zoom into real life events and quantify the relation between these events and returns. The list of events that possibly affect the stock price is extremely large, but in the remaining paragraphs we focus on a limited number hereof, considered to be of increased relevance in financial markets.

A management change event is defined in (Warner et al., 1988) as a change in the set of individuals holding the title Chief Executive Officer (CEO), president, or chairman of the board. The reaction after a change in management can indicate whether the market considers this event as important. A stock return after a management change contains:

- The information effect (negative): the management performance is worse than expected by the market. Shareholders did not realize the company performed this bad so the news is received negatively.
- The real effect (positive): the change is in shareholders' interest. If a company performs very bad, a management change could mean a new vision, strategy, etc. so the expectations about the companies' future results could be revised. The news is received positively.

The authors did not find a general relation between a management change and abnormal stock returns. Only on the day of the announcement a statistically significant price movement was noticed, but the direction was both positive and negative.

The real effect of a management change is studied in (Bonnier and Bruner, 1989). The authors find that the average excess return from day -1 (the day before announcement) until day 0 is 2.479% (positive significant). They also found that the power of the title, the size of the company and the type of the manager have a positive, significant impact on the real effect. Also, in general, a management change conveys bad news about the performance of the company. A management change is received positively if the company performance is bad.

The trading activity and price movements before, on, and after the day of a merger announcement are studied in (Keown and Pinkerton, 1981). A significant increase in trading activity can be found before, on, and after the merger announcement. In 79% of the acquired firms a significant increase in trading volume is noticed one week before the announcement, compared to 3 months before that date. Approximately half of the reactions occur before the official public announcement - they start one month before the merger. The strongest reaction in 1 day is on the announcement day itself: the market reacts immediately.

The price momentum following a merger announcement is investigated in (Rosen, 2006), and relates to an initial response of the market to a merger announcement and its propagation through time, i.e., if the initial reaction is positive, then it will tend to continue. The results indicate that if a company had successful mergers in the past, then this will positively influence momentum.

Stock splits and their effect on the price are investigated in (Ikenberry and Ramnath, 2002). A NYSE sample from 1988 until 1997 is used with over 3000 stock splits. The authors find a

9% positive difference of abnormal returns between the split stocks and a control group, a year after the split.

The price reaction after dividend initiations and omissions is investigated in (Michaely et al., 1995). The short term (3 days) and the long term (several years) reactions are investigated. A buy and hold strategy is used to measure the returns after certain periods. In the three days around the initiation announcement a significant excess return of 3.4% is identified. In the year before, the excess return is 15.1%. The companies with a dividend omission perform very poor in the year before the announcement, apparent from an excess return of -31.8%. Around the announcement, an excess return of -3.1% is identified. These trends continue also in the next 1 year and the next 3 years after the announcement.

These findings come to support our assumption that events that can be identified in news messages have a significant impact on stock prices and trading volumes. For this reason, we consider it worthwhile to use such events in our analysis. In the final part of this section we focus on different properties of (public) information in the context of financial markets.

The degree of uncertainty of information is explored in (Zhang, 2006). The hypothesis is that greater information uncertainty will lead to higher expected stock returns after good news and lower expected returns after bad news. This implication is based on results from behavioural finance studies, i.e., psychological biases such as overconfidence are increased when there is more uncertainty. News is defined in this research as an analyst forecast revision. Good news is an upward revision and bad news is a downward revision. The authors find evidence that the market reaction directly after the news announcement is not complete. This means that bad news implies relatively lower future returns and good news predicts higher future returns.

An investigation into whether certain forms of rumours have influence on trading activity on the stock market is performed in (Bommel, 2003). To test the possible influence, a dynamic model is presented with three kinds of rumours: honest, bluffing, and cheating. The conclusion is that spreading rumours makes economic sense. It increases the demand for stocks and it can drive the price above the price the rumourmonger knows with his private information. If the followers know a certain trader cheated with a false rumour, they will not take notice of his rumours and causing the rumourmonger to lose his reputation.

It should be noted that our approach does not focus on representing and reasoning with complex knowledge contained in the news messages, but rather focuses on single events. One possible (future) direction could consist of designing an ontology of events, in Resource Description Framework (RDF) or RDF Schema (RDFS) (Klyne and Carroll, 2004; Brickley and Guha, 2004), or the Web Ontology Language (OWL) (Patel-Schneider et al., 2004) if the context is static, although a temporal web ontology language such as tOWL (Milea et al., 2008a; Frasincar et al., 2010) should be more suited for representing and reasoning with the facts exposed by the news messages. However, as we focus on single events extracted from news, we do not rely on an approach based on ontologies.

Similar work regarding the extraction of optimal trading rules based on technical indicators related to price is presented in (Allen and Karjalainen, 1999). However, unlike the current research, news are not used for determining these optimal trading strategies.

The approach we take in this chapter is novel, in that it does not focus on a particular type of event, but rather on a whole thesaurus of events that, as shown in this section, play a significant role in financial markets. Rather than focussing on the volume of news, we extract the relevant events from the market announcements and attempt to include them in trading rules. For this purpose, we use a genetic algorithm that can choose between different variables in creating profitable trading rules. The variables are all originating in technical analysis, with the exception of the news-related variable.

3.3 A Preliminary Analysis of the Relationship Between News and the Stock Market

In this section we provide a preliminary analysis of the relationship between news and the stock market, as apparent from the dataset we have collected. The analysis is focussed on discovering the influence that news have on the share price of the concerned companies, as well as on whether this influence can be captured through the extraction of events from news messages and using a predefined impact for determining the direction of this influence on prices. We start off by describing the process of extracting events from the news messages in Section 3.3.1. Descriptive

statistics of the dataset used are presented in Section 3.3.2. Finally, the relationship between events and share prices of the companies involved in these events is presented in Section 3.3.3.

3.3.1 Event Information Extraction

The event information extraction from the news messages is based on recognizing a predefined set of events as well as the affiliated entities. For this we rely on the ViewerPro tool[1], a proprietary application able to extract events from text-based data.

ViewerPro is an application created by SemLab that enables the identification of events in news messages. These events can be used to identify an impact of a news item on an equity. ViewerPro turns enormous amounts of unstructured news into structured trading information. Once the unstructured news information is fed in the ViewerPro system, it undergoes several kinds of (proprietary) processing steps in order to filter out unwanted information and select solely that which is relevant. These steps include metadata filtering, parsing, gazetteering, stemming, natural language processing, and automatic pattern matching. Large numbers of news messages are filtered for equity-specific news and the semantic analysis system of ViewerPro interprets the impact of every individual news message. We provide a couple of examples of the type of information that can be extracted from a news message by using ViewerPro.

Example 1

A sentence in a news message states: "Nokia, scrambling to revise its handset strategy and reverse deep declines in its core business, may announce as soon as Friday a partnership with Microsoft to use the Windows Phone 7 system on its mobile phones." Given this piece of text, ViewerPro is able to recognize the event 'Company collaboration consideration' and associate it to equity 'Nokia'.

Example 2

A sentence in a news message states: "Arena Leisure PLC (ARE.LN), the operator of horse-racing fixtures, said Thursday it has successfully tendered to The London Organizing Committee

[1] `http://viewerpro.semlab.nl/pages/p.view?id=212`

of The Olympics Games and Paralympic Games Limited, or LOCOG, and has been awarded a contract to provide catering services at both the Eton Dorney and Greenwich Park venues as part of the Olympic and Paralympic Games in 2012." From this sentence, ViewerPro extracts the event 'Company contract win' for equity 'Arena Leisure PLC'.

3.3.2 Descriptive Statistics of the Dataset

The dataset we use consists of a database of historical company share prices as well as a collection of news messages related to these companies. The company dataset consists of all firms included in the FTSE350 stock index at August 1st, 2008. The news dataset is collected through the Reuters news feed, and concerns all 350 companies listed under FTSE350. Both datasets cover the period January 1st, 2007 until April 30th, 2007. The news dataset provides a set of 5157 events. However, only a subset hereof is used for our study. The selection of these relevant events is based on three criteria:

- News articles issued on days when the stock exchange is closed are not considered;
- Duplicate events are not considered;
- Rarely occurring events ($< 0.5\%$ of all events) are not considered.

Not including the articles issued on days when the stock exchange is closed relates to the fact that the events contained in these messages will not have an immediately quantifiable impact on the stock price. Since several events can occur during the period when the stock exchange is closed, associating these events with the changes in price over this period will introduce additional variance with regard to which event precisely influences the change in price. Thus, in ordered to limit this aggregate effect, those news messages are excluded from the analysis.

At times, news messages may be repeated for the purpose of providing updates on an event described in a previous message. This results in storing, in our event database, events that are on the same day, concern the same company, and are identical to another event previously extracted on the same day. Since these news messages describe the same events, it suffices to only consider them once and thus incorporate the associated impact for the event in the stock price projection only once.

Infrequently occurring events, i.e., events occurring in less than 0.5% of the news messages, are removed from the dataset due to the fact that considering them would negatively influence the statistical validity of our conclusions. Moreover, the impact of such isolated events is difficult to asses with confidence.

Upon considering these four criteria on the event dataset we have collected, the original sample of 5157 events is reduced to 2112. An overview of these events, as well as their frequencies in the event dataset, is presented in Figure 3.1.

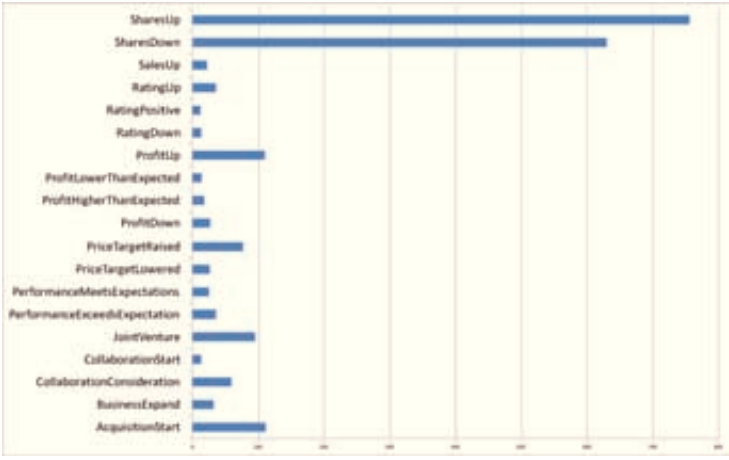

Figure 3.1: Frequency of events in the dataset

3.3.3 Relationship Between News and Share Prices

The real impact of news on stock prices is assessed by means of relative returns, based on end-of-day data, i.e., closing prices. For a single asset, a return is computed as:

$$r_i = \frac{P_{i+n} - P_i}{P_i} * 100 \tag{3.1}$$

where i represents the day before the event and n represents the number of days over which the return is calculated, with $n > 0$.

In case multiple events occur across the same day, regarding the same asset, the return is corrected for the number of events, as follows:

$$R_i = \frac{\sum_{j=1}^{N} r_j}{N} \tag{3.2}$$

where N is the number of events.

In order to correct the returns for the general market sentiment, we focus on excess returns. The excess return is calculated as the individual return of an asset that is achieved in excess of the market return, i.e., the return of the main index in which the asset is included. This is computed as follows, where r_i^I denotes the return of the index used as benchmark:

$$a_i = r_i - r_i^I \tag{3.3}$$

When dealing with multiple events and excess returns, we correct these returns for the number of events as follows, where N is the number of events:

$$A_i = \frac{\sum_{i=1}^{N} a_i}{N} \tag{3.4}$$

For the results presented in this section, the benchmark index used to compute excess returns is the FTSE350 index. An overview of the results obtained for each of the events included in the analysis, in terms of absolute returns, as well as excess returns, is presented in Tables 3.1 and 3.2. For each event we compute the returns for the day of the event, denoted R_0 in the case of absolute returns, or A_0 in the case of excess returns, as well as the returns following the event one, two, five, and ten days after the event is made public. For each of the events, we compute the percentage of events for which the direction of the return corresponds to the sign of the impact, i.e., positive returns in the case of positive impacts and negative returns in the case of negative impacts, and we denote this by d. Additionally, we compute the two-tailed significance for the returns obtained for each of the events, and report this as s. It should be noted that the impacts reported in Tables 3.1 and 3.2 have been manually determined by experts. Finally, three remarks are in place when interpreting these results:

- Multiple events may determine asset prices, and thus the returns reported in the following tables, while not all these events could be captured through the news messages used for the analysis. However, we assume that the largest share of the reported returns is captured by the reported events;

- Reactions to events, in terms of price changes, may initiate before the actual event is public. By relying on the closing price of the asset on the day previous to the event, we incorporate most of the anticipation preceding an event;

- When manually assessing the impact of an event type on stock prices (reported under the *impact* column in Tables 3.1 and 3.2) the assumption is made that no other interactions, involving, for example, other events, have a significant influence on the price.

Table 3.1: Average returns for different time intervals after an event

Event	Impact	Count	R_0	d	s	R_1	d	s	R_2	d	s	R_5	d	s	R_{10}	d	s
SharesUp	2	756	1.63	85	0.00	1.65	80	0.00	1.53	73	0.00	1.74	72	0.00	2.27	69	0.00
RatingUp	2	36	1.55	83	0.00	1.82	72	0.00	1.89	75	0.00	2.52	69	0.00	2.31	72	0.01
CollaborationStart	2	13	1.06	46	0.32	1.73	62	0.18	1.95	62	0.14	1.95	77	0.10	1.30	62	0.20
RatingPositive	1	12	0.84	75	0.25	0.96	58	0.32	1.51	75	0.28	3.23	75	0.02	4.26	92	0.02
ProfitHigherThanExpected	3	18	0.74	50	0.39	1.99	61	0.11	1.72	61	0.19	2.87	56	0.08	2.61	61	0.12
AcquisitionStart	3	111	0.40	62	0.08	0.56	59	0.06	0.53	55	0.12	0.90	59	0.02	0.90	62	0.06
SalesUp	2	22	0.33	64	0.44	0.56	55	0.37	0.35	55	0.65	1.17	64	0.29	1.44	59	0.23
PriceTargetRaised	2	77	0.24	51	0.31	0.54	56	0.13	0.65	58	0.11	0.83	56	0.08	2.35	71	0.00
BusinessExpand	1	32	0.21	53	0.70	0.83	56	0.16	0.69	56	0.34	1.52	66	0.08	2.13	72	0.02
JointVenture	1	95	0.18	53	0.31	0.12	47	0.59	0.10	52	0.72	0.49	56	0.14	0.78	53	0.05
PerformanceExceedsExpectation	3	36	0.11	58	0.82	0.35	50	0.52	0.25	47	0.70	1.49	53	0.11	0.58	56	0.53
ProfitUp	2	110	0.08	50	0.80	0.22	50	0.53	0.26	52	0.54	1.25	56	0.03	1.73	64	0.01
CollaborationConsideration	1	59	-0.05	49	0.81	-0.08	54	0.83	-0.17	49	0.69	0.19	59	0.67	0.11	59	0.85
PerformanceMeetsExpectations	1	25	-0.21	40	0.66	-0.11	48	0.79	0.31	56	0.57	0.55	64	0.46	0.75	56	0.42
ProfitDown	-2	27	-0.94	48	0.21	-0.52	52	0.49	0.05	48	0.95	0.33	52	0.78	0.83	52	0.56
RatingDown	-2	13	-0.96	54	0.15	-0.98	62	0.14	-1.32	69	0.11	-0.65	54	0.54	0.04	46	0.98
PriceTargetLowered	-2	26	-1.17	73	0.16	-1.44	77	0.12	-1.77	73	0.07	-1.50	65	0.14	-1.51	50	0.20
SharesDown	-2	630	-1.38	81	0.00	-1.49	71	0.00	-1.44	69	0.00	-1.20	62	0.00	-0.91	59	0.00
ProfitLowerThanExpected	-3	14	-2.52	64	0.06	-2.33	71	0.04	-2.04	71	0.08	-1.38	57	0.40	-1.29	64	0.35
		2112		60			60			61			62			62	

Impact represents a predefined predicting factor for future price movement.
Count represents the frequency of the event in the dataset.
R_x represents the average return after x days.
d represents the percentage of returns which went into the right direction (positive if impact is positive).
s represents the two-tailed significance at the 95% level.

Table 3.2: Average abnormal returns after x days of an event

Event	Impact	Count	A_0	d	s	A_1	d	s	A_2	d	s	A_5	d	s	A_{10}	d	s
RatingUp	2	36	1.49	81	0.00	1.88	72	0.00	1.89	72	0.00	2.46	69	0.00	2.02	69	0.03
SharesUp	2	756	1.32	80	0.00	1.31	74	0.00	1.26	68	0.00	1.59	68	0.00	1.58	64	0.00
CollaborationStart	2	13	1.02	46	0.26	1.62	69	0.15	1.76	62	0.12	1.74	77	0.10	0.83	62	0.23
ProfitHigherThanExpected	3	18	0.98	67	0.24	1.90	61	0.13	1.52	61	0.23	2.38	67	0.12	1.48	50	0.35
RatingPositive	1	12	0.83	67	0.26	1.21	67	0.17	1.69	75	0.11	2.88	75	0.03	3.35	67	0.03
AcquisitionStart	3	111	0.44	60	0.03	0.59	56	0.03	0.59	52	0.05	0.61	50	0.10	0.49	49	0.25
PerformanceExceedsExpectation	3	36	0.37	53	0.33	0.46	61	0.29	0.44	53	0.38	1.15	53	0.17	0.22	47	0.79
PriceTargetRaised	2	77	0.30	52	0.15	0.69	53	0.04	0.85	58	0.02	0.85	49	0.05	1.70	61	0.00
ProfitUp	2	110	0.19	46	0.48	0.43	55	0.15	0.45	55	0.17	1.21	55	0.01	1.44	56	0.01
SalesUp	2	22	0.16	59	0.69	0.39	59	0.47	0.26	55	0.72	1.04	55	0.27	0.94	59	0.35
BusinessExpand	1	32	0.15	38	0.74	0.66	56	0.23	0.46	50	0.46	1.17	53	0.10	1.11	53	0.18
JointVenture	1	95	0.14	49	0.36	0.18	54	0.39	0.13	51	0.58	0.13	49	0.65	-0.02	44	0.96
CollaborationConsideration	1	59	-0.14	44	0.48	-0.10	53	0.75	-0.20	49	0.57	-0.24	58	0.49	-0.63	41	0.16
PerformanceMeetsExpectations	1	25	-0.21	40	0.64	-0.32	40	0.47	-0.01	44	0.99	-0.27	48	0.67	-0.58	52	0.51
PriceTargetLowered	-2	26	-0.94	73	0.24	-1.36	85	0.16	-1.59	73	0.12	-2.24	73	0.04	-2.34	58	0.07
RatingDown	-2	13	-1.11	69	0.08	-0.91	69	0.13	-1.10	69	0.20	-0.40	46	0.71	-1.06	54	0.42
ProfitDown	-2	27	-1.14	63	0.14	-0.75	52	0.31	-0.19	52	0.82	0.22	63	0.83	0.12	63	0.93
SharesDown	-2	630	-1.18	80	0.00	-1.30	73	0.00	-1.30	73	0.00	-1.49	68	0.00	-1.46	64	0.00
ProfitLowerThanExpected	-3	14	-2.42	64	0.09	-2.42	71	0.04	-2.31	79	0.05	-1.36	71	0.30	-2.19	79	0.12
		2112		59			62			61			60			57	

Impact represents a predefined predicting factor for future price movement.
Count represents the frequency of the event in the dataset.
R_x represents the average abnormal return after x days.
d represents the percentage of abnormal returns which went into the right direction (positive if impact is positive).
s represents the two-tailed significance at the 95% level.

An initial inspection of Table 3.1 reveals that in nearly 90% of the types of events, the direction of the R_0 returns corresponds with the sign of the impact assessed by experts. The two events where this is not the case are the *collaboration consideration* event and the *performance meets expectations* event. However, the expert impact associated with these events is only slightly positive, while the generated returns are slightly negative. Thus, based on the small number of events on which this impact is assessed, as well as on the assumptions listed in the previous paragraph, and the small difference between the expert impact and the generated returns, we consider the impact assigned by experts to be trustworthy in the absence of additional data. Finally, it should be noted that the slightly negative returns are not significant at the 95% level.

Considering R_0, the event that generates the highest return is the *shares up* event, producing an average of 1.63%, backed up by the fact that 85% of this type of events generated a positive return. Presumably, not all events in this category generate a positive return due to the fact that, in some occasions, this event co-occurs with another event that generates a decrease in price that dominates the increase associated with the *shares up* event.

In the short run, i.e., when considering the R_0, R_1, and R_2 returns, we find three events for which the returns are both statistically significant, as well as showing the same direction as the impact determined by experts, in all three cases: *shares up, shares down*, and *rating up.*

In the long run, i.e., when considering the R_5 and R_{10} returns, we find more events for which both returns are statistically significant as well as being correctly captured by the manually determined impacts: *shares up, rating up, rating positive, profit up,* and *shares down.* From this we conclude that, for most events, the impacts are observable at longer time intervals after the event is reported.

When excess returns are considered, the short run exhibits four events for which the returns are significant at the 95% level and the direction of the return corresponds with the sign of the impact: *rating up, shares up, acquisition start,* and *shares down.* For the long run, the same is found for the events: *rating up, shares up, rating positive, price target raised, profit up,* and *shares down.*

The linear relationship between the predefined impact of events and the generated absolute returns is quantified by means of Person's correlation, and reported in Table 3.3. For all time intervals, we find a significant, positive correlation between the returns and impacts, respectively.

Table 3.3: Pearson's correlation of impact and return

Return	Correlation	Sign.
R_0	.844	.000
R_1	.851	.000
R_2	.804	.000
R_5	.789	.000
R_{10}	.662	.002

In Table 3.4 we report the values for Pearson's correlation for the predefined impact and abnormal returns. Again, for all time horizons, we find strong positive correlations that are significant at the 99% level. We note that in the case of excess returns, the values for Pearson's correlation are higher than in the case of absolute returns, indicating that correcting for the index leads to results that are more in line with the expectations.

Table 3.4: Pearson's correlation between impact and abnormal return

Abnormal Return	Correlation	Sign.
A_0	.878	.000
A_1	.863	.000
A_2	.827	.000
A_5	.753	.000
A_{10}	.712	.001

Based on the results presented in this section, two main conclusions can be drawn. First, the events that are selected and extracted from the text of the news messages can be used in trading strategies, as in most cases these events provide the ability to generate positive returns. Second, the predefined impact associated with the extracted events is a good reflection of the actual impact of these events on stock prices, as apparent from the Pearson correlation test presented in this section.

3.4 Technical Trading

The focus of this section is on presenting the technical trading indicators that are used in the trading strategies generated through genetic programming. The indicators including in the study are: the simple moving average (SMA), the Bollinger band (BB), the exponential moving average (EMA), the rate of change (RoC), momentum (MOM), and moving average convergence divergence (MACD). The choice for these indicators is based on their widespread use in trading.

3.4.1 Simple Moving Average and Bollinger Bands

The SMA averages the last 20 days of the price of a stock (Achelis, 2000), and is computed as:

$$M_i = \frac{\sum_{i=1}^{N} P_i}{N} \tag{3.5}$$

where P_i represents the price on day i. The average is calculated over a fixed period of 20 days prior to the day for which the average is calculated, i.e., $N = 20$, which is standard for this indicator. A buy signal is generated when the price crosses the moving average in an upward movement, while a sell signal is generated when the price crosses the moving average in a downward movement.

The Bollinger band is a technical indicator which creates two "bands" around a moving average (Achelis, 2000). These bands are based on the standard deviation of the price. It is assumed that the price will move within these bands, around the moving average. If the volatility is high, the bands are wide and when there is not much volatility the bands are narrow. The lower and upper bands of the Bollinger band can be calculated as, respectively:

$$L = M - 2 * \sigma_M \tag{3.6}$$

$$U = M + 2 * \sigma_M \tag{3.7}$$

where σ_M stands for the volatility of the moving average M.

A buy signal is generated when the price is below the lower band, which is regarded as an oversold situation. A sell signal is generated at an overbought situation, when the price is above the upper band.

3.4.2 Exponential Moving Average

The exponential moving average (EMA) aims to identify trends by using a short and a long term average (Achelis, 2000). When the averages cross each other, it is the start of a new trend. The short term average is set at 5 days and the long term average at 20 days. The EMA is computed as:

$$E_i = \frac{2}{N+1} * (P_i - E_{i-1}) + E_{i-1} \tag{3.8}$$

where P_i represents the price on day i, and N is the number of days. The initial EMA is calculated using the SMA, in our case for 5 and 20 days respectively starting from the first observation, as previously described.

When the short term average crosses the long term average upwards, a buy signal is generated. A sell signal is generated when the short term average crosses the long term average downwards.

3.4.3 Rate of Change

The rate of change (RoC) is an indicator that calculates the difference between the closing price of the current day and the closing price 10 days ago (Achelis, 2000), according to the following equation:

$$C_i = \frac{P_i - P_{i-10}}{P_{i-10}} \tag{3.9}$$

where P represents the closing price, i the moment in time.

If the RoC starts decreasing above 0 (a peak was reached), a sell signal is generated. If it starts increasing below 0, a buy signal is generated.

3.4.4 Momentum

The momentum indicator uses exactly the same formula as the RoC. Instead of creating a buy signal after a peak, it creates a buy signal when the momentum crosses the 0 level upwards (Achelis, 2000). A sell signal is generated when the RoC crosses the 0 level downwards.

3.4.5 Moving Average Convergence Divergence

The moving average convergence divergence (MACD) is a technical indicator that extracts two exponential averages from each other, namely the 12 and the 26 day exponential average (Achelis, 2000).

The mathematical formula for the MACD is:

$$D_i = E[12]P_i - E[26]P_i \tag{3.10}$$

where P represents the closing price at day i. A buy signal is generated when the MACD reaches the 0 level in an upward motion. A sell signal is generated when the MACD breaks through the 0 level in a downward motion.

3.4.6 Performance of Technical Trading Indicators

In this section we look at how the technical indicators perform on our dataset when considered separately from any other indicators, such as news. In Table 3.5, we present the returns generated by each technical indicator over different time intervals. The focus of this table is on the buy signals generated by the indicators. The frequency shows how many buy signals are generated by the indicator. The returns show the average return surrounding a buy signal in the given time frame, i.e., R_{10} represents the 10 day return. The best performing technical indicator is the simple moving average, followed by the exponential moving average.

Table 3.6 displays the performance of the technical indicators when sell signals are considered. Note that negative returns in this table are desirable, since higher magnitudes of a negative

Table 3.5: Returns, buy signals generated by technical indicators

	Frequency	R_0	R_1	R_2	R_5	R_{10}
SMA (20)	1,663	**1.927**	**2.069**	**2.197**	**2.463**	**3.512**
Rate of Change (10)	4,387	-0.417	-0.290	-0.245	-0.053	**0.245**
Momentum	1,988	**1.499**	**1.444**	**1.637**	**1.938**	**2.759**
Bollinger Band	2,014	-1.608	-1.374	-1.170	**0.110**	**0.180**
EMA (5, 20)	870	**1.717**	**1.860**	**1.859**	**2.122**	**3.350**
MACD (12, 26)	667	**1.310**	**1.324**	**1.309**	**1.568**	**2.882**

Table 3.6: Returns, sell signals generated by technical indicators

	Frequency	R_0	R_1	R_2	R_5	R_{10}
SMA (20)	1,737	**-1.888**	**-1.932**	**-1.923**	**-1.496**	**-0.805**
Rate of Change (10)	2,937	0.723	0.637	0.953	1.564	2.370
Momentum	2,049	**-1.310**	**-1.151**	**-1.248**	**-1.021**	**-0.629**
Bollinger Band	2,971	1.563	1.530	1.499	1.595	1.827
EMA (5, 20)	922	**-1.662**	**-1.654**	**-1.586**	**-0.933**	**-0.488**
MACD (12, 26)	581	**-1.276**	**-1.106**	**-1.162**	**-0.106**	0.317

return indicate better performance of the sell signal. Again, the two best performing technical indicators are the simple moving average and the exponential moving average.

3.5 A Framework for Trading Based on News

In this section we introduce a framework for incorporating news in stock trading strategies. The framework presented here assumes that events have been extracted from news messages and are available together with the date on which the events took place. Additionally, a predefined impact should be assigned to each event, allowing the news variable to be included in the trading strategies. For deriving the optimal trading strategies, we rely on genetic programming.

Genetic programming (Koza, 1992) is a technique based on genetic algorithms where the potential solutions are represented as computer programs rather than numerical values encoded in some manner. Starting from a (usually randomly generated) initial population, genetic programs attempt to improve the fitness of individuals over successive generations through a process inspired by natural evolution. During this process, individuals are altered, usually based on their fitness values, by combining them with other individuals (crossover), or by slightly altering some

parts of the individual with a predefined probability (mutation). In this chapter, genetic programming is used for finding optimal trading strategies based on technical indicators and news. Genetic programming has previously been used in the design of decision support systems, see for example (Zhao, 2007; Fan et al., 2006).

The trading strategies we determine take the form of trees that, when evaluated, return a Boolean value: true, when a trading signal is generated, or false, when no signal is generated and thus no action has to be taken. The trading strategies include at least one technical indicator or a news variable. Most often, the trading strategies include multiple variables, that may be either technical indicators or the news variable, connected by the logical operators *and*, and *or*. An example of a trading strategy that may be generated is given in Figure 3.2. This rule would generate a trading signal when the short moving average generates a trading signal simultaneously with at least one of the exponential moving average and rate of change indicators.

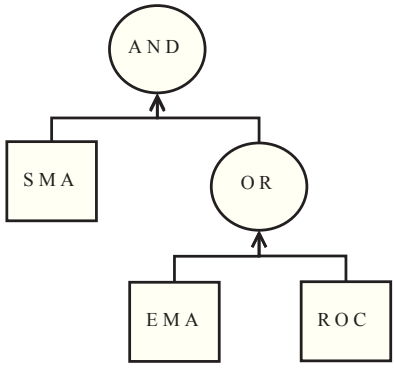

Figure 3.2: Trading rule

The fitness of a trading strategy is computed based on the return that it generates on the dataset that we use. Thus, for each day when the strategy generates a signal, we compute the n-day return as:

$$q_i = \frac{P_{i+n} - P_i}{P_i} * 100, \tag{3.11}$$

where P_{i+n} represents the price n days after day i, and P_i represents the price on day i.

Multiple days can be taken into consideration when computing the fitness, which is achieved by computing the following equation:

$$f = \frac{\sum_{i=1}^{N} q_i}{N}, \tag{3.12}$$

where N represents the number of consecutive days after the generation of the signal that are taken into consideration. Thus, when N equals 2, the day following the signal and the following day are taken into consideration. The result is an average over the number of consecutive days that are considered, i.e., N.

Since most rules generate several signals across the dataset, we obtain the fitness of a trading strategy by averaging the returns of all signals issued by the strategy, as follows:

$$F = \frac{\sum_{i=1}^{Z} f}{Z} \tag{3.13}$$

with Z denoting the number of signals generated by the trading strategy.

The genetic program we use for determining the optimal trading strategies starts from a random initial population of trees, and generates new populations of trading strategies by applying crossover and mutation on the population from the previous iteration. Crossover consists of selecting two trading strategies, and determining a random crossover point, i.e., one of the nodes of the tree. Next, the subtrees generated under the crossover point are exchanged between the two trading strategies, thus resulting in two new rules that are added to the new population. Mutation only relates to the technical indicators included in a trading strategy, and consists of a slight change in the parameters of the randomly selected technical indicator, e.g., changing the number of days used by the simple moving average from 5 to 7.

The genetic programming algorithm that we use is presented in Algorithm 1. The end condition for the algorithm relates to the improvement in the best solution found, i.e., when the optimal solution cannot be improved in a number of generations, the algorithm stops.

Algorithm 1

(∗ Genetic programming algorithm ∗)

Input: Size of initial population, n, probability of mutation, p_m

Output: Optimal trading rule(s)

1. Initialize population of size n
2. terminate = false
3. **while** not terminate
4. **do** evaluate the fitness $f_{(x)}$ of each individual
5. select the best individual and place it directly into the new population (elitism)
6. **while** new population is not completely generated
7. **do** randomly select two individuals from the current population according to $f_{(x)}$
8. crossover the parents to form a new offspring
9. **for** each new offspring
10. **if** individual selected for mutation based on p_m
11. **then** mutate individual
12. place new offspring in the new population
13. return best solution

We summarize the proposed framework in Figure 3.3. As illustrated in the figure, the events are extracted from the news messages represented in free text format, and constitute input to the genetic algorithm. The historical price data constitutes an individual input to the genetic algorithm, used for computing the performance of the trading strategies, but is simultaneously used to derive the values for technical trading indicators, another input to the genetic algorithm. Finally, the optimal trading strategies are determined through genetic programming.

3.6 Experiments and Results

In this section we provide an overview of the validation of our proposed framework for including news in stock trading strategies. First, we focus on the performance of the news variable taken individually, and then in combination with each of the technical indicators we consider. We then

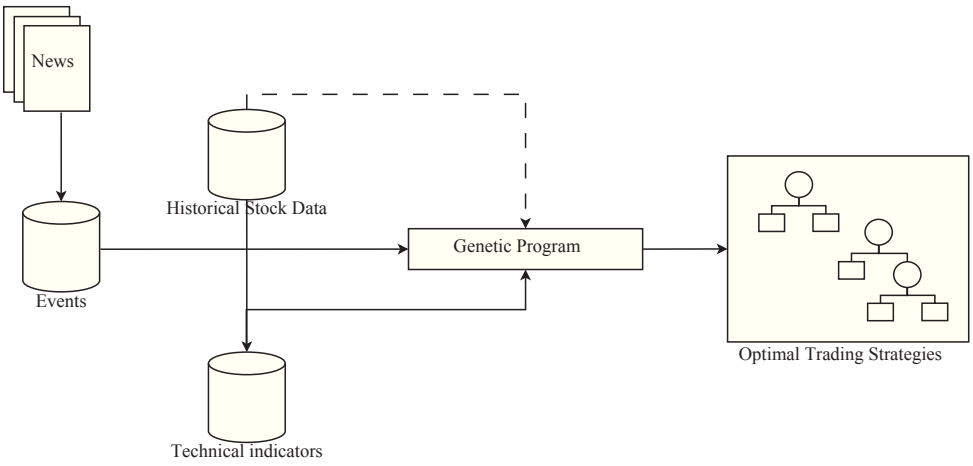

Figure 3.3: News-based trading framework

Table 3.7: Returns, signals generated by news

	Frequency	R_0	R_1	R_2	R_5	R_{10}
Buy	818	1.563	1.625	1.531	1.808	2.301
Sell	579	-1.578	-1.640	-1.606	-1.332	-1.061

present the performance of optimal trading rules as determined through genetic programming, and discuss these results.

3.6.1 Performance of Individual Events

When trading strategies are built only by using the news variable, we generate the returns displayed in Table 3.7. Here, a buy signal is generated when events are encountered that are known produce an R_0 of at least 0.5%, as shown in Table 3.1. Similarly, a sell signal is generated when the R_0 is below -0.5%. In Table 3.7 we present the results for buy and sell signals individually, for different time horizons.

A comparison with the results obtained by using trading strategies based only on technical indicators, as presented in Tables 3.5 and 3.6, reveals that the news variable performs relatively well. In the case of buy signals, the trading strategies based solely on the news variable are consistently outperformed by the simple moving average and the exponential moving average, but

Table 3.8: Returns, buy signals

	Frequency	R_0	R_1	R_2	R_5	R_{10}
News & SMA (20)	108	**3.005**	**2.888**	**2.748**	**3.236**	**3.954**
News & Rate of Change (10)	68	-0.098	**0.036**	-0.194	-0.060	**0.369**
News & Momentum	100	**3.058**	**2.857**	**2.976**	**3.339**	**3.874**
News & Bollinger Band	18	-0.841	-0.587	**0.011**	**0.374**	-0.174
News & EMA (5, 20)	54	**3.284**	**3.081**	**3.013**	**3.433**	**4.246**
News & MACD (12, 26)	33	**3.566**	**3.610**	**3.563**	**4.040**	**5.371**

only slightly. In the case of sell signals, the situation is similar, although the exponential moving average is not consistenly outperforming the trading strategies based on the news variable.

The good overall performance obtained by using technical indicators and news, respectively, suggests that combining these indicators might result in yet better trading strategies. In the next section, we look at how each of the technical indicators performs when considered in combination with the news variable.

3.6.2 News and Technical Indicators

In this section we consider the performance of the individual technical indicators when trading strategies combine each of them with the news variable. Again, buy and sell signals are considered separately, and the returns are presented for different time horizons.

Table 3.8 presents the returns generated with buy signals when news and technical indicators are considered together. Positive returns generated through buy signals are displayed in bold, as these are considered desirable results. An overall conclusion is that when news and technical indicators are considered together, the generated returns are higher than when these variables are considered individually. Out of the six combinations, four consistently generate positive returns at all time horizons. The highest observed return is generated by the combination of news and moving average convergence divergence, for time horizon R_{10}.

In Table 3.9 we present the returns generated through sell signals when the individual technical indicators are considered together with news. Again, results are presented for different time horizons. The overall conclusion is that the combination of the two indicators generally outperforms the trading strategies based on the indicators taken separately. From the six combinations, four consistently generate desirable returns at all time horizons. Again, the highest

Table 3.9: Returns, sell signals

	Frequency	R_0	R_1	R_2	R_5	R_{10}
News & SMA (20)	93	**-3.177**	**-3.700**	**-3.849**	**-3.353**	**-3.166**
News & Rate of Change (10)	28	0.211	**-0.414**	0.231	0.319	0.196
News & Momentum	78	**-3.304**	**-3.748**	**-3.969**	**-3.417**	**-3.369**
News & Bollinger Band	19	0.702	0.445	0.485	0.311	**-1.163**
News & EMA (5, 20)	42	**-3.790**	**-3.894**	**-3.977**	**-3.529**	**-3.654**
News & MACD (12, 26)	31	**-3.796**	**-4.217**	**-4.307**	**-4.555**	**-3.240**

return is achieved through the combination of news and moving average convergence divergence, but this time at time horizon R_5.

The results presented in this section allow us to conclude that combining individual technical indicators with the news variable for determining trading strategies enables higher returns than when technical indicators and the news variable are considered separately. We next move on to generating more complex trading strategies, that rely on multiple technical indicators, possibly in combination with the news variable, for generating trading signals.

3.6.3 Optimal Trading Strategies

The optimal trading strategies are determined through genetic programming, as outlined in Section 3.5. The initial population used by the genetic algorithm consists of 50 randomly generated trading rules. Additionally, for this analysis, we only consider the performance of buy signals.

In Table 3.10 we show the results obtained when the fitness of the generated trading rules is computed as the relative return one day after the generation of the signal. One of the optimal rules includes news as relevant variable, with a generated return of 2.65%, making it the second best performing rule. The simple moving average is included in all the rules, confirming the performance this indicator achieved when trading rules were considered that only take into account the individual technical indicators.

In Table 3.11 we provide an overview of the optimal trading strategies and the generated returns of these strategies when the fitness of a trading rule is computed as the return three days after the generation of the signal. We again note that the simple moving average is included in all trading strategies. However, when the time horizon over which returns are calculated consists

Table 3.10: Optimal trading strategies if stocks are held 1 day

Tree	Frequency	R_1
SMA (21) and MACD (26, 12)	14	2.891
NEWS and (SMA (27) or SMA (20))	36	2.650
SMA (20)	192	1.775
EMA(100, 27) or SMA (20)	196	1.770
MOM (10) or SMA (20)	349	1.459

Table 3.11: Optimal trading strategies if stocks are held 3 days

Tree	Frequency	R3
(MOM (4) AND SMA (24)) AND (BB (15) OR SMA (17))	65	5.122
SMA (14) AND SMA (26)	87	3.680
MOM (10) AND SMA (26)	44	3.491
(SMA (14) AND SMA (20)) AND SMA (17)	123	3.102

of three days, the news variable is not included in any of the optimal trading strategies. Finally, we note that the return of the best performing trading rule exceeds, on average, 5% three days after the generation of the signal.

Table 3.12 presents the optimal trading strategies found by the genetic program and the corresponding returns when the fitness of the trading strategies is computed as the return five days after the generation of the signal. Again, the simple moving average is included in all optimal trading strategies. The combination of news and momentum can also be found amongst the optimal trading strategies, generating a return of 2.29%.

Table 3.12: Optimal trading strategies if stocks are held 5 days

Tree	Frequency	R5
SMA (14) AND SMA (26)	87	3.680
SMA (28)	148	3.017
(NEWS AND MOM (7)) OR (NEWS AND (MOM (3))	44	2.290
MACD (26, 12) OR SMA (27)	212	2.266

3.7 Conclusions and Future Work

In this chapter we presented a framework for incorporating news into stock trading strategies. The trading strategies that we consider may include (in addition to the news variable) any number of technical trading indicators. The news variable is quantified based on the extraction of events from the text of news messages and the assignment of a predefined impact to each of these events. Our results indicate that the assigned impact correlates highly with the returns generated by these events when tested against actual data.

The selected technical indicators are also tested against actual data, and the individual performance of each indicator is reported. Additionally, combinations of individual technical indicators and the news variable are investigated. The results indicate that adding the news variable to each of the indicators generates higher returns than when each of the variables is considered alone. This indicates that considering the news variable indeed can lead to higher returns, thus making it worthwhile to consider trading rules that, next to technical indicators, consider the events that are relevant for a certain company.

Finally, a genetic program is used to discover more complex trading rules based on technical indicators and the news variable. For this purpose, we consider three time horizons when computing the fitness of the trading strategies based on the generated returns, namely 1, 3, and 5 days after the generation of the signal. The optimal trading strategies that we find include the news variable in two of the three time horizons considered, namely 1 and 5 days. We can thus conclude that, in the majority of the cases, news is indeed a relevant variable for trading rules, and its inclusion in trading strategies leads to higher returns than when this variable is not considered. We also conclude that the framework that we proposed is appropriate for including news in technical trading strategies.

Our results indicate that the inclusion of news into stock trading strategies can be achieved by extracting the events from the text of the news messages and associating an impact with the latter. This impact can later be used in the derivation of optimal trading strategies, where the news variable, consisting of the predefined impact, is used next to technical indicators. Returning to the two hypotheses stated in the introduction, namely that news will be included in the optimal trading strategies if news is a relevant variable and that these trading rules should

generate positive returns, we conclude that the news variable has been quantified in a meaningful way, since this variable is included in the optimal trading strategies. Additionally, all trading strategies that include generate a positive return, thus confirming our second hypothesis.

Future work will focus on including more indicators, technical or non-technical in nature, in the variable pool from which trading strategies are generated. Additionally, a more fine-grained analysis of the news messages, e.g., identification of event-related information such as the actors involved in an event, should provide more information that can be used in generating trading strategies. Finally, considering the interaction between individual events occurring within the same day, or within small intervals, will provide a deeper understanding of the way that news impact stock prices and may lead to more profitable trading strategies.

Together with the previous chapter, this chapter reveals the need for a systematic approach for the representation of the knowledge contained in news. The knowledge that we seek to represent relates to both the structure of the domain being studied – in our case the financial domain – and more concrete, ephemeral knowledge emerging from the information sources used – in our case news messages. Additionally, the representation must enable automated inference on the knowledge base, and provide the ability to discover inconsistencies.

Chapter 4

tOWL: A Temporal Web Ontology Language [1]

Through its interoperability and reasoning capabilities, the Semantic Web opens a realm of possibilities for developing intelligent systems on the Web. The Web Ontology Language (OWL) is the most expressive standard language for modelling ontologies, the cornerstone of the Semantic Web. However, up until now no standard way of expressing time and time-dependent information in OWL has been provided. In this chapter, we present a temporal extension of the very expressive fragment SHIN(D) of the OWL-DL language resulting in the tOWL language. Through a layered approach we introduce three extensions: i) Concrete Domains, which allows the representation of restrictions using concrete domain binary predicates, ii) Temporal Representation, which introduces timepoints, relations between timepoints, intervals, and Allen's 13 interval relations into the language, and iii) TimeSlices/Fluents, which implements a perdurantist view on individuals and allows for the representation of complex temporal aspects, such as process state transitions. We illustrate the expressiveness of the newly introduced language by using an example from the financial domain.

[1] An article based on this chapter has appeared in: Viorel Milea, Flavius Frasincar, Uzay Kaymak. tOWL: A Temporal Web Ontology Language. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, Volume 42, Number 1, pages 268–281, IEEE Computer Society, 2012.

4.1 Introduction

The considerable and increasing need to access the large volume of data present on the World Wide Web today motivates a migration from free-text representations of data to semantically rich representations of information. Endeavours in this direction are being undertaken under a common denominator: the Semantic Web (Berners-Lee et al., 2001). The state-of-the-art tools and languages provided under this umbrella, such as Resource Description Framework (RDF), RDF Schema (RDFS) (Klyne and Carroll, 2004; Brickley and Guha, 2004) and OWL (Patel-Schneider et al., 2004), go beyond the standard Web technology and provide the means for data sharing and reusing outside this platform, i.e., in the form of semantic applications.

Focused on the inference of implicit knowledge from explicitly represented information, Semantic Web approaches are currently centered around static abstractions of the world. However, conceptualizations lacking a temporal dimension are not only rather artificial, but are also impractical in environments that require temporal awareness. Examples of such environments are the financial domain, scheduling, marketing, etc. Within the financial domain, for example, one can envision the need for representing ephemeral knowledge, contained for instance in news messages (e.g., stock price and other financial variables), or more fundamental aspects of the financial domain (e.g., mergers and acquisitions, financial processes, etc.).

Consider, for example, the temporary relation between a person and a company in which that person is the Chief Executive Officer (CEO) of that company. Such a relation is described by a temporal interval across which a person fulfills the function of CEO for a company. For example, until 16 October 2008, Jack Dorsey was the CEO of Twitter. On that date Jack Dorsey stepped down and Evan Williams became the new CEO of the company. In the standard Semantic Web approach based on OWL-DL once the CEO of the company changes, there is no way to represent both CEOs and the times associated with them. Furthermore, one would like to reason with temporal information. For example, assuming that Evan Williams is only CEO for a limited amount of time, and that the ending point of him being a CEO of Twitter is known, we would like to consistently represent all this information in our temporal language, without any loss of information regarding the *ceoOf* relationship. In graphical terms, what we would like to represent in a temporal Semantic Web language is illustrated in Figure 4.1.

For this representation, we would like to be able to define temporal constraints such as that the starting point of a time interval should be before the ending point of the interval (in our example $t_1 < t_2$ and $t_2 < t_3$). Such knowledge and constraints cannot be be enforced using OWL-DL semantics. In this chapter, we propose a temporal extension to the OWL language that allows us to represent and reason with temporal information in the Semantic Web.

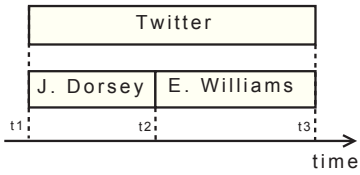

Figure 4.1: Change of CEO in the Twitter example

In general, addressing temporality in abstract representations of the world requires dealing with the aspect of time. One aspect is that of reference system - bringing an order into sequences of events. In this respect, time can be instant-based or interval-based, with instants denoting basic points in time with no duration, and intervals being represented as pairs of distinct instants denoting some period of time.

A second aspect of time regards temporal concepts such as the ephemeral character of relationships between individuals. In this context, representations of change should be possible. These representations include descriptions of individuals that take variable values for some property at different points in time and state transitions, enabling the representation of processes and corresponding transition axioms. In this context, time is somewhat implicit to the representation, i.e., the conceptualization evolves relative to the temporal reference system and requires the latter.

The main goal pursued in this chapter is an extension of a fragment of OWL-DL with time. The fragment of OWL-DL considered is $\mathcal{SHIN}(\mathcal{D})$, which represents OWL-DL without the use of nominals. In the remainder of this article, we shall denote the fragment of OWL-DL based on the $\mathcal{SHIN}(\mathcal{D})$ description logic as OWL-DL$^-$. We focus on this particular subset due to the fact that it is the most expressive fragment of OWL-DL extended with concrete domains for which a terminating, sound, and complete reasoning algorithm is known (Lutz, 2002).

In temporal terms, the extension of OWL-DL$^-$ that we envision addresses time in the sense of a reference system as well as covering more complex temporal aspects, such as change and state transitions. This materializes in a syntactic and semantic extension of OWL-DL$^-$ in the form of a temporal web ontology language (tOWL) (Milea et al., 2008b,c,a, 2007; Frasincar et al., 2010). Hence, the tOWL language is an extension of OWL-DL$^-$ that enables the representation of and reasoning with time and temporal aspects.

The extension of OWL that we present in this chapter is mainly aimed at extending the communication between machines to contexts that require temporality. Building upon the main goals of Semantic Web languages, tOWL is not only aimed at enabling the inference of implicit knowledge when a temporal dimension is involved, but aims at representing information, especially information that is temporal in nature, in an unambiguous fashion, in a unified way that ensures the preservation of meaning across different machines. Due to the desirable computational properties of the language, which is based on a decidable description logic extended with concrete domains, we enable temporal reasoning in tOWL knowledge bases that extends well beyond the capabilities of any of the Semantic Web languages developed until now.

Generally speaking, ontologies are separated into a TBox and an ABox. The TBox contains the terminological knowledge in the ontology, and refers to classes and properties. The ABox is the assertional part of the ontology, and relates to individuals - instances of the classes described in the TBox. The issue of time is also relevant in the context of TBoxes and ABoxes of ontologies. Rather than focusing on the evolution of ontologies, i.e., changes at TBox level, we solely focus on changes in the ABox (we assume that the domain structure is known).

The outline of the chapter is as follows. In Section 4.2 we provide an overview of work related to the current endeavor. Section 4.3 introduces different layers of the tOWL language built on top of OWL-DL$^-$. An extensive example of how the expressiveness of tOWL can be used for the representation of a financial process is provided in Section 4.4. We give a discussion Section 4.5, where we place our proposed language into a broader context. Finally, we conclude in Section 4.6.

4.2 Temporal Representations

World representations may be synchronic or diachronic in the way the temporal perspective is considered within the representation (The Stanford Encyclopedia of Philosophy, 2003). Synchronic representations consider a single point in time, with no regard for temporal evolution. Diachronic representations take into consideration the existence of a history, and thus take into account change through time. Regardless of the form of representation chosen, one must invariably deal with the problem of identity. Synchronic identity regards identity holding at one single time. Diachronic representations, which are our current focus, must deal with the problem of diachronic identity, or put differently, establish how change affects the identity of entities existing at different times.

Leibniz provided two principles regarding the issue of identity (Leibniz, 1969). The first one, regarding the *identity of indiscernibles*, states that entities for which all properties are common and identical are, in turn, identical. Additionally, *indiscernibility of identicals* states that entities being identical implies that the entities have all properties in common, and the values thereof are identical. Such choices are mainly concerned with the field of logics. As our focus is on the Semantic Web, special attention is given to approaches related to Description Logics (Baader et al., 2003). The issue of temporality has also been addressed in this context, presenting several choices regarding the handling of the temporal dimension. The main distinction separating approaches in temporal description logics can be made in terms of whether the temporal language offers explicit time, or whether the temporal dimension is only implicitly present in the language by providing the means to talk about an order of events and/or states. Following (Artale and Franconi, 2000), these different approaches are categorized as *explicit* and *implicit*, respectively.

Additionally, philosophy presents us with two main theories regarding the persistence of objects through time: endurantism and perdurantism. Endurantism refers to the view that objects are three-dimensional, and persist through time, i.e., are always present. Perdurantism, or four-dimensionalism, regards objects as being composed of temporal parts. The identity of a four dimensional object then consists of all the temporal parts of that object, i.e., all instances of that object through time. An approach related to incorporating perdurants through the use of timeslices and fluents is presented in (Welty and Fikes, 2006), where the authors develop a

reusable ontology for fluents in OWL-DL. In this approach, timeslices represent the temporal parts of a specific entity at intervals in time and the concept itself is then defined as all of its timeslices. Fluents are properties that hold at a specific moment in time, or at a specific interval in time. One of the drawbacks of this approach is the proliferation of objects in the ontology due to the creation of two timeslices each time something is changing, which, in turn, must be associated to the static individuals they represent and linked to each other by a fluent. Further, no solution is provided for the temporal equivalent of the cardinality construct, which cannot be modeled in the case of overlapping timeslices (Welty and Fikes, 2006). Finally, the time associated to timeslices relies on the OWL-Time ontology, rather than on a more expressive approach based on concrete domains.

The expressiveness of description logics is usually denoted with a series of letters, such as $\mathcal{SHOIN}(\mathcal{D})$, where each letter stands for a level of expressiveness. A language that allows functional properties will contain the letter $\mathcal{F}$ in its name, while $\mathcal{S}$ stands for the $\mathcal{ALC}$ language (attributive language with complement) with transitive roles, $\mathcal{H}$ stands for role hierarchy, $\mathcal{O}$ for nominals, $\mathcal{I}$ for inverse roles, and $\mathcal{N}$ for number restrictions. The $\mathcal{SHOIN}$ language thus is the language that incorporates the expressiveness associated with each letter as described above, and $\mathcal{SHOIN}(\mathcal{D})$ additionally provides support for data types, as indicated by $\mathcal{D}$.

When logics, and especially description logics, are considered, concepts such as decidability, concept satisfiability, and subsumption play an important role. Decidability relates to whether a method exists such that formulas validity can always be determined in a finite number of steps. Concept satisfiability consists of checking whether an assertion has a model, i.e., its interpretation is non-empty. Finally, subsumption relates to being able to determine whether one concept is more general than some other concept (Baader et al., 2003). Decidability is highly relevant, especially in the context of the Semantic Web, where machines must be endowed with the capability to reason on the knowledge that is being presented to them. The focus of the Semantic Web on description logics comes from the fact that the DL community has always placed emphasis on the decidability of the logics introduced. As our temporal extension of the web ontology language is also intended for the Semantic Web, special attention is given to the

decidability of the language. When considering relevant literature on temporal extensions to description logics, we also focus on the decidability of the temporal extension.

The interval-based $\mathcal{TL}$-$\mathcal{ALCF}$ description logic (Artale and Franconi, 1998, 1994) enables the representation of temporal interval networks through Allen's interval temporal logic in the context of the static $\mathcal{ALCF}$ description logic. The resulting logic is an aggregation of a temporal and static logic, thus making this approach external as the temporal dimension is external to the $\mathcal{ALCF}$ description logic. Returning to our current focus, OWL-DL$^-$ and the very expressive underlying description logic $\mathcal{SHIN}(\mathcal{D})$, it can be concluded that an approach not moving beyond the expressiveness of $\mathcal{ALCF}$ is insufficient for our goal. This relates mostly to the fact the $\mathcal{ALCF}$ description logic is much less expressive than $\mathcal{SHIN}(\mathcal{D})$, which is our main focus for the temporal extension. In the temporal language that we propose, role hierarchy ($\mathcal{H}$) and inverse roles ($\mathcal{I}$) play an important role. For example, the fluent *ceoOf* is a more specific variant of the *worksFor* fluent, which can only be represented in the knowledge base when role hierarchy is enabled. Additionally, the inverse of the *ceoOf* fluent, *hasCEO*, can only be represented in the knowledge base when inverse roles are enabled by the language.

Approaches similar to $\mathcal{TL}$-$\mathcal{ALCF}$, but relying on a point-based temporal structure rather than an interval-based one, provide the means to represent temporal dependencies between entities. An example of one such logic consists of the $\mathcal{DLR}$ description logic extended with the temporal operators *Until* and *Since*, resulting in the $\mathcal{DLR}_{\mathcal{US}}$ temporal description logics (Artale et al., 2001). Similarly, the $\mathcal{ALCT}$ temporal description logic is an extension of $\mathcal{ALC}$ with the temporal connectives of tense logic, such as existential and universal future (Artale and Franconi, 2000). These approaches are not sufficient for our current purpose mainly because extending the expressiveness of the static DL in this way easily leads to undecidability.

Time can also be incorporated in a formalism by making it part of the latter, in what constitutes an internal approach. One such approach consists of extending a DL formalism with concrete domains. Initially proposed in (Baader and Hanschke, 1991), concrete domains allow abstract concepts to be related to concrete values through functional roles (roles that take exactly one value). Description logics extended with concrete domains maintain decidability, provided that the concrete domain satisfies the property of admissibility or ω-admissibility (Baader and

Hanschke, 1991; Lutz and Milicic, 2007). For a number of constraint systems, special types of concrete domains based on binary domain predicates that are jointly-exhaustive and pairwise disjoint, ω-admissibility has been proved in (Lutz and Milicic, 2007), such as a constraint system based on a domain consisting of intervals and Allen's 13 interval relations that may hold between pairs of intervals. This constraint system approach to introducing time in DL-based formalisms is less restricted by the expressiveness of the static DL which it extends. Indeed, results are known for $\mathcal{SHIQ}(\mathcal{C})$, description logic for which a terminating, sound and complete reasoning algorithm is known (Lutz, 2002).

Linear time temporal logic (LTL) has also been considered as a temporal extension of DLs (Lutz et al., 2008), particularly with regard to the $\mathcal{ALC}$ description logic, resulting in the $LTL_{\mathcal{ALC}}$ temporal DL. However, extensions of this TDL to more expressive DLs, such as $\mathcal{SHIQ}$, easily results in undecidability (Lutz et al., 2008), thus making such an approach unsuitable due to the limited expressiveness of the underlying language.

Different aspects regarding the representation and management of time-varying data have also been addressed within the broad area of temporal databases. A common way of regarding time in such a context relates to the type of time that is addressed by the system. This has resulted in three types of time (Jensen et al., 1994) that may be considered in a temporal database: i) valid time, the time when a fact is true in the real world, ii) transaction time, the time when the fact is known in the database, and iii) user-defined time, which can represent any temporal attribute for which the temporal semantics is only known to the user and has no particular meaning in the database. Combinations of these types are also possible. When valid time and transaction time are considered together, this results in bitemporal data models (Jensen et al., 1994). Regarding the structure of the time domain, a further distinction may be made between linear time - one time flow from past, through present, to future - and branching time, where the representation of possible, alternative futures is allowed (Ozsoyoglu and Snodgrass, 1995).

In the context of the Semantic Web, a number of approaches have already been designed, addressing different temporal aspects in relation to ontology languages. A rather extensive approach towards extending ontology languages with a temporal dimension is Temporal RDF

(Gutierrez et al., 2007). This work is similar to our approach as it concerns the ability to represent temporal information in ontologies, but differs in that the language considered is the Resource Description Framework (RDF). Another approach is OWL-Time (Hobbs and Pan, 2004), which focuses on the Web Ontology Language rather than RDF. The initial purpose behind the design of a time ontology (OWL-Time) was to represent the temporal content of Web pages and the temporal properties of Web Services (DAML-Time) (Hobbs and Pan, 2004). This approach is rather extensive in describing quantitative time and the qualitative relations that may exist among instants and intervals. Being based on OWL-DL, it uses the underlying $\mathcal{SHOIN}(\mathcal{D})$ description logic and thus relies on data types rather than concrete domains for the description of instants and intervals. Due to this fact, proper intervals, i.e., intervals for which the starting point is strictly smaller than the end point, cannot be represented in this approach. Semantic Web approaches similar to ours also include (Batsakis and Petrakis, 2010, 2011), relating to a 4d fluents approach for representing change, and (Motik, 2010) focusing on the representation of valid time in RDF and OWL. In the following, we discuss our proposed temporal extension of the web ontology language.

4.3 The Temporal Web Ontology Language

Designing a temporal extension of OWL-DL$^-$ begins with a clarification of what is understood under the general, common denominator *time*. We consider a couple of fundamental aspects hereof, namely: i) temporal infrastructure, and ii) change. The first aspect, *temporal infrastructure*, regards the representation of time in the form of instants and/or intervals. From this perspective, we aim for an approach that incorporates both a point-based as well as an interval-based time representation. Such an approach should provide not only the temporal entities that constitute the temporal infrastructure of the language, but also the relations that may hold between these entities, e.g., the *before* relation that may hold between intervals.

Regarding the second aspect, *change*, we note that there are two types of changes in OWL ontologies: changes at the terminological level (TBox), and changes at the assertional level (ABox). For the tOWL language, the focus is solely on changes that concern individuals; in

other words, tOWL enables the representation of change at the ABox level. We allow three types of change: i) change in a concrete attribute value of an individual, such as a change of hair color, ii) a change in the relationship between entities, such as a product that is built by a company, and iii) state transitions in processes, such as the transition from the liquid state to a bankruptcy state in the case of companies. In this context, we refer only to valid time, as known from temporal databases, rather than transaction time. Therefore, we seek to represent when certain changes take place in the actual world rather than the time when they are represented in the ontology.

In the following we discuss the details of our proposed tOWL language. The design uses the results from temporal logic, temporal databases, and Semantic Web research where possible. The design choices are explained in Section 4.3.1. The individual tOWL layers are presented, one by one, in Sections 4.3.2 through 4.3.4. A discussion on reasoning in the tOWL language is presented in Section 4.3.5.

4.3.1 Design Choices

For the tOWL language there are a number of choices regarding the most suitable approach(es) for the representation of the two temporal aspects considered above. At the level of temporal infrastructure, we seek to enable point-based as well as interval-based representations. Additionally, we seek to extend the expressiveness of OWL-DL$^-$ and the underlying $\mathcal{SHIN}(\mathcal{D})$ description logic without constraining the latter. From the approaches known in the literature, the only method suitable for our goals is the one based on concrete domains. The temporal infrastructure then becomes internal to the language, and covers both the point-based time and the interval-based time. For a point-based representation of time, we rely on a concrete domain based on the set $\mathbb{Q}$ of rational numbers and the set of binary concrete domain predicates $\{<, \leq, =, \neq, \geq, >\}$. Results are known for such an extension to the description logic $\mathcal{SHIQ}$, where the concrete domain is also extended with an additional unary predicate $=_q$ for denoting equality with $q \in \mathbb{Q}$, resulting in the $\mathcal{SHIQ}(\mathcal{C}^+)$ (Lutz, 2002, 2003). Introducing such a concrete domain in the language has the advantage of not only enabling the representation of dates and times in terms of a translation between the *xsd:dateTime* XML data type and rational numbers,

but enables also the description of any numerical attribute through a direct reference to the concrete domain.

In our approach, we seek to enable an interval-based representation of time satisfying the previously mentioned constraints. For this purpose, we aim to add intervals and Allen's 13 interval relations (Allen, 1983) to the tOWL language. As known from (Allen, 1983), all 13 Allen's interval relations may be translated in terms of equivalent relations on the intervals' endpoints. For this reason, the concrete domain based on the set $\mathbb{Q}$ of rational numbers and the set of binary concrete domain predicates $\{<, \leq, =, \neq, \geq, >\}$ is sufficient for such representations. Thus, intervals and Allen's 13 interval relations are not introduced in the language by means of a concrete domain, but rather as syntactic sugaring over the concrete domain $\mathbb{Q}$ with the respective relations. By only introducing one concrete domain into the language, we build upon known decidability results (Baader and Hanschke, 1991; Lutz, 2003) for description logics extended with concrete domains and ensure the language decidability.

The representation of change in a temporal ontology language poses several problems that need to be addressed. We consider diachronic representations that take history into account rather than synchronic ones, and are thus faced with the problem of diachronic identity, as mentioned in Section 4.2. The second principle of Leibniz, indiscernibility of identicals, poses an additional restriction on the choice of representation and the perspective on identity when change is involved. Finally, as the temporal language we develop is aimed at the Semantic Web, one must invariably be able to say what holds true at a certain moment in time. The Semantic Web, and OWL-DL$^-$ in this context, further restrict the flexibility of designing an approach for the representation of change due to the restriction of the underlying $\mathcal{SHIN}(\mathcal{D})$ description logic.

The straight-forward approach of associating a valid time to the binary predicate (similar to solutions from temporal databases and temporal RDF) is not suited in our case, as ternary predicates are not directly supported in OWL-DL. The W3C Semantic Web Best Practices working group provides three alternative ways of representing n-ary relationships on the Semantic Web (Noy and Rector, 2005), namely: i) representing a relationship as a class rather than as a property, ii) representing the individuals participating in the relation in the form of a collection

or ordered list, and iii) RDF reification. The first two approaches share the drawbacks of proliferation of objects and the reduced meaning of the actual representation of instances, especially in the case of OWL-DL. Regarding the third, it should be noted that RDF reification is not appropriate when "the intent is to talk about instances of a relation, not about statements about such instances" (Noy and Rector, 2005). Besides the fact that the RDF "reification of a triple does not entail the triple, and is not entailed by it" (Hayes and McBride, 2004), reification is not supported at all in OWL-DL$^-$. Since we are extending OWL-DL$^-$, such an approach is not suitable.

Another approach for associating valid time with a binary relation relates to the addition of a meta-logical (McCarthy and Hayes, 1969) predicate that takes as arguments the binary relationship and the time when this relationship holds. However, as also discussed in (Welty and Fikes, 2006), such predicates are not supported in any of the OWL species. The fluents approach presented in (Welty and Fikes, 2006) and discussed in Section 4.2 is consistent with the second principle of Leibniz and enables the maintenance of identity through change by introducing a 4D view of the world in OWL ontologies. By moving the temporal argument to the level of timeslices rather than the fluent itself, it circumvents the issue of n-ary relationships, while still enabling the determination of what holds true at a particular time. This approach also has the advantage of not restricting the expressiveness of the description logic it extends, as it is more concerned with syntactic sugaring rather than being a semantic extension. As introduced in (Welty and Fikes, 2006), this 4D approach can be achieved in the form of an OWL ontology, which although insufficient for extending the OWL-DL$^-$ language, should prove a good starting point in addressing the representation of change in the tOWL language.

For the design of the language we choose a layered approach. On top of the foundational OWL-DL$^-$ layer, we add a concrete domains layer, a temporal reference layer, and a 4d fluents layer, as described in the following sections.

4.3.2 Concrete Domains Layer

The representation of complex restrictions, regardless of whether they describe some temporal aspect, or relate to some static expression, can be achieved through the composition of roles. In

what follows, we denote by *feature chain* a composition of features (functional roles). Following common denomination from Description Logics and the Semantic Web, we make a distinction between abstract features, that point to something in the abstract domain, and concrete features, that take values from the concrete domain. Additionally, in tOWL we allow the feature chains to be composed with one concrete feature g, forming what is commonly denoted as a *concrete feature path* (CFP), and which is mathematically equivalent to the composition:

$$f_1 \circ f_2 \circ ... \circ f_n \circ g, \tag{4.1}$$

where $n \in \mathbb{N}$. Note that for $n = 0$, by convention, the set of abstract features is empty.

An example of such a CFP could consist of the composition of the *time* abstract feature and the *start* concrete feature, resulting in a composition of type $f_1 \circ g$, where f_1 is the *time* feature and g is represented by the *start* feature, as follows:

$$time \circ start. \tag{4.2}$$

A construction such as the one in (4.2) would denote the starting point of an interval by first applying the *time* abstract feature to obtain the interval associated with an individual and then the *start* concrete feature to obtain the starting point of that interval.

Letting u_i denote a CFP, we allow existential and universal quantification of the following form in tOWL, where p_d denotes a binary concrete domain predicate:

$$\exists(u_1, u_2).p_d, \tag{4.3}$$

$$\forall(u_1, u_2).p_d. \tag{4.4}$$

For such constructs, u_i may arbitrarily denote a CFP of length m, with $m \in \mathbb{N}^*$. Such constructs are useful for defining, for example, that the starting point of an interval should be strictly smaller than the ending point of that interval. Such a definition of an interval would take the following form:

Table 4.1: Semantics for the concrete domains layer

tOWL Abstract syntax	Model-Theoretic Semantics
`ConcreteFeatureChain`($f_1\ f_2\ \ldots\ f_n\ g$)	$\{(a_1, b) \in \Delta^{\mathcal{I}} \times \Delta_{\mathcal{D}} \mid \exists! a_2 \in \Delta^{\mathcal{I}}, \ldots, \exists! a_{n+1} \in \Delta^{\mathcal{I}} \wedge$ $\wedge\ \exists! b \in \Delta_{\mathcal{D}} : (a_1, a_2) \in f_1^{\mathcal{I}}, \ldots$ $(a_n, a_{n+1}) \in f_n^{\mathcal{I}} \wedge g^{\mathcal{I}}(a_{n+1}) = b\}$.
`dataSomeValuesFrom` ($u_1\ u_2\ p_d$)	$\{x \in \Delta^{\mathcal{I}} \mid \exists! q_1 \in \Delta_{\mathcal{D}}, \exists! q_2 \in \Delta_{\mathcal{D}} :$ $u_1^{\mathcal{I}}(x) = q_1 \wedge u_2^{\mathcal{I}}(x) = q_2 \wedge\ (q_1, q_2) \in p_d^{\mathcal{I}}\}$.
`dataAllValuesFrom` ($u_1\ u_2\ p_d$)	$\{x \in \Delta^{\mathcal{I}} \mid \forall q_1 \in \Delta_{\mathcal{D}}, \forall q_2 \in \Delta_{\mathcal{D}} :$ $u_1^{\mathcal{I}}(x) = q_1 \wedge u_2^{\mathcal{I}}(x) = q_2 \wedge\ (q_1, q_2) \in p_d^{\mathcal{I}})\}$.

$$\exists(time \circ start, time \circ end). < \tag{4.5}$$

where we use the $<$ concrete domain predicate (p_d) to state that the starting point of some interval is strictly smaller than its ending point.

We summarize the semantics introduced by this layer in Table 4.1, with reference to the tOWL abstract syntax constructs we propose. In Table 4.1, f_n denotes an abstract feature, g a concrete feature, u_i is a concrete feature chain, a_i and x are individuals from the abstract domain, b, q_1, and q_2 are concrete values, and p_d is a concrete domain predicate. For a complete overview of the tOWL language we refer the reader to the appendix.

The first definition in this table, that of a `ConcreteFeatureChain`, states that the interpretation of such a concept consists of all those pairs of individuals of the abstract domain and the concrete domain, respectively, such that each of these abstract individuals is in the interpretation of the f_1 abstract feature together with exactly one other abstract individual, a_2, which in turn is in the interpretation of f_2, together with exactly one other individual, a_3, and so on to a_{n+1}. Finally, the interpretation of the concrete feature g on the individual a_{n+1} should be defined and should take on exactly one concrete value, namely b.

The `dataSomeValuesFrom` construct, states that the interpretation of such a concept consists of all those individuals from the abstract domain such that, when the two concrete feature chains u_1 and u_2 are interpreted over these individuals, the result consists of the q_1 and q_2 unique concrete values that are in the interpretation of the p_d concrete domain predicate. Hence, they

can be described by p_d. Since this is an existential quantification, the values involved should exist, i.e., be explicitly defined.

The `dataAllValuesFrom` construct is similar to the `dataSomeValuesFrom`, with the exception that this time a universal quantification is involved, which means that the p_d relation should hold true for all values of q_1 and q_2. The difference between these two constructs is that, in the case of `dataAllValuesFrom`, the relationship can hold true when q_1 and q_2 are missing.

4.3.3 Temporal Reference Layer

The concrete domain in the tOWL context, as presented in the previous section, enables the representation of new restrictions in the language. In the *Temporal Reference* layer we include basic representations of time, both point-based and interval-based, as well as a number of temporal relations between instants and intervals, as discussed in Section 4.3.1. This forms the basis for our approach, as it allows the definition of complex restrictions, such as the ones described in the previous section, but this time presenting a temporal character. The concrete domain used for the current purpose is a concrete domain based on the set $\mathbb{Q}$ of rational numbers and the set of binary concrete domain predicates $\{<, \leq, =, \neq, \geq, >\}$.

This concrete domain also enables the representation of intervals and Allen's 13 interval relations through a translation scheme between interval relations and equivalent relations in terms of the intervals' endpoints (Lutz, 2000). Rather than being a concrete domain, this extension is achieved by means of syntactic sugaring at language level, while at reasoner level we rely on the concrete domain $\mathbb{Q}$ and the corresponding relations for dealing with representations based on intervals.

Another issue regarding time in this context relates to its representation in tOWL ontologies. The actual representation of time in tOWL ontologies is based on XML Schema data types, namely *dateTime* as enabled by the concrete domain based on rational numbers and relations that may exist between these numbers. Finally, it should be noted that the definition of intervals as introduced by tOWL goes beyond the expressiveness of OWL-DL$^-$ by relying on the concrete domain predicate $<$ and the two concrete features *start* and *end* for stating that the starting point of an interval should always be strictly smaller than its ending point:

$$\texttt{ProperInterval} \equiv \exists(\texttt{begin, end}).< \tag{4.6}$$

Although not directly enforced on the user, this restriction on proper definitions of temporal intervals is checked by the reasoner. All intervals that do not satisfy this restriction are not considered proper intervals, which will be indicated to the user through the reasoning service following a consistency check.

4.3.4 4d Fluents Layer

The concrete domain approach that enables a temporal infrastructure in ontologies as presented in the previous sections forms the basis for our approach. Building further upon this, we seek to represent temporal aspects of entities other than timespan. In this context, the final level of expressiveness that we enable in tOWL considers different aspects of change: i) change in a concrete attribute value of an individual, ii) a change in the relationship between entities, and iii) state transitions in processes.

A perdurantist approach forms the foundation of this type of features. Up to a certain level, it can be argued that the fluents and timeslices used for the representation of temporal information do not go beyond the expressiveness of OWL-DL$^-$. Rather, fluents and timeslices represent a vocabulary used for the representation of temporal parts of individuals that change some property in time. However, the semantics of fluents as envisioned for tOWL enforces a number of restrictions on tOWL specific concepts, and most importantly on fluents and timeslices. Some interesting features emphasize the interdependence between the concrete domain and the timeslices/fluents approach and relate mostly to the restrictions this approach imposes on the very concepts it introduces.

One such restriction imposes that fluents only relate timeslices that hold over the same time interval. Representing such a restriction involves the concept of equality of concrete values, and such a representation can thus only be enabled through the use of a concrete

domain. We illustrate this idea through an example that we graphically depict in Figure 4.2. In this example, we define two OWL classes, namely *Company* and *Product*. For each of these classes, we instantiate an individual, namely *iGoogle* and *iChrome*, respectively, representing the company Google and Chrome, the web-browser from Google. For each of these individuals, we instantiate a timeslice, namely *iGoogle_TS1* and *iChrome_TS1*, respectively, representing the static individuals over the periods *iInterval1* and *iInterval2*. Here, the two timeslices *iGoogle_TS1* and *iChrome_TS1* share the same time interval, i.e., *iInterval1* is equal to *iInterval2*, as denoted by the *towl:equal* relationship. Finally, the two timeslices are connected by the fluent *hasProduct* that indicates that over the period *iInterval1* (equivalent to the period *iInterval2*) Google Chrome is a product of Google.

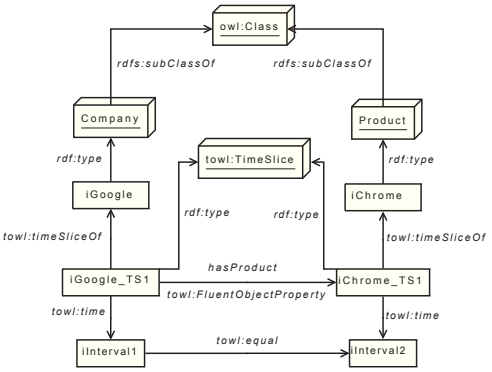

Figure 4.2: Temporal restrictions on timeslices connected by fluents

In Table 4.2 we present an overview of the tOWL TBox axioms corresponding to the timeslices/fluents layer. The tOWL axioms and facts are described in the appendix.

In Table 4.2, the concept of `TimeSlice` is defined as all those individuals for which the `time` property is defined and takes a value of type `Interval`, and for which the `timeSliceOf` property is defined and takes a value that is not an `Interval`, a `TimeSlice`, or a `Literal`. The concept of `Interval` is defined as all those individuals for which the `start` and `end` properties are defined and take a value from XML Schema `dateTime` such that the value associated to the starting point is smaller than the value associ-

Table 4.2: tOWL axioms for the *4DFluents* layer

tOWL 4dFluents Construct	tOWL Axioms in OWL-DL
Class(TimeSlice)	$\exists$time.Interval $\sqcap$ (= 1 time) $\sqcap$ $\exists$timeSliceOf.$\neg$(TimeSlice $\sqcup$ Interval$\sqcup$ $\sqcup$ rdfs:Literal) $\sqcap$ (= 1 timeSliceOf)
Class(Interval)	$\exists$(start, end). $\leq$ $\sqcap\exists$start.dateTime $\sqcap$ $\exists$end.dateTime$\sqcap$ $\sqcap$(= 1 start) $\sqcap$ (= 1 end)
Class(FluentProperty)	FluentProperty $\sqsubset$ rdf:Property
Class(FluentObjectProperty)	FluentObjectProperty $\sqsubset$ FluentProperty
Class(FluentDatatypeProperty)	FluentDatatypeProperty $\sqsubset$ FluentProperty
Property(timeSliceOf)	$\geq$ 1 timeSliceOf $\sqsubseteq$ TimeSlice $\top \sqsubseteq$ $\forall$timeSliceOf.$\neg$(TimeSlice $\sqcup$ Interval $\sqcup$ rdfs:Literal)
Property(time)	$\geq$ 1 time $\sqsubseteq$ TimeSlice $\top \sqsubseteq$ $\forall$time.Interval
Property(start)	$\geq$ 1 start $\sqsubseteq$ Interval $\top \sqsubseteq$ $\forall$start.dateTime
Property(end)	$\geq$ 1 end $\sqsubseteq$ Interval $\top \sqsubseteq$ $\forall$end.dateTime

ated to the ending point. The concept of `FluentProperty` is defined as a subclass of the RDF `Property` class, and is in turn a superclass for the `FluentObjectProperty` and `FluentDatatypeProperty` constructs. The `timeSliceOf` property is defined as that property that can be applied to timeslices and that only takes values that are not timeslices, intervals, or literals. The `time` property is defined as that property that takes values only of type `Interval` and can be applied to individuals of type `TimeSlice`. The `start` and `end` properties are defined as those properties that are defined for intervals and that take values from XML Schema `dateTime`.

4.3.5 Reasoning

The tOWL language extends OWL-DL$^-$ through the addition of constructs that support the representation of time and temporal aspects. The $\mathcal{SHIN}(\mathcal{D})$ description logic, on which OWL-DL$^-$ is based, is insufficient for the expressiveness introduced by the tOWL layers. Currently, a reasoner has been implemented for the Lite version of the tOWL language. The tOWL-Lite language is based on the $\mathcal{ALC}(\mathcal{C})$ description logic, and is thus limited in expressiveness. However, this logic is sufficient for representing fairly complex cases, such as the Leveraged Buy Out example in Section 4.4. The reasoner is based on the algorithm described in (Lutz and Milicic, 2007), extended with a number of optimization techniques meant to enhance the efficiency of the algorithm. The implemented

optimizations are: Normalization and Simplification Normalization, TBox Absorption, RBox Absorption, Lazy Unfolding, Dependency-directed Backjumping, and Top-Bottom Search for Classification (Maggiore, 2008).

The complexity of ontology entailment in $\mathcal{SHIQ}(\mathcal{C})$ and thus also of tOWL is ExpTime-complete (Lutz, 2002, 2003), and for $\mathcal{ALC}(\mathcal{C})$ and tOWL Lite it is as well ExpTime-complete (Lutz, 2000, 2003), provided that the satisfiability in $\mathcal{C}$ (the concrete domain) can be decided in ExpTime. Additionally, the timeslices/fluents extension proposed for the tOWL language (the 4d fluents layer) is merely syntactic sugaring, and does not incur reasoning cost when regarded from the perspective of language complexity.

Rather than extending existing reasoners, the tOWL-Lite reasoner consists of a new C++ implementation containing the tableau algorithm for the unrestricted version of the $\mathcal{ALC}(\mathcal{C})$ description logic as described in (Lutz and Milicic, 2007). The execution of algorithms based on tableaux as an inference procedure for expressive logics requires a massive use of dynamic structures thus motivating the implementation of a new reasoner from scratch using C++. The decision to implement a new reasoner from scratch has been taken due to the lack of documentation, or very poor documentation when present, of existing reasoners, thus not fostering extensions and making the choice for the design of a new reasoner necessary.

The tOWL reasoner enables different temporal inferences on tOWL knowledge bases. For example, given a time instant, we can determine what holds true at the moment in time based on the *inside* relationship between an instant and an interval. In this way, it can be determined, at any point in time, which timeslices hold true, since each timeslice has an interval associated with it. Thus, we can determine what facts are true at any moment in the knowledge base. Additionally, based on the relationships between intervals, we can, for example, determine, how intervals relate to each other in temporal terms, and thus the facts that we represent in the knowledge base. More concrete examples of reasoning in a practical application are provided in Section 4.4.3.

The correctness of the reasoner has been tested by using the benchmark suite proposed for Description Logics systems (Horrocks and Patel-Schneider, 1998). The test procedure consists of four categories of tests, as outlined in (Horrocks and Patel-Schneider, 1998): concept satisfiability, artificial TBox classification, realistic TBox classification, and synthetic ABox tests. The concept satisfiability tests are focused on the performance of computing satisfiability of large concept expressions without reference to a TBox. The artificial TBox classification tests investigate the performance of classifying an artificially generated TBox, while the realistic TBox classification tests perform the same investigation but on knowledge bases related to the GALEN medical terminology knowledge base (Horrocks, 1998). Finally, the synthetic ABox tests look at the system's performance when realising a synthetic ABox.

4.3.6 RDF/XML Serialization

We present the RDF/XML serialization of the tOWL abstract syntax as a separate document available as an electronic attachment to this chapter. The serialization is relevant as the RDF/XML syntax is the lingua franca of Semantic Web applications. By providing the serialization we enable different users to use the tOWL language in their applications. These applications can export the tOWL data for reuse in interoperability scenarios.

4.4 Example Application

In this section we illustrate the use of the tOWL language in a temporal context. For this purpose, we focus on a complex process - Leveraged Buy Outs (LBO) in financial applications. In Section 4.4.1 we present LBO processes in general, and introduce the Alliance Boots LBO. In Section 4.4.2 we illustrate how such a process can be represented in the tOWL language. We conclude this section by providing some reasoning examples for the LBO application in Section 4.4.3.

A different implementation of the tOWL language, next to the one presented in this chapter, is described in (Milea et al., 2008a). Here, we use the tOWL language for the representation of company market recommendations in a system that aggregates these recommendations for the generation of buy/hold/sell advices for the stock market. In this example, tOWL proves valuable in representing the different recommendations that may hold at different points in time, and overlap each other, and in determining which advices hold true at any point in time.

4.4.1 Leveraged Buy Outs in General

A Leveraged Buy Out is a special type of an acquisition of a company by another company by relying mostly on loans for the price of the acquisition. Additionally, often the assets of the company that is to be acquired are used, partly or wholly, as collateral for the loans. This type of process is of particular interest in the current case for two reasons: i) its complexity is adequate for illustrating the main features of the tOWL language, and ii) the ability to deal with such a process in an automated fashion is also of interest in the economic domain, due to the high impact that the different stages have on the share prices of the involved companies. An activity diagram of an LBO process is presented in Figure 4.3.

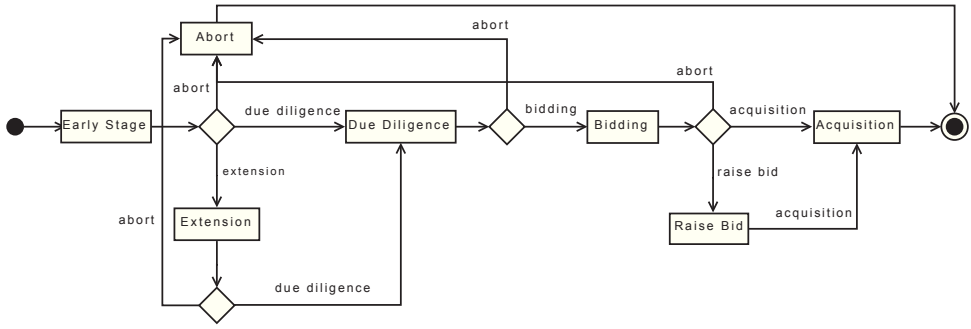

Figure 4.3: Stages of an LBO process

An LBO process can be divided into 4 main stages:

1. Early Stage.
2. Due Diligence.
3. Bidding.
4. Acquisition.

The transition between stages is not straightforward, as after nearly each stage the process can be aborted. Additionally, some of the stages may be extended before the transition into a different stage. In the bidding stage, the extension leads to a raise of the current bid. The initial state of an LBO process is the *Early Stage*. From this stage, a transition can be made into the next state - *Due Diligence*, or this state may be extended, or the whole process can be aborted. Whether an extension is granted or not, the process may evolve to the *Due Diligence* stage. In case the process is not aborted in this stage, the LBO can continue with the *Bidding* phase. Again, besides the process being aborted, the LBO can continue with a *Raise Bid* phase in which the companies involved increase the amount they are prepared to lay down for the target company. When the final bid is made and accepted, even in the case when no counter bids are made, the process moves into the *Acquisition* phase and ends.

In the following we consider a model of the biggest LBO acquisition in Europe. In the March and April of 2007, two hedge funds competed for the acquisition of one target company. From the two hedge funds, KKR and Terra Firma, the first won the bidding and acquired the target company Alliance Boots.

4.4.2 The Alliance Boots LBO in tOWL

The focus of this section is to illustrate how the information regarding the LBO process can be represented in tOWL abstract syntax, both at an abstract level as well as in the particular example presented here. The main focus is on illustrating the main concepts that are relevant from a language perspective.

TBox

At TBox level we represent conceptual information that is known about LBO processes in general. In this context, two types of companies that take part in an LBO are known: `HedgeFund` and `Target`, which we define as subclasses of the `Company` class.

```
Class(Company)
Class(HedgeFund partial Company)
Class(Target partial Company)
```

The different stages of an LBO process are represented as subclasses of the `Stage` class, such as for example in the case of the `Bidding` stage.

```
Class(Stage)
Class(Bidding partial Stage)
```

All stages are pairwise disjoint, which we represent as follows.

```
DisjointClasses(EarlyStage, DueDiligence, Bidding,
RaiseBid, Acquisition, Abort, Extension)
```

We define the class of all timeslices of an LBO Process as follows.

```
Class(LBOProcess_TS complete
     restriction(timeSliceOf(someValuesFrom
      LBOProcess)))
```

In similar fashion, we define, for each stage, the class of all timeslices of that stage. For the `EarlyStage` this is achieved as follows.

```
Class(EarlyStage_TS complete
     restriction(timeSliceOf(someValuesFrom
      EarlyStage)))
```

For each stage, we define a functional property that links a particular LBO process timeslice to the timeslice of the stage belonging to it.

```
ObjectProperty(earlyStage
     domain(LBOProcess_TS)
     range(EarlyStage_TS))
Func(earlyStage)
```

Next, we move on to define the `inStage` fluent that points, for each timeslice of a company, to the stage in which the company finds itself.

```
FluentObjectProperty(inStage
 domain(
  restriction(timeSliceOf(someValuesFrom
   Company)))
 range(
  restriction(timeSliceOf(someValuesFrom
   Stage))))
```

Timeslices of an LBO process are defined by the sequence of stages that a company may follow in this process. Representing such sequences relies on functional role chains, and reduces to assessing the order of the intervals associated with the different stages. For example, representing that the `EarlyStage` always starts an LBO process can be represented as follows.

```
Class(LBOProcess_TS partial
 restriction(
  dataSomeValuesFrom(
   ConcreteFeatureChain(earlyStage time),
    time, starts)))
```

ABox

At ABox level we represent particular information that is known about the specific LBO process presented in this section. We start by instantiating the relevant individuals that are known to play a role in the LBO process. First, we represent the participating companies.

```
Individual(iAllianceBoots type(Target))
Individual(iKKR type(HedgeFund))
Individual(iTerraFirma type(HedgeFund))
```

For each of the hedgefunds involved, we instantiate a process and define its stages, such as in the case of TerraFirma.

```
Individual(iLBOProcess1_TS1
  type(LBOProcess_TS)
  value(timeSliceOf iLBOProcess_1)
  value(earlyStage iEarlyStage1_TS1)
  value(dueDiligence iDueDiligence1_TS1)
  value(bidding iBidding1_TS1)
  value(abort iAbort1_TS1))
```

Next, we represent the information contained by the individual news messages associated with the LBO process. We illustrate this by using a news message that describes the hedge fund `TerraFirma` entering the `EarlyStage` phase. Here, we only present a summary of the actual news message and indicate the stage that is signaled by it. The date and time associated to the news message is the one as specified on `http://www.marketwatch.com/`, and represents the time when the news message was issued and thus became available to the wide public.

Buyout firm Terra Firma mulls Boots bid

Sun Mar 25, 2007 8:42am EDT

This news message signals the beginning of the LBO, mentioning that Terra Firma is considering a bid for Alliance Boots (*EarlyStage*).

For representing the information contained in the news message we create a timeslice for the hedge fund and the target, respectively, a time interval associated to the stage, and use the `inStage` fluent to associate the companies to the stage.

```
Individual(t1 type(Interval))
Individual(iEarlyStage1
  type(EarlyStage_TS))
Individual(iEarlyStage1_TS1
  type(TimeSlice)
  value(timeSliceOf iEarlyStage1)
  value(time t1))
Individual(iAllianceBoots_TS1
  type(TimeSlice)
  value(timeSliceOf iAllianceBoots)
  value(time t1)
  value(inStage iEarlyStage1_TS1))
Individual(iTerraFirma_TS1
  type(TimeSlice)
  value(timeSliceOf iTerraFirma)
  value(time t1)
  value(inStage iEarlyStage1_TS1))
```

In this section we have shown how a dynamic process, in the form of an LBO process, can be represented in the tOWL language. Due to the temporal expressiveness of the language, we were able to define the order of stages of the considered process, as well as the transitions that the companies involved make through this process.

4.4.3 Reasoning Examples

In this section we present an example of how reasoning can be used in the LBO application previously described. We show how it can be inferred that a company is in a certain stage based on information on other stage transitions in which a company was involved.

The different paths that a company may follow when involved in an LBO process have been described in Figure 4.3. A set of axioms, as presented in Section 4.4.2, have been used to represent this in a tOWL knowledge base, exhaustively describing all acceptable stage sequences in an LBO process. Based on this knowledge, and in the presence of incomplete information, one can, in a certain number of cases, infer this missing information from the facts already present in the knowledge base. In this section we illustrate this by means of an example.

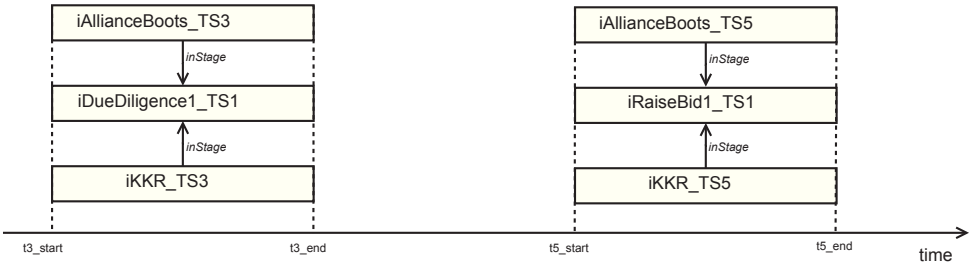

Figure 4.4: Overview of explicit knowledge on LBO stages

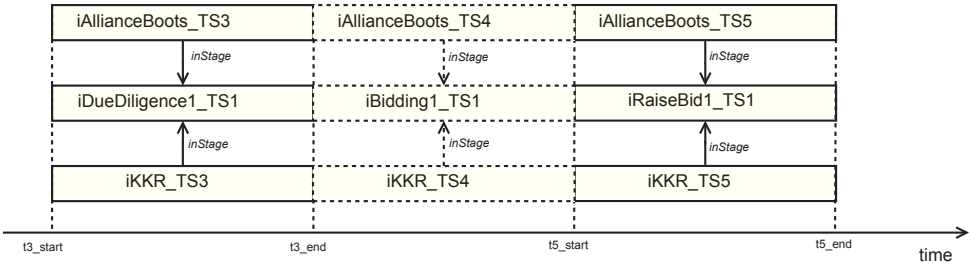

Figure 4.5: Overview of explicit and implicit knowledge on LBO stages

Assuming the existence of a news message, N_1, reporting that the company KKR, seeking to acquire Alliance Boots, has entered the Due Diligence stage of an LBO process,

we can represent the following information in the tOWL knowledge base, illustrated here by means of tOWL abstract syntax.

```
Individual(t3 type(Interval))
Individual(iDueDiligence1
  type(DueDiligence_TS))
Individual(iDueDiligence1_TS1
  type(TimeSlice)
  value(timeSliceOf iDueDiligence1)
  value(time t3))
Individual(iKKR_TS3
  type(TimeSlice)
  value(timeSliceOf iKKR)
  value(time t3)
  value(inStage iDueDiligence1_TS1))
Individual(iAllianceBoots_TS3
  type(TimeSlice)
  value(timeSliceOf iAllianceBoots)
  value(time t3)
  value(inStage iDueDiligence1_TS1))
```

The representation just introduced defines a new time interval, $t3$, that is associated with timeslices of the two companies involved in the Due Diligence phase as well as with a

timeslice of this stage. Additionally, we associate the timeslices of the two companies with the timeslice of the Due Diligence stage through the *inStage* fluent, thus indicating that, over interval $t3$, KKR and Alliance Boots find themselves in the Due Diligence phase.

Following the N_1 news message, another news message is issued, N_2, reporting that KKR and Alliance Boots have entered the Raise Bid stage of the LBO process, over some interval $t5$. The knowledge relating to these two stages in the tOWL knowledge base is depicted in Figure 4.4.

Having previously described all the possible paths through which an LBO process can be instantiated for a certain company, it is apparent that a direct transition from the Due Diligence phase to the Raise Bid phase is not possible. We derive that, between the two stages, the two companies must have transitioned the Bidding phase before moving on to Raise Bid, since no other path is possible between the two stages described in the news messages N_1 and N_2, respectively. We depict the new, relevant, snapshot of the knowledge base in Figure 4.5, where the inferred knowledge is represented in dotted lines.

4.5 Discussion

We have seen in Section 4.4 that the tOWL language can be used to represent rather complicated processes in which temporal aspects such as time and change play an important role. The tOWL language meets shortcomings of various previous approaches, such as OWL-Time (Hobbs and Pan, 2004) and the OWL ontology for fluents (Welty and Fikes, 2006) that only address temporality to a limited extent. The approach presented in (Hobbs and Pan, 2004), for example, only deals with the representation of time in the form of intervals and instants. However, ensuring that intervals are properly defined (e.g., starting point is always strictly smaller than the ending point) is not possible in this approach. Additionally, no support is offered for reasoning on the temporal constructs introduced other than the standard OWL-DL reasoning. For example, in OWL-Time it is also not possible to enforce a particular order of state transitions in a process.

The approach in (Hobbs and Pan, 2004) is limited in the representation of temporal aspects such as change. The approach taken in (Welty and Fikes, 2006) builds upon (Hobbs and Pan, 2004) by addressing this limitation, namely: the representation of temporal aspects such as change. However, it is limited in another sense which relates to the definition of fluent properties as being symmetric, i.e., if the pair (x,y) is the interpretation of a symmetric property, then the pair (y,x) is also an instance of this property. This is more often than not false, as in the very simple example of the *employeeOf* relation: although it holds that x is an employee of y, it is certainly not the case that y is also an employee of x. Therefore, enforcing symmetry on fluent properties is usually too restrictive.

Building upon the approach in (Welty and Fikes, 2006), tOWL enables differentiations between fluents that take values from the *TimeSlice* class and fluents that indicate changing values (data types). This is achieved through the use of the *FluentObjectProperty* and *FluentDatatypeProperty* properties. For the representation of time, the tOWL language relies on an approach based on concrete domains, thus enabling higher temporal expressiveness when compared to the approaches in (Hobbs and Pan, 2004) and (Welty and Fikes, 2006).

The approach proposed in this chapter can also provide a strong logical base for temporal extensions of conceptual languages. In the ER model (Chen, 1976), for example, the relationship between concepts can be represented as a fluent in order to denote a time-varying relationship. Adaptations of the fluent approach presented in this chapter, such as the differentiation between fluents relating to objects and fluents relating to data types, can help refine existing conceptual models where time is taken into account. For example, the approach in (Zimanyi et al., 1997), presenting the TERC+ temporal conceptual model, could incorporate such a refinement in the language specification. Additionally, the approach taken in tOWL for the representation of time could be incorporated in

temporal conceptual models when the representation of processes and state transitions is envisioned.

From an application perspective, our work comes to enable temporal representations in systems where this was not previously possible, such as often the case in designs based on computational intelligence methods combined with Semantic Web approaches. Many such applications have been developed, as for example the application of a fuzzy ontology to news summarization (Lee et al., 2005), an ontology-based computational intelligent multi-agent system applied to Capability Maturity Model Integration assessment (Lee and Wang, 2009), and project monitoring (Lee et al., 2008). Ontology-based approaches have also been applied in the development of systems based on computing with words approach (Marek and Ly, 2009). Fuzzy concept networks and their evolution are analysed in (Calegari and Farina, 2007), while the fuzzy matchmaking of Semantic Web services is described in (Fenza et al., 2008). A fuzzy markup language (FML) based on XML is applied in the context of an adaptive domotic framework in (Acampora and Loia, 2005a), and in the more general context of Ambient Intelligence, together with other fuzzy technologies, in (Acampora and Loia, 2005b). tOWL can enhance these systems by providing a formalism for the representation of time and change.

4.6 Conclusions and Future Work

The tOWL language is an extension of OWL-DL$^-$ that enables the representation and reasoning with time and temporal aspects. It comes to meet shortcomings of previous approaches, such as (Hobbs and Pan, 2004; Welty and Fikes, 2006) that only address this issue to a limited extent. It extends the OWL-DL$^-$ language with concrete domains, and enables class axioms that rely on binary concrete domain predicates that can also be used in combination with property chains. The language provides a concrete domain based on the set $\mathbb{Q}$ of rational numbers and the set of binary concrete domain predicates $\{<, \leq, =, \neq, \geq, >\}$. By means of syntactic sugaring we also introduce intervals and Allen's

Table 4.3: tOWL constructs

tOWL Abstract Syntax	DL Syntax	Semantics
A (URI Reference)	A	$A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$
`owl:Thing`	$\top$	$\Delta^{\mathcal{I}}$
`owl:Nothing`	$\bot$	$\{\}$
`intersectionOf`$(C_1\ C_2 ...)$	$C_1 \sqcap C_2$	$C_1^{\mathcal{I}} \cap C_2^{\mathcal{I}}$
`unionOf`$(C_1\ C_2 ...)$	$C_1 \sqcup C_2$	$C_1^{\mathcal{I}} \cup C_2^{\mathcal{I}}$
`complementOf`(C)	$\neg C$	$\Delta^{\mathcal{I}} \setminus C^{\mathcal{I}}$
`restriction`$(R$ `someValuesFrom`$(C))$	$\exists R.C$	$\{x \in \Delta^{\mathcal{I}} \mid \exists y.\langle x,y\rangle \in R^{\mathcal{I}}$ and $y \in C^{\mathcal{I}}\}$
`restriction`$(R$ `allValuesFrom`$(C))$	$\forall R.C$	$\{x \in \Delta^{\mathcal{I}} \mid \forall y.\langle x,y\rangle \in R^{\mathcal{I}} \rightarrow y \in C^{\mathcal{I}}\}$
`restriction`$(R$ `minCardinality`$(n))$	$\geq n\ R$	$\{x \in \Delta^{\mathcal{I}} \mid \sharp(\{y.\langle x,y\rangle \in R^{\mathcal{I}}\}) \geq n\}$
`restriction`$(R$ `maxCardinality`$(n))$	$\leq n\ R$	$\{x \in \Delta^{\mathcal{I}} \mid \sharp(\{y.\langle x,y\rangle \in R^{\mathcal{I}}\}) \leq n\}$
`restriction`$(U$ `someValuesFrom`$(D))$	$\exists U.D$	$\{x \in \Delta^{\mathcal{I}} \mid \exists y.\langle x,y\rangle \in U^{\mathcal{I}}$ and $y \in D^{\mathcal{D}}\}$
`restriction`$(U$ `allValuesFrom`$(D))$	$\forall U.D$	$\{x \in \Delta^{\mathcal{I}} \mid \forall y.\langle x,y\rangle \in U^{\mathcal{I}}$ and $y \in D^{\mathcal{D}}\}$
`restriction`$(U$ `minCardinality`$(n))$	$\geq n\ U$	$\{x \in \Delta^{\mathcal{I}} \mid \sharp(\{y.\langle x,y\rangle \in U^{\mathcal{I}}\}) \geq n\}$
`restriction`$(U$ `maxCardinality`$(n))$	$\leq n\ U$	$\{x \in \Delta^{\mathcal{I}} \mid \sharp(\{y.\langle x,y\rangle \in U^{\mathcal{I}}\}) \leq n\}$
`ConcreteFeatureChain`$(f_1\ \ldots\ f_n\ g)$	$f_1 \circ ... \circ f_n \circ g$	$\{(a_1,b) \in \Delta^{\mathcal{I}} \times \Delta_{\mathcal{D}} \mid \exists! a_2 \in \Delta^{\mathcal{I}}, ..., \exists! a_{n+1} \in \Delta^{\mathcal{I}} \wedge$ $\wedge\ \exists! b \in \Delta_{\mathcal{D}} : (a_1,a_2) \in f_1^{\mathcal{I}}, \ldots$ $(a_n,a_{n+1}) \in f_n^{\mathcal{I}} \wedge g^{\mathcal{I}}(a_{n+1}) = b\}$.
`restriction`$((u_1,u_2)$ `someValuesFrom`$(p_d))$	$\exists(u_1,u_2).p_d$	$\{x \in \Delta^{\mathcal{I}} \mid \exists! q_1 \in \Delta_{\mathcal{D}}, \exists! q_2 \in \Delta_{\mathcal{D}} : u_1^{\mathcal{I}}(x) = q_1 \wedge$ $\wedge\ u_2^{\mathcal{I}}(x) = q_2 \wedge\ (q_1,q_2) \in p_d^{\mathcal{I}}\}$
`restriction`$((u_1,u_2)$ `allValuesFrom`$(p_d))$	$\forall(u_1,u_2).p_d$	$\{x \in \Delta^{\mathcal{I}} \mid \forall q_1 \in \Delta_{\mathcal{D}}, \forall q_2 \in \Delta_{\mathcal{D}} : u_1^{\mathcal{I}}(x) = q_1 \wedge$ $\wedge\ u_2^{\mathcal{I}}(x) = q_2 \wedge\ (q_1,q_2) \in p_d^{\mathcal{I}})\}$

13 interval relations that may hold between intervals in the language. Additionally, a fluents approach is used for the representation of the different aspects of change, such as state transitions. Building on the approach presented in (Welty and Fikes, 2006), it extends the latter by making a difference between fluents that point to data types and fluents that point to objects, thus limiting the proliferation of objects inherent to this approach, since less timeslices are created in the case of data type fluents. The tOWL language can be used for representation and reasoning in a wide variety of dynamic domains, such as the financial one as shown in this chapter.

Future work will focus on further optimizing the representation for automated reasoning. Additionally, an investigation is required of how the meaning of different static language constructs, such as cardinality, changes when the temporal dimension is being considered.

Chapter 5

Temporal Optimizations and Temporal Cardinality in the tOWL Language [1]

The tOWL language is a temporal web ontology language based on OWL-DL without nominals. The language enables the representation of time and time-related aspects, such as state transitions. The design choices of the language pose new challenges from a temporal perspective. One such challenge is the representation of temporal cardinality. Another challenge consists of optimising the temporal representations in order to reduce the number of axioms. One such optimisation is temporal coalescing, which merges concepts that are associated with time intervals that either meet or share at least one instant with each other. In this chapter we formally introduce these concepts into the tOWL language and illustrate how they can be applied.

[1] An article based on this chapter has appeared in: Viorel Milea, Flavius Frasincar, Uzay Kaymak, Geert-Jan Houben. Temporal Optimizations and Temporal Cardinality in the tOWL Language. International Journal of Web Engineering and Technology, Volume 7, Number 1, pages 45–64, Inderscience Publishers, 2012.

5.1 Introduction

tOWL is a temporal Web ontology language based on the $\mathcal{SHIN}(\mathcal{D})$ description logic, a subset of OWL-DL (OWL-DL without nominals) (Milea et al., 2007, 2008a; Frasincar et al., 2010). It addresses shortcomings of OWL-DL (Patel-Schneider et al., 2004) by enabling temporal representations in ontologies. The tOWL language is focussed on the representation of concrete time, in the form of instants and intervals, and change, such as involved in the representation of processes. For the representation of concrete time, the tOWL language relies on a concrete domains extension to the language. For the representation of change, tOWL relies on a four-dimensional representation based on fluents and timeslices.

The language enables different temporal representations in ontologies, from more simple aspects such as the change in the attribute value of a property, to more complex aspects such as processes and the associated state transitions. Such representations are not possible in OWL-DL as this ontology language does not provide any semantics of time. Therefore, any representation of change or processes in OWL-DL will be limited by static semantics, and will not enable any inference related to the temporal aspects of the representation. In a tOWL representation of a process, for example, one could infer the state of the process in which an entity finds itself based on information of the states an entity has already transitioned.

In this chapter we develop the tOWL language further by discussing temporal coalescing in this context, and increasing the expressiveness of the language through temporal cardinality. Temporal coalescing is an operation similar to duplicate elimination in databases, applied to a temporal context (Bohlen et al., 1996). In the case of the tOWL language, this relates to merging timeslices that represent the same relation over time intervals that meet or share at least one instant. The motivation for temporal coalescing in tOWL knowledge bases is two-fold. One aspect relates to the proliferation of objects in the knowledge base, inherent to a timeslice approach where new timeslices are created

every time something is changing. By temporally coalescing the knowledge base, the number of timeslices can be greatly reduced. Another aspect relates to posing temporal queries upon the knowledge base.

Representing change by means of a 4d approach based on timeslices and fluents in the tOWL language also poses some interesting challenges to some of the static OWL-DL concepts. One such concept is the *cardinality* construct, that is used in OWL for restricting the number of attribute values that a property is allowed to take. For example, an imaginary *hasBiologicalFather* property in OWL-DL, indicating the individual representing the biological father of another individual, should have a cardinality of exactly one, as any person can have exactly one biological father. Such a cardinality construct is also relevant from a temporal context, where one might want to represent that, *at any point in time*, a person may have exactly one biological father. Introducing the temporal cardinality construct in tOWL increases the expressiveness of the language when regarded from a temporal context and allows for more accurate representations of the world.

Several optimizations are possible in the tOWL language, optimizations that can be related directly to the language or optimizations at the reasoner level. At language level, a possible optimization relates to making a distinction between datatype fluents and object fluents. By not requiring the creation of timeslices for datatypes everytime a concrete value is changing, the number of timeslices is considerably reduced. Another optimization at language level could involve the definition of the time for which timeslices hold as a union of intervals, thus preventing the creation of separate timeslices for each interval in the union and reducing the number of timeslices in the knowledge base. This latter optimization is not considered in the current work. At reasoner level we consider temporal coalescing as a technique for reducing the number of timeslices in the knowledge base and enabling a large number of queries to be correctly answered when posed on the coalesced knowledge base.

Temporal cardinality and temporal coalescence are also of practical relevance in a Semantic Web context. The concept of temporal cardinality, for example, can be used in a variety of applications, such as online retailers or corporate knowledge bases, to name a couple. In the case of online stores, one could envision restrictions posed in the store's knowledge base, such as each customer having exactly one customer number at any point in time. In corporate knowledge bases one might want to enforce the restriction that at any point in time, an employee has at least one direct manager. In the same context, temporal coalescing comes to increase the number of queries that can be correctly answered when posed upon a coalesced knowledge base. For example, one employee, e.g., consultant, works for a company for two years, during which period he has a temporary contract. In a tOWL knowledge base, timeslices are instantiated for this fact and the employment relationship is represented. After two years, the employee's contract is extended for an additional period of three years. In the tOWL knowledge base, new timeslices are created to represent this employment relationship for a period of three years. Provided the tOWL knowledge base is uncoalesced, queries posed by, for example, clients of the company on the company's website, seeking to retrieve the consultants with at least four years of experience in the company, will fail to retrieve the employee discussed in this paragraph, since no relationship describes the employment relationship for a period longer than three years. The coalesced knowledge base would enable this consultant to be recongnized as an individual working for the company for five years.

This chapter is organized as follows. In Section 5.2 we discuss work related to the research presented in this chapter. In Section 5.3 we present an overview of the tOWL language and illustrate how the envisioned optimizations and new constructs that we seek to introduce affect the language. In Section 5.4 we present temporal coalescing in the context of the tOWL language. Section 5.5 discusses temporal cardinality for the tOWL language. Finally, we conclude in Section 5.6.

5.2 Temporal Coalescing and Temporal Cardinality

In the context of the Semantic Web, a number of approaches have already been designed, addressing different temporal aspects in relation to ontology languages. A rather extensive approach towards extending ontology languages with a temporal dimension is Temporal RDF (Gutierrez et al., 2007). This work is similar to the tOWL language as it concerns the ability to represent temporal information in ontologies, but differs in that the language considered is the Resource Description Framework (RDF). Another approach is OWL-Time, which focusses on OWL rather than RDF. The initial purpose behind the design of a time ontology (OWL-Time) (Hobbs and Pan, 2004) was to represent the temporal content of Web pages and the temporal properties of Web Services. This approach is rather extensive in describing quantitative time and the qualitative relations that may exist among instants and intervals. Being based on OWL-DL, it uses the underlying $\mathcal{SHOIN}(\mathcal{D})$ description logic and thus relies on datatypes rather than concrete domains for the description of instants and intervals, while tOWL uses concrete domains for the representation of concrete time. The problem of change in ontologies has also been addressed in the context of ontology evolution, such as in (Noy and Klein, 2004) and (Haase and Stojanovic, 2005). However, the problem of ontology evolution relates to changing TBoxes, while tOWL focusses on representing change at ABox level. These approaches do not attempt to deal with concepts such as temporal cardinality or temporal coalescing.

The concept of temporal coalescing can best be illustrated with the example provided in tables 5.1 and 5.2. Table 5.1 describes John as being an employee of Yahoo over two adjacent intervals. Clearly, queries posed to this knowledge base of the form *employees of Yahoo who have worked for this employer for at least 3 years* cannot be answered correctly in the case of John, since both intervals associated with this description are shorter than 3 years. In such cases, and in cases where the intervals share at least one instant instead of being adjacent, in the presence of equivalent attribute values, the

Name	Employer	Time
John	Yahoo	[01/01/2001 14/02/2003)
John	Yahoo	[15/02/2003 16/07/2005)

Table 5.1: Uncoalesced relation

Name	Employer	Time
John	Yahoo	[01/01/2001 16/07/2005)

Table 5.2: Coalesced relation

relations can be summarized into a single relation without loss of information. The desired result is presented in Table 5.2.

A similar definition of coalescence is used in the XBit datamodel (Wang and Zaniolo, 2004), while a discussion of temporal coalescence in RDF is provided in (Grandi, 2009). The problem of temporal coalescence was addressed in the context of temporal databases (Bohlen et al., 1996; Dyreson, 2003). In (Bohlen et al., 1996), the authors address the coalescing of timestamped tuples where the attribute values are equal and the timestamps associated with the tuples either meet or overlap (for a formal description of the temporal relations that may exist between time intervals, such as meet and overlap, we refer the reader to (Allen, 1983)). As is the case for this chapter, in (Bohlen et al., 1996) the focus is on valid time. It should also be noted that the tOWL language, our current context for temporal coalescing, only focuses on valid time in the representation of temporal relations. The work in (Bohlen et al., 1996) is similar to parts of the work presented in this chapter, with the crucial difference that we address temporal coalescence in the context of timeslices connected by fluents, rather than tuples with identical attribute values. Thus, our focus is on those timeslices that represent the same individuals and are connected by equivalent fluents, and hold over temporal intervals that either meet or share at least one instant amongst them. Finally, an SQL implementation of temporal coalescence for Oracle is provided in (Bohlen et al., 1996).

Since tOWL allows different time granularities in the language, the approach described in (Dyreson, 2003) presents some overlap with the current work. The authors discuss the problem of temporal coalescing, and also take into account the possibility that the coalesced relations may be associated with timestamps at different levels of granularity.

Temporal cardinality is a constraint used to limit the number of values an attribute can have over the life-time of its entity (Zimanyi et al., 1997; Spaccapietra et al., 1998). The problem of expressing temporal cardinality in the context of the tOWL language has been preliminarily addressed in (Milea et al., 2008c). Here, the authors develop an approach for moving beyond the expression of fluent cardinality, enabled by OWL, and represent cardinality also when overlapping timeslices are involved. The issue of overlapping timeslices is also briefly discussed in (Welty and Fikes, 2006) in the context of a reusable OWL ontology for fluents. A distinction can be made in terms of temporal cardinality between snapshot cardinality, a temporal cardinality that holds over a limited period of time, and lifetime cardinality, a constraint holding across the whole lifetime of an entity (Tauzovich, 1991; Wijsen, 1999). The current work focuses on lifetime cardinality. Some discussion on the concept of snapshot cardinality is also provided in (Artale et al., 2010) in the context of the DL-Lite description logic. The same concept in the context of the Temporal ER model is described in (Gregersen and Jensen, 1999).

5.3 The tOWL Language

In this section we provide a general description of the tOWL language. This description is not meant to be exhaustive, and for more in-depth discussions of tOWL we refer the reader to (Milea et al., 2007, 2008a; Frasincar et al., 2010). This description is presented in Section 5.3.1. A discussion of the issues that we address in this chapter relative to the tOWL language is presented in Section 5.3.2.

5.3.1 General Description of the Language

The tOWL language (Milea et al., 2007, 2008a; Frasincar et al., 2010) is a temporal Web ontology language based on the $\mathcal{SHIN}(\mathcal{D})$ description logic, an expressive subfragment of OWL-DL (Patel-Schneider et al., 2004). An overview of the different layers introduced by the tOWL language on top of OWL-DL is provided in Figure 5.1. The language enables the representation of time and time-related aspects, such as change. For the representation of time, the tOWL language relies on concrete domains, and enables both instant-based as well as interval-based representations, as well as the relations that may exist between instants and intervals (such as Allen's 13 interval relations (Allen, 1983) in the case of intervals). For the representation of more complex aspects, such as change, the tOWL language is designed around a 4-dimensional view of the world. In this view, timeslices are used to represent otherwise static OWL individuals across temporal intervals, and fluents, a type of temporal property, are used to indicate what is changing. This design enables the representation of, for example, processes, and the associated state transition axioms. An example of how a complex process, in this case a leveraged buyout process, can be represented in the tOWL language is given in (Frasincar et al., 2010). The focus of the language is solely on valid time, i.e., the time when an axiom is true in the real world.

The timeslices-based representation has the ability to determine, at any point in time, what holds true. In order to use this representation, one has to create timeslices for the static individuals that are involved in a relation that is ephemeral in nature. For example, if one wants to represent the changing CEO of a company, and the ontology contains static individuals that represent both the person that is a CEO, as well as the company, then timeslices have to be instantiated for both these static concepts. Upon having done this, the two timeslices can be connected by a fluent, such as the *hasCEO* fluent, to indicate that, over the time interval associated with the timeslices, the two timeslices are in the *hasCEO* relationship. This is illustrated in Figure 5.2.

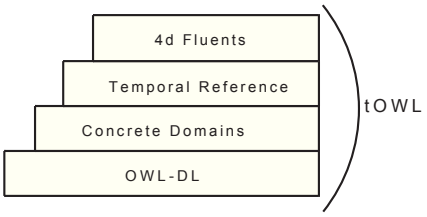

Figure 5.1: tOWL layer cake

In the example presented in Figure 5.2, two OWL classes have been defined, namely *Company* and *Person*. For each of these classes, one individual is instantiated, namely *iGoogle*, representing the company Google, an instance of the *Company* class, and *iEricSchmidt*, representing the individual with the name Eric Schmidt, an instance of the *Person* class. For each of these individuals, a timeslice is instantiated, namely *iGoogle_TS1* and *iEricSchmidt_TS1*, respectively. These timeslices both hold over the same interval, *iInterval1*, a consequence of the design of the tOWL language (fluents can only connect timeslices that hold over the same interval), and thus represent the static individuals with which they are associated over that interval. To denote that Eric Schmidt is the CEO of Google over the period denoted by *iInterval1*, we connect the two timeslices by the *hasCEO* fluent. In this way, we represent that Eric Schmidt was the CEO of Google over the interval *iInterval1* by using two timeslices and a fluent property.

5.3.2 Representational and Reasoning Issues in the tOWL Language

Temporal coalescing is an operation similar to duplicate elimination in databases, applied to a temporal context (Bohlen et al., 1996). In the case of the tOWL language, this relates to merging timeslices that represent the same relation over temporal intervals that either meet or share at least one instant with each other. Both the case of meeting timeslices, as well as the case of overlapping timeslices, are depicted in Figure 5.3.

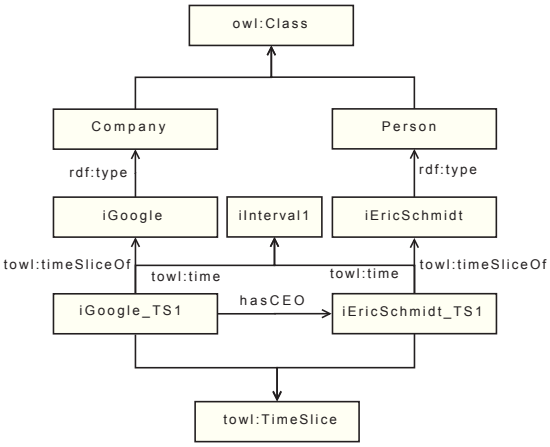

Figure 5.2: Representing change in tOWL

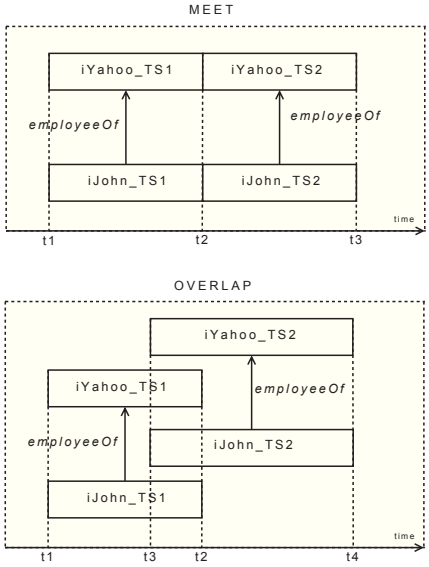

Figure 5.3: tOWL coalesce candidates

The desired result of temporal coalescing in the case of the tOWL language is depicted in Figure 5.4. Here we depict the cases of meeting as well as overlapping timeslices, and the resulting, merged timeslice. For the case of meeting timeslices, a new timeslice is created

holding from the beginning timestamp of the timeslice that appears chronologically first (t_1), until the ending timestamp of the timeslice that appears chronologically second (t_3). Upon removal of the two original timeslices, the knowledge base is simplified with no loss of information. The lack of loss of information comes from the design of the language, since timeslices are not reused in the representation of temporal information. Thus, when removing coalesced timeslices, we do not encounter references to other timeslices (such as a fluent relation between the removed timeslice and another timeslice in the knowledge base) that might be lost once the coalesced timeslices are removed. This is due to the fact that timeslices are created everytime something is changing, thus not relying on the reuse of timeslices for representation purposes. For this reason, we do not consider the problem of referential integrity (Widagdo, 2007; Steiner and Norrie, 1997) in our current approach, and assume that the timeslices that are coalesced have no references to/from other timeslices.

The same procedure applies to the case of overlapping timeslices, where the resulting, merged timeslice again holds from the beginning of the timeslice that appears chronologically first (t_1) until the end of the timeslice that appears chronologically second (t_4). A similar mechanism is applied when the intervals can be described by the equal, starts, finishes, and during Allen relationships, as well as the inverses of the latter three relationships.

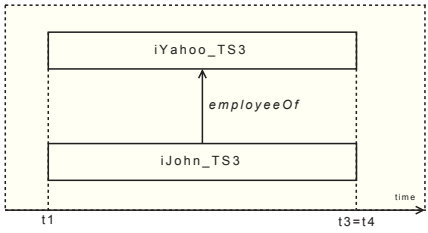

Figure 5.4: tOWL coalesced timeslices

The motivation for temporal coalescing in tOWL knowledge bases is two-fold. One aspect relates to the proliferation of objects in the knowledge base, inherent to a timeslice approach where new timeslices are created every time something is changing. By temporally coalescing the knowledge base the number of timeslices can be greatly reduced. Another aspect relates to posing temporal queries upon the knowledge base. In the case of the *employeeOf* relationship where the two timeslices meet, one could query whether John was an employee of Yahoo during the $[t_1, t_3)$ time interval. Since, in the original state, no relationship describes the link between John and Yahoo over the whole time interval, but rather subintervals hereof, this query would be evaluated as the empty set. The temporally coalesced version of this knowledge base correctly describes John as being an employee of Yahoo during the whole time interval, thus enabling more queries on the knowledge base than previously.

Representing change by means of a 4d approach based on timeslices and fluents in the tOWL language also poses some interesting challenges to some of the static OWL-DL concepts. One such concept is the *cardinality* construct, that is used in OWL for restricting the number of attribute values that a property is allowed to take. For example, an imaginary *hasBiologicalFather* property in OWL-DL, indicating the individual representing the biological father of another individual, should have a cardinality of exactly one, as any person can have exactly one biological father. Such a cardinality construct is also relevant from a temporal context. We might want to state, returning to our employee example, that John may be an employee of not more than one company, at a time. With this assumption in mind, we depict an example violating this temporal cardinality constraint in Figure 5.5. As can be observed from this depiction, there exists an interval $[t_3, t_2)$ over which John is both an employee of Yahoo, as well as Google, thus violating our temporal cardinality constraint stating that John may be an employee of at most one company, at any time (assuming that Google is different from Yahoo).

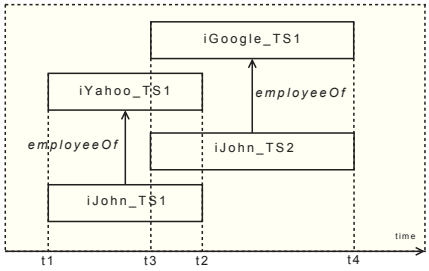

Figure 5.5: tOWL temporal cardinality example

The static cardinality construct that tOWL inherits from OWL-DL can be applied directly to the fluent *employeeOf*, stating that this fluent may take at most one value. However, as depicted in Figure 5.5, our cardinality constraint may be violated in a temporal context without violating the static cardinality constraint attached to the fluent, thus rendering the OWL-DL cardinality construct not expressive enough in a temporal setting.

The motivation for the representation of temporal cardinality in tOWL arises mainly from the increased expressiveness that this construct has to offer. Returning to the employee example, many such situations can be pictured where the concept of cardinality is not only relevant from the static perspective, but also from the temporal one. Increasing the expressiveness of the language with temporal cardinality will provide for more precise descriptions of the world, at least when regarded from a temporal point of view.

5.4 Temporal Coalescing in tOWL

In this section we introduce temporal coalescing in the tOWL language. Temporal coalescing implies merging timeslices that represent the same individual and are connected to timeslices of the same individual by the same fluent, over temporal intervals that either meet or share at least one instant amongst them. This operation is similar to duplicate elimination in databases (Bohlen et al., 1996). Temporal coalescing in tOWL has two de-

sirable consequences, namely the reduction of the proliferation of objects in the knowledge base, and resolving temporal queries in an appropriate way.

We introduce the concept of *individual equivalence*, which is illustrated in Figure 5.6, where the timeslices *iEricSchmidt_TS1* and *iEricSchmidt_TS2* are individual equivalent.

Definition 1 (Individual equivalence)

Two timeslices are individual equivalent (`ie`) if they are connected to the same static individual, i.e., they represent the same individual over arbitrary temporal intervals.

$$\begin{aligned} \texttt{ie}(m,n) \quad \equiv \quad & m,n \in \texttt{towl:TimeSlice}^{\mathcal{I}} \wedge (\exists p,q \in (\neg(\texttt{towl:TimeSlice} \sqcup \\ & \texttt{towl:Interval} \sqcup \texttt{rdfs:Literal}))^{\mathcal{I}}, (m,p) \in \texttt{towl:timeSliceOf}^{\mathcal{I}} \\ & \wedge (m,q) \in \texttt{towl:timeSliceOf}^{\mathcal{I}} \wedge (p,q) \in \texttt{owl:sameAs}^{\mathcal{I}}) \end{aligned}$$

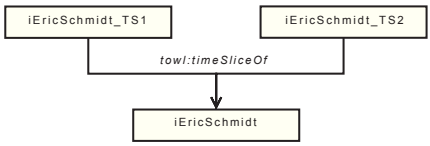

Figure 5.6: Individual equivalence in tOWL

We also introduce the concept of *fluent equivalence* and illustrate this concept in Figure 5.7, where *iEricSchmidt_TS1* and *iEricSchmidt_TS2* are fluent equivalent.

Definition 2 (Fluent equivalence)

Two timeslices are fluent equivalent (`fe`) if the values of any object fluent that is connected to these timeslices are individual equivalent and the values of any datatype fluent are the same.

$$\begin{aligned}\texttt{fe}(m,n) \equiv\ & m,n \in \texttt{towl:TimeSlice}^{\mathcal{I}} \wedge (\forall f \in \texttt{towl:FluentObjectProperty}^{\mathcal{I}} \\ & \exists p,q \in \texttt{towl:TimeSlice}^{\mathcal{I}}\ (m,p) \in f^{\mathcal{I}} \wedge (n,q) \in f^{\mathcal{I}} \wedge \texttt{ie}(p,q)) \wedge \\ & (\forall f \in \texttt{towl:FluentDatatypeProperty}^{\mathcal{I}}\ \exists p,q \in \texttt{owl:Datatype}^{\mathcal{I}} \\ & (m,p) \in f^{\mathcal{I}} \wedge (n,q) \in f^{\mathcal{I}} \wedge (p=q))\end{aligned}$$

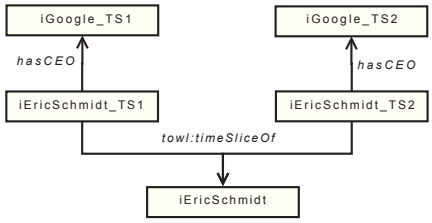

Figure 5.7: Fluent equivalence in tOWL

Finally, we introduce the concept of *temporal relatedness* as a temporal relationship between timeslices. This concept has already been illustrated in Figure 5.3, where, for example, the timeslices *iJohn_TS1* and *iJohn_TS2* are temporally related.

Definition 3 (Temporal relatedness)

Any pair of timeslices is temporally related (`tr`) if the intervals over which they are defined can be described by one of following Allen's interval relations: equal, meet, overlaps, starts, during, or finishes. We use only the direct Allen relations and not their inverses, as the order of timeslices m and n, and thus of variables p and q, is not important here (we can always swap these to obtain the direct relations if needed).

$$\begin{aligned}\texttt{tr}(m,n) \equiv\ & m,n \in \texttt{towl:TimeSlice}^{\mathcal{I}} \wedge (\exists p,q \in \texttt{towl:TimeInterval}^{\mathcal{I}} \\ & (m,p) \in \texttt{towl:time}^{\mathcal{I}} \wedge (n,q) \in \texttt{towl:time}^{\mathcal{I}} \wedge \\ & (p,q) \in \texttt{allen:meets}^{\mathcal{I}} \vee (p,q) \in \texttt{allen:overlaps}^{\mathcal{I}} \vee \\ & (p,q) \in \texttt{allen:starts}^{\mathcal{I}} \vee (p,q) \in \texttt{allen:finishes}^{\mathcal{I}} \vee \\ & (p,q) \in \texttt{allen:during}^{\mathcal{I}} \vee (p,q) \in \texttt{allen:equal}^{\mathcal{I}})\end{aligned}$$

Having defined individual equivalence, fluent equivalence and temporal relatedness, we can move on to introduce the concept of timeslices that are apt for temporal coalescence, i.e., timeslices that can be merged into a new timeslice representing the previous two timeslices over the union of the intervals over which the latter are defined.

Definition 4 (Binary aptness for temporal coalescence)
Any pair of timeslices is apt for temporal coalescing (`coal2`) if the timeslices are individual equivalent, fluent equivalent, and temporally related.

$$\texttt{coal2}(m,n) \equiv m,n \in \texttt{towl:TimeSlice}^{\mathcal{I}} \wedge \texttt{ie}(m,n) \wedge \texttt{fe}(m,n) \wedge \texttt{tr}(m,n)$$

Having defined aptness for temporal coalescence when two timeslices are involved, we can define this aptness for each individual timeslice, such that, given any timeslice in the knowledge base, it can be determined whether this timeslice can be merged with any other timeslice that satisfies the given conditions.

Definition 5 (Unary aptness for temporal coalescence)

Any individual timeslice is apt for temporal coalescing (`coal1`) if there exists some other timeslice such that the pair consisting of these two timeslices is apt for temporal coalescing.

$$\texttt{coal1}(m) \equiv m \in \texttt{towl:TimeSlice}^{\mathcal{I}} \wedge (\exists n \in \texttt{towl:TimeSlice}^{\mathcal{I}}\ \texttt{coal2}(m,n))$$

Provided that two timeslices, m and n satisfy the $\texttt{coal2}(m,n)$ relationship, we perform temporal coalescing according to the algorithm presented in Algorithm 2. Starting from these 2 timeslices, we begin by creating a new timeslice that will represent the coalesced relation over the merged time intervals, and attach the static individual describing m and n to the new timeslice. Additionally, we merge the intervals associated to m and n and attach the thus obtained interval to the new timeslice. Finally, we attach all fluents describing the 2 timeslices to the newly created timeslice, and complete the process by deleting the timeslices m and n that initiated the process.

Algorithm 2

($*$ Temporal coalescing algorithm $*$)

Input: m and n timeslices which are apt for temporal coalescing

Output: p timeslice which is obtained by coalescing m and n

1. create new timeslice *ts*
2. attach the static individual linked to m and n as static individual of *ts*
3. start(i) = min(start(time(m)), start(time(n)))
4. end(i) = max(end(time(m)), end(time(n)))
5. attach interval i as interval of *ts*
6. **for** all fluents f attached to m
7. **do** attach fluents f to *ts*

8. delete timeslices m and n

5.5 Temporal Cardinality in tOWL

The concept of cardinality in OWL-DL, which we denote as static cardinality, relates to the number of values that may be assigned to a property. It is used for representing the number of attribute values that may describe, through some static property, an OWL individual. In this static context, one might for example want to represent that a bicycle has exactly two wheels, namely the front wheel and the back wheel. Thus, when assigning attribute values to a property *hasWheels*, one must invariably assign both a front wheel and a back wheel to the individual being described. Static cardinality comes in three flavours. In addition to describing the exact cardinality one might also define the minimum cardinality, i.e., the minimum number of attribute values describing an individual through some property, or, conversely, the maximum cardinality. Formally, these three concepts can be described as in (Bechhofer et al., 2004):

- *minCardinality*: if stated to have the value a on a property P, with respect to a class C, then any instance of C will be related through P to at least a semantically distinct values (individuals or data values) (of which the type may further be restricted by the range of P);

- *maxCardinality*: if stated to have the value a on a property P, with respect to a class C, then any instance of C will be related through P to at most a semantically distinct values (individuals or data values) (of which the type may further be restricted by the range of P);

- *cardinality*: if stated to have the value a on a property P, with respect to a class C, then any instance of C will be related through P to exactly a semantically distinct values (individuals or data values) (of which the type may further be restricted by

the range of P). In other words, both a *minCardinality* of a and a *maxCardinality* of a are simultaneously satisfied.

The static cardinality implies the ability to determine when timeslices are equal, thus avoiding counting the latter more than once. Equality of timeslices can be determined by the following formula, where two timeslices m, n are considered to be equal if they hold over the same interval and describe the same static individual.

$$\begin{aligned}
\texttt{equal}(m,n) \equiv\ & (\exists o,p \in \texttt{towl:TimeInterval}^{\mathcal{I}}\ (m,o) \in \texttt{towl:time}^{\mathcal{I}} \wedge \\
& (n,p) \in \texttt{towl:time}^{\mathcal{I}} \wedge (o,p) \in \texttt{allen:equal}^{\mathcal{I}}) \wedge \\
& (\exists q,r \in \texttt{owl:Class}^{\mathcal{I}}\ (m,q) \in \texttt{towl:timeSliceOf}^{\mathcal{I}} \wedge \\
& (n,r) \in \texttt{towl:timeSliceOf}^{\mathcal{I}} \wedge (q,r) \in \\
& (\neg(\texttt{towl:TimeSlice} \sqcup \texttt{towl:Interval} \sqcup \texttt{rdfs:Literal}))^{\mathcal{I}})
\end{aligned}$$

In a temporal context in the tOWL language, the concept of static cardinality can be used to describe the cardinality of a fluent. Returning to the example presented in Figure 5.2, the concept of static cardinality can be applied to the fluent *hasCEO*, stating that this fluent can point to exactly one individual, since a company may have only one CEO. This restriction would thus be violated when we assign, for example, two different CEO's to the same fluent holding for the same timeslice. This is graphically depicted in Figure 5.8, where the left-hand side represents a non-violating example of the static cardinality applied to a fluent, and the right-hand side represents a violation of this restriction.

Thus, a straight-forward extension of the static cardinality concept might be applied directly to fluents in a temporal setting, instead of static properties, and maintain its semantics in the context of the tOWL language. However, due to the timeslices representation used in tOWL, this use of the static cardinality concept proves insufficient. For

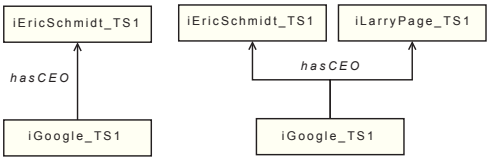

Figure 5.8: Static cardinality applied to fluents

illustrating why such a concept is insufficient, we consider the example already presented in Figure 5.5. Translating this to our current *hasCEO* example results in the illustration presented in Figure 5.9.

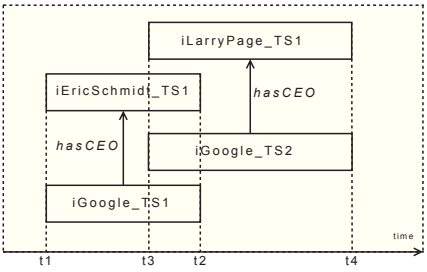

Figure 5.9: Violation of *hasCEO* temporal cardinality constraint

As can be seen from this figure, the fluent cardinality constraint is not violated, as, at all times, the *hasCEO* fluent is pointing to exactly one individual. However, we can clearly see that over the time interval $[t_3, t_2)$, the company Google has two CEOs, which is against the spirit of the cardinality contraint that we envision (a company may have exactly one CEO at a time). Thus, when discussing the concept of cardinality in a temporal context, we deem it necessary to make a distinction between two types of temporal cardinality.

1. *Fluent cardinality*: the (static) cardinality of the *hasCEO* fluent should be equal to one, following the description above. In other words, the *hasCEO* fluent must be associated to exactly one timeslice (of a static individual of type *Person*) each time it is defined for a timeslice of an individual of type *Company*. This issue can easily

be addressed by using the OWL-DL cardinality construct, as shown in (Welty and Fikes, 2006).

2. *Overlapping timeslices*: the (temporal) cardinality of the *hasCEO* fluent should be equal to 1. In other words, at any point in time, the *hasCEO* relation must be described by one timeslice of a static individual of type *Person*.

Clearly, simply addressing fluent cardinality in a temporal context is insufficient for expressing truly temporal cardinality constraints in tOWL. Therefore, we seek to extend the static cardinality constructs presented in this section in a temporal setting, and thus introduce the concepts of *temporalMinCardinality*, *temporalMaxCardinality*, and *temporalCardinality*. We define these concepts as follows.

Definition 1 (temporalMinCardinality)

Given a fluent property f, and a value a such that $a \in \mathbb{N}$, we represent by *temporalMinCardinality(f,a)* the restriction on timeslices of an arbitrary individual so that f is defined such that, at any point in time, the number of individuals that are associated to timeslices referred to by the fluents f is **at least** a.

Definition 2 (temporalMaxCardinality)

Given a fluent property f, and a value a such that $a \in \mathbb{N}$, we represent by *temporalMaxCardinality(f,a)* the restriction on timeslices of an arbitrary individual so that f is defined such that, at any point in time, the number of individuals that are associated to timeslices referred to by the fluents f is **at most** a.

Definition 3 (temporalCardinality)

Given a fluent property f, and a value a such that $a \in \mathbb{N}$, we represent by *temporalCardinality(f,a)* the restriction on timeslices of an arbitrary individual so that f is defined such that, at any point in time, *temporalMinCardinality(f,a)* and *temporalMaxCardinality(f,a)*

simultaneously hold.

We next focus on giving a formal semantic representation of the three types of temporal cardinality we introduced. In achieving this, we first define a function g that, given a fluent f, a static individual i and a point in time t, returns the number of different individuals j, for which f refers from a timeslice of i to a timeslice of j, for an interval that includes t. The result of this function is a natural number, obtained by counting the unique individuals obeying the above constraints, as shown next.

$$
\begin{aligned}
g_{(f,i,t)} \;=\; & |\{j \in (\neg(\texttt{towl:TimeSlice} \sqcup \texttt{towl:Interval} \sqcup \texttt{rdfs:Literal}))^{\mathcal{I}} \mid \\
& \forall x \in \texttt{towl:TimeSlice}^{\mathcal{I}} \exists y \in \texttt{towl:TimeSlice}^{\mathcal{I}} \wedge \\
& s \in \texttt{xsd:dateTime}^{\mathcal{I}} \wedge e \in \texttt{xsd:dateTime}^{\mathcal{I}} \wedge \\
& p \in \texttt{towl:TimeInterval}^{\mathcal{I}} \wedge \\
& (x,i) \in \texttt{towl:timeSliceOf}^{\mathcal{I}} \wedge (y,j) \in \texttt{towl:timeSliceOf}^{\mathcal{I}} \wedge \\
& (x,y) \in f^{\mathcal{I}} \wedge (y,p) \in \texttt{towl:time}^{\mathcal{I}} \wedge (p,s) \in \texttt{towl:start}^{\mathcal{I}} \wedge \\
& (p,e) \in \texttt{towl:end}^{\mathcal{I}} \wedge s \leq t < e\}|
\end{aligned}
$$

This function enables the definition of the three temporal constructs we seek to introduce in the language, namely *temporalMinCardinality*, *temporalMaxCardinality*, and *temporalCardinality*. In the following, $\geq_{\mathcal{T}}$, $\leq_{\mathcal{T}}$, and $=_{\mathcal{T}}$ denote the three constructs we introduce, a is a natural number larger than 0, f denotes a fluent and t denotes a point in time. It should be noted that the definition of temporal cardinality includes (is stronger than) the definition of static cardinality.

$$
\begin{aligned}
(\geq_{\mathcal{T}} a\ f)^{\mathcal{I}} \;=\; & \{x \in \texttt{towl:TimeSlice}^{\mathcal{I}} \mid \exists i \in (\neg(\texttt{towl:TimeSlice} \sqcup \\
& \texttt{towl:Interval} \sqcup \texttt{rdfs:Literal}))^{\mathcal{I}}, \\
& \exists p \in \texttt{towl:TimeInterval}^{\mathcal{I}}, \exists s, e \in \texttt{xsd:dateTime}^{\mathcal{I}} \wedge \\
& (x, i) \in \texttt{towl:timeSliceOf}^{\mathcal{I}} \wedge (x, p) \in \texttt{towl:time}^{\mathcal{I}} \wedge \\
& (p, s) \in \texttt{towl:start}^{\mathcal{I}} \wedge (p, e) \in \texttt{towl:end}^{\mathcal{I}} \wedge \\
& \forall t \in \texttt{xsd:dateTime}^{\mathcal{I}}, s \leq t < e, g_{(f,i,t)} \geq a\}
\end{aligned}
$$

$$
\begin{aligned}
(\leq_{\mathcal{T}} a\ f)^{\mathcal{I}} \;=\; & \{x \in \texttt{towl:TimeSlice}^{\mathcal{I}} \mid \exists i \in (\neg(\texttt{towl:TimeSlice} \sqcup \\
& \texttt{towl:Interval} \sqcup \texttt{rdfs:Literal}))^{\mathcal{I}}, \\
& \exists p \in \texttt{towl:TimeInterval}^{\mathcal{I}}, \exists s, e \in \texttt{xsd:dateTime}^{\mathcal{I}} \wedge \\
& (x, i) \in \texttt{towl:timeSliceOf}^{\mathcal{I}} \wedge (x, p) \in \texttt{towl:time}^{\mathcal{I}} \wedge \\
& (p, s) \in \texttt{towl:start}^{\mathcal{I}} \wedge (p, e) \in \texttt{towl:end}^{\mathcal{I}} \wedge \\
& \forall t \in \texttt{xsd:dateTime}^{\mathcal{I}}, s \leq t < e, g_{(f,i,t)} \leq a\}
\end{aligned}
$$

$$
(=_{\mathcal{T}} a\ f)^{\mathcal{I}} \;=\; (\geq_{\mathcal{T}} a\ f)^{\mathcal{I}} \cap (\leq_{\mathcal{T}} a\ f)^{\mathcal{I}}
$$

Returning to the *hasCEO* example presented in this section, the newly introduced concepts can be used for representing the fact that a company may only have one CEO at a time. For representing this example we rely on tOWL abstract syntax. Upon defining the two static OWL classes, `Company` and `Person`, we define the class of all timeslices of a person, namely `Person_TS`. The class `Company_TS` is defined as the class of all timeslices of an individual of type `Company` for which the `hasCEO` fluent takes a value that is a timeslice of `Person`. Finally, the temporal cardinality of the `hasCEO` fluent is defined to be equal

to 1, i.e., one company may have exactly one CEO at a time. It should be noted that, since the formal definition of temporal cardinality includes static cardinality, there is no need to additionally define the static cardinality of the `hasCEO` fluent.

```
Class(Company)
Class(Person)

Class(Person_TS partial TimeSlice
   restriction(timeSliceOf someValuesFrom(Person)))

Class(Company_TS partial TimeSlice
   restriction(timeSliceOf someValuesFrom(Company))
   restriction(hasCEO someValuesFrom(Person_TS))
   restriction(hasCEO temporalCardinality(1)))
```

Figure 5.10: Using temporal cardinality in tOWL abstract syntax

5.6 Conclusions and Future Work

In this chapter we have introduced two novel concepts in the tOWL language: temporal coalescing and temporal cardinality. The first concept, temporal coalescing, ensures a reduction of the proliferation of objects in the knowledge base, while also ensuring that a larger number of temporal queries can be resolved than previously. Temporal cardinality comes to address the limitations of the concept of static (OWL-DL) cardinality when timeslices are involved. Rather than focussing on the range of a property, the temporal cardinality introduced in tOWL involves overlapping timeslices that may violate a cardinality constraint, such as a company with two CEOs at a moment in time.

Temporal coalescence, as introduced in this chapter in the context of the tOWL language, is a reasoner-based optimization aimed at reducing the number of timeslices contained in the knowledge base. Although this does not impact the representational power available to the user, we deem this optimization relevant in a practical context, where the size of the knowledge is of crucial importance in determining the speed of the applications

based on the tOWL language. Also, it allows to answer correctly more temporal queries than using tOWL without this optimization.

As future work we focus on further extending the expressiveness of the tOWL language in a temporal context, as well as the optimizations that may be envisioned in such a context. A possible research direction is introducing multi-dimensional time (e.g., transaction time, user-defined time, etc.) in the language. An example of an optimization that we plan to investigate is the ability to relate multiple intervals to the same timeslice.

The tOWL language and the lessons learned in the first two chapters of this dissertation regarding the topic of news-based algorithmic trading, provide a good foundation for the development of truly time-aware financial decision support systems. The next chapter is focused on a general framework that is better suited at dealing with the dynamic nature of financial markets.

Chapter 6

A General Framework for Time-Aware Decision Support Systems [1]

In this chapter we present a time-based framework for decision support systems. The system uses the state-of-the-art tOWL language for the representation of temporal knowledge and enables temporal reasoning over the information that is represented in the knowledge base. Our approach is the first to use state-of-the- art Semantic Web technology for handling temporal data. We illustrate the applicability of our framework by building a market recommendations aggregation system. This system automatically collects market recommendations from online sources, and, based on the past performance of the analysts that issued a recommendation, generates an aggregated recommendation in the form of a buy, hold, or sell advice. We illustrate the flexibility of our proposed system by implementing multiple methods for the aggregation of market recommendations.

[1] An article based on this chapter has appeared in: Viorel Milea, Flavius Frasincar, Uzay Kaymak. A Time-Based Framework for Decision Support Systems. Expert Systems with Applications. DOI information: 10.1016/j.eswa.2012.08.001

6.1 Introduction

Decision systems often rely on historical information for the formulation of a best course of action. With the increasing number of information sources, today's systems must deal with large volumes of data. Storing, retrieving and checking this information for consistency represents one of the main challenges in building decision support systems that use historical data.

Although flat representations of data used in, for example, business intelligence, such as databases, provide intelligent storage and retrieval of data (Batini et al., 1991), automated inference, as needed in consistency checks, is limited in approaches based on such formalisms due to the rather inexpressive semantics of the underlying structures. Modern knowledge representation approaches provide for more finely grained semantics and additional expressiveness from a semantic perspective. From these approaches, Semantic Web (Lee et al., 2001; Shadbolt et al., 2006) languages such as Resource Description Framework (RDF) and RDF Schema (RDFS) (Klyne and Carroll, 2004; Brickley and Guha, 2004), and the Web Ontology Language (OWL) (Patel-Schneider et al., 2004; Motik et al., 2009) provide the most expressive choices when the problem of automated inference is considered. When historical data is used, some time-enabled formalism is required. In this context, an approach such as the tOWL language (Milea et al., 2012a,b), a temporal web ontology language based on OWL, can provide the appropriate formalism for the representation of time-varying knowledge.

These properties of Semantic Web languages make them suitable for use in decision support systems where some of the intelligence of the system is already incorporated in the representation language and does not need to be explicitly accounted for in the main system. For example, in a system where future product prices are predicted based on past prices, the restriction that a product may only have one price at any point in time can be enforced generically (through a temporal cardinality restriction), at the level of the whole knowledge base, and outside the main application. This can be achieved by using a

temporal reasoner that is able to check the knowledge base for consistency, thus eliminating the need to incorporate such checks in the main system. In this way, the designer of the system can focus on the application intelligence rather than enforcing/checking data related restrictions manually. Also, there is an increased support for reuse of temporal reasoning tools across applications.

By relying on the state-of-the-art tOWL language, in this chapter we propose a framework for designing semantic, temporal decision support systems. The systems that we propose provide the means to efficiently store and retrieve data, and allow (temporal) inference on the represented data. Restrictions on this data can be represented in a generic way at the level of the temporal language. Different (temporal) properties of the entities in the knowledge base can also be represented, both at abstract as well as concrete level. In this way, data-related operations are separated from the main intelligence component of the decision support system.

The framework that we introduce is deployed in a practical context. We illustrate how such semantic, time-enabled decision support systems can be used by means of an example from the financial domain. More precisely, we choose the aggregation of market recommendations as a proof-of-concept. Market recommendations are advices, in the form of indicated courses of action, issued by financial analysts, regarding the stock of a certain company. These recommendations most often materialize in *buy*, *hold*, or *sell* advices, and are issued at different times. Since multiple analysts can issue such recommendations, more often than not, at a specific point in time, a company may have recommendations issued by different analysts. When these recommendations diverge, in the sense that there is no consensus within the analyst group whether an asset should be bought, held, or sold, choosing the appropriate course of action might not be obvious. The system that we present investigates which aggregation method for market recommendations gives the best results.

In investigating which approach provides the best results for the aggregation of market recommendations, we consider two different alternatives: a *majority voting* approach, in which we choose the recommendation to which most analysts concur, and an approach that takes into account the *analysts' past performance* when deciding the course of action with the highest expected performance. For measuring the analysts' past performance, we rely on the Sharpe ratio (Sharpe, 1966).

The example that we present contains several features that make it interesting to consider. First, the recommendations issued by analysts either have a limited validity in time, or hold until a new recommendation is issued. By relying on the tOWL language for the representation of recommendations, we can define default durations for advices, and also set as ending valid time for a recommendations the time when a new recommendation is issued by the same analyst, regarding the same company. Determining which recommendations hold at any point in time, regarding one or more companies, can also be achieved due to the timeslices representation used by the tOWL language. Last, this enables determining, at any point in time, which recommendations have been issued by an analyst in the past, which in turn enables determining the past performance of analysts.

The outline of the chapter is as follows. In Section 6.2 we provide an overview of research related to our current endeavour. The tOWL language, one of the main building bricks of our approach, is presented in Section 6.3. The semantic, time-enabled framework that we propose is presented in Section 6.4. The application, as well as the methodology we use for aggregating market recommendations is presented in Section 6.5. Our results and a discussion of the results are presented in Section 6.6. Finally, we conclude in Section 6.7.

6.2 Related Work

A thorough approach towards the design and evaluation of temporal expert systems is presented in (Chinn and Madey, 1997). The approach starts by evaluating specific char-

acterics of expert systems and temporal applications separately, and then formulating a framework that brings both of them together. The application area that is considered for this framework consists of business problems. One of the temporal requirements that the authors formulate relates to support for a time-line view of events as well as the ability to maintain a historical repository of events, requirements that are both supported by the tOWL language. Other requirements formulated in (Chinn and Madey, 1997) relate to being able to define and use time in different knowledge base constructs and the ability to represent temporal relationships. By relying on the tOWL language for the representation of temporal knowledge, these requirements are fulfilled due to the ability to represent intervals and Allen's interval relationships (Allen, 1983) in the language, as well as timeslices and fluents for representing what is changing.

The importance of a temporal dimension in knowledge bases is also identified in (Juarez et al., 2008) in the context of context-dependent temporal diagnosis. The temporal dimension is also considered in the context of recommender systems in (Arroyo-Figueroa et al., 1998). The temporal dimension in medical expert systems is discussed in (Kohane, 1987), while temporal reasoning in expert systems in a more general sense is discussed in (Perkins and Austin, 1990).

Although the temporal dimension in decision support systems has been investigated to a certain extent in the literature, approaches relying on Semantic Web technologies have not yet been considered. Semantic Web languages, such as the state-of-the-art tOWL language, fulfil the domain-independent requirements formulated until now in different studies.

In the last part of this section, we focus on work related to the proof-of-concept application that is used to illustrate the framework that we introduce. We give an overview of work related to market recommendations and their influence on the value of stocks.

The estimation of abnormal returns by using everyday portfolio balancing based on the consensus of market recommendations is researched in (Barber et al., 2001). Profit can

be generated by buying the most recommended stocks and selling the less favoured ones. In (Bjerring et al., 1983) the authors conclude that following the advices given by the broker provides a better result than following the TSE 300 or the S&P 500. In (Bjerring et al., 1983), (Morgan and Stocken, 2003) and (Barber et al., 2001), it is stated that investors can yield abnormal returns by following recommendations, although in (Barber et al., 2001) investors yield these high returns only when following recommendations with high consensus among the analysts. We note that (Ramnath et al., 2006) have done a large literature study of 250 papers. The authors suggest that no individual broker has enough information to always give correct advices. Traders will aggregate advices given by brokers and other information about a certain company to make a decision to buy, hold or sell a stock.

Previous studies have thus shown that market recommendations do have an impact on developments regarding stock prices, i.e., abnormal returns can be yielded by following stock advices. Additionally, consensus plays an important role in the lucrativeness of these advices, and a meaningful way of aggregating the individual advices might lead to improved results in terms of abnormal returns. Last, taking into account the possible bias of brokers can further help improve the performance of an investment strategy based on market consensus.

6.3 The tOWL Language

The tOWL language (Milea et al., 2007, 2008a; Frasincar et al., 2010; Milea et al., 2012a,b) is a temporal web ontology language based on the $\mathcal{SHIN}(\mathcal{D})$ description logic, which is an expressive subfragment of OWL-DL (McGuinness and Van Harmelen, 2004). tOWL is built on top of OWL-DL, the current state-of-the-art ontology language and W3C standard. An overview of the different layers introduced by the tOWL language on top of OWL-DL is provided in Figure 6.1. The language enables the representation of time and time-related aspects, such as change. For the representation of time, the tOWL language

relies on concrete domains, and enables both instant-based as well as interval-based representations, as well as the relations that may exist between instants and intervals (such as Allen's 13 interval relations (Allen, 1983) in the case of intervals). For the representation of more complex aspects, such as change, the tOWL language is designed around a 4-dimensional view of the world. In this view, timeslices are used to represent, otherwise static, OWL individuals across temporal intervals, and fluents are used to indicate what is changing. This design enables the representation of, for example, processes, and the associated state transition axioms. An example of how a complex process, in this case a leveraged buyout process, can be represented in the tOWL language is given in (Frasincar et al., 2010). The focus of the language is solely on valid time, i.e., the time when an axiom is true in the real world.

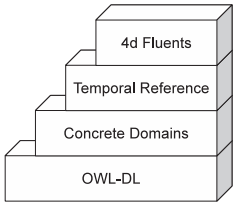

Figure 6.1: tOWL layer cake

The timeslice-based representation has the ability to determine, at any point in time, what holds true. In order to use this representation, one has to create timeslices for the static individuals that are involved in a relation that is ephemeral in nature. For example, if one wants to represent market recommendations issued for a company, and the ontology contains static individuals that represent both the analyst issuing the recommendation, as well as the company, then timeslices have to be instantiated for both these static concepts. Upon having done this, the two timeslices can be connected by a fluent, such as the *hasRecommendation* fluent, to indicate that, over the time interval associated with the

timeslices, the two timeslices are in the *hasRecommendation* relationship. This example is illustrated in Figure 6.2.

In the example presented in Figure 6.2, two OWL classes have been defined, namely *Company* and *Analyst*. For each of these classes, one individual is instantiated, namely *iIBM*, representing the company IBM, an instance of the *Company* class, and *iBarclays*, representing the Barclays bank, an instance of the *Analyst* class. For each of these individuals, a timeslice is instantiated, namely *iIBM_TS1* and *iBarclays_TS1*, respectively. These timeslices both hold over the same interval, *iInterval1*, a consequence of the design of the tOWL language (fluents can only connect timeslices that hold over the same interval), and thus represent the static individuals with which they are associated over that interval. To denote that Barclays has a recommendation for IBM over the period denoted by *iInterval1*, we connect the two timeslices by the *hasRecommendation* fluent.

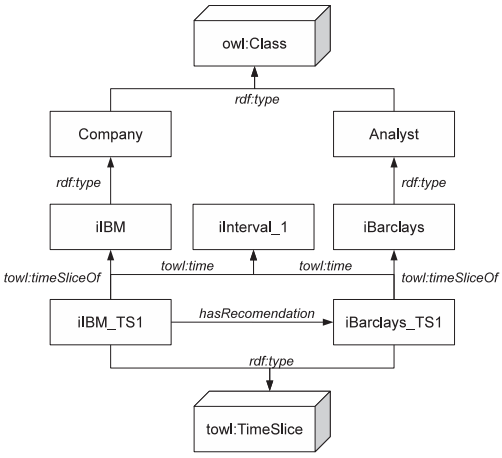

Figure 6.2: Representing change in tOWL

In this chapter, we use the tOWL language both for the representation of static knowledge, such as company names and ticker symbols, as well as temporal knowledge, such as the interval for which an advice holds true. The tOWL knowledge is thus used for representing all information that is considered relevant for the system, and forms the

backbone of all experiments that we perform. The tOWL formalism supports the representation of market recommendations and the time they are valid through intervals and Allen's temporal relationships between intervals. Additionally, the timeslice-based representation supports determining which advices hold true at any point in time.

6.4 The Framework

In this section we describe the general architecture of the temporal framework for decision support systems that we propose. The different subsections discuss the design considerations and functionality of each component. Additionally, we describe the design of each component for the case study used in this chapter, i.e., the aggregation of market recommendations. The general architecture of the system is presented in Figure 6.3.

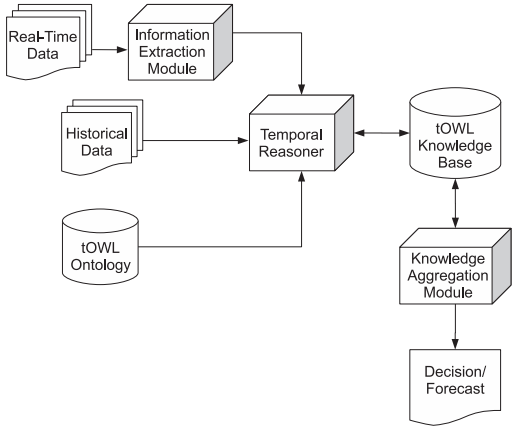

Figure 6.3: General architecture of the time-enabled decision support system

The input of the system we propose consists of three main sources: *real-time data*, *historical data*, and a *temporal domain ontology* (designed in the tOWL language) describing the input data at an abstract level. Assuming that the real-time data is extracted from a raw source, i.e., the data is not annotated, the *information extraction module* is used to extract the relevant information from the source. All input information is fed

to the *temporal reasoner* for consistency checks and the update of the *tOWL knowledge base*. The *knowledge aggregation module* uses data from the tOWL knowledge base for the generation of a forecast or recommended decision - the output of the system.

6.4.1 Information Extraction Module

The purpose of the information extraction module is the retrieval of information from sources that provide data that is not annotated. The motivation behind such a component is that most information is provided in raw format, and quick processing of this data is crucial for timely forecasts and, generally speaking, decision systems. This module can consist of various components aimed at the processing of raw data. When raw text data is considered, components such as a part-of-speech tagger can be considered, as well as components that rely on patterns for the extraction of knowledge. When the textual data is only analysed at a superficial level, different content analysis components can be considered for implementation in the information extraction module.

For the application studied in this chapter, we rely on Analist.nl, a Dutch language source for international market recommendations. Extracting the information contained in the advices consists of two main parts, namely: Part-of-Speech (POS) tagging and pattern-based extraction. The POS-tagger annotates every word in an advice with the part-of-speech it represents from the following eight categories: verbs, nouns, pronouns, adjectives, adverbs, prepositions, conjunctions and interjections (Banko and Moore, 2004). There are different implementations available of POS-Taggers, such as Microsoft's HMM Tagger (Banko and Moore, 2004) and the Stanford POS-Tagger (Toutanova et al., 2003; Toutanova and Manning, 2000).

Due to the fact the we use a non-English source for the advices, a Dutch POS-tagger is used, namely TreeTagger (Schmid, 1995), which implements the probabilistic tagging method as explained in (Schmid, 1997). An example of a non-tagged advice in XML as

extracted from Analist.nl is provided in Figure 6.4, while the tagged advice is displayed in Table 6.1.

```
<item>
<title>RBC Capital Markets: APPLE INC. kopen.</title>
<link>http://rss.feedsportal.com/c/637/f/8254/s/1047096/stor□01.htm</link>
<description>(Analist.nl) Toronto - Op 6-5-2008 herhalen de analisten van RBC Capital
Markets hun koopadvies voor APPLE INC. (ISIN: US0378331005 / TICKER: AAPL).Het
12-maands koersdoel voor APPLE INC. wordt opwaarts bijgesteld van 200.00 USD naar
220.00 USD. In 2006 bedroeg.. (lees verder op: www.analist.nl)</description>
<guid isPermaLink="false">http://www.analist.nl/rss.php?id=51910</guid>
</item>
```

Figure 6.4: RSS feed of advice *51910*

Using the example presented in Table 6.1, we proceed to extract the information as follows. The first cardinal numeric value is the issuing date. The verb following the cardinal number indicates whether the advice is upgraded, downgraded, or held constant. Next, we search for the advice type, which represents words like “kopen”, “houden”, or “verkopen” (or any of their synonyms), i.e., buy, hold, or sell. In case of an upgrade/downgrade, search for a second occurrence of these words “kopen”, “houden” or “verkopen” (or any of their synonyms). The second occurrence of the advice type indicates the new recommendation in the case of an upgrade or downgrade. The price target is found by searching for the first cardinal number after the word “koersdoel”. If the price target is followed with the word “niet”, there is no price target. In case of upgrading or downgrading the price target, the final price target will be preceded with the word “naar”. The ISIN identifier can be found by searching for the word ISIN and extracting the cardinal number following this word. The ISIN number is used for uniquely identifying a company. Finally, we extract the issuer of the advice. This is done by searching for the first occurrence of the word “van”. Since the length of the name of the analyst is unknown, all the words between the first occurrence of the word “van” after the word “analist” and the next occurrence of a

Table 6.1: Advice 51910 POS-tagged by TreeTagger

Word	POS	Lemma
Op	nounsg	<unknown>
6-5-2008	num__card	@card@
herhalen	verbprespl	herhalen
de	det__art	de
analisten	nounpl	analist—analiste
van	prep	van
RBC	adj	<unknown>
Capital	nounsg	<unknown>
Markets	nounpl	<unknown>
hun	det__poss	hun
koopadvies	nounpl	<unknown>
voor	prep	voor
APPLE	adj	<unknown>
INC	nounpl	<unknown>
ISIN	nounsg	<unknown>
US0378331005	nounsg	<unknown>
AAPL	nounsg	<unknown>
wordt	verbpressg	worden
opwaarts	adj	opwaarts
bijgesteld	verbpapa	bijstellen
van	prep	van
200.00	num__card	@card@
USD	adj	<unknown>
naar	prep	naar
220.00	num__card	@card@
USD	adj	<unknown>

verb, common noun, adjective, or adverb is chosen as the name of the broker. Extensive tests of this method provided no errors in extracting the name of the issuer.

6.4.2 Temporal Reasoner

In our framework, the temporal reasoner represents the interface for populating the tOWL knowledge base. The creation of the required instances is based on a set of input sources, namely: the real-time data extracted through the Information Extraction Module, historical data, and the tOWL domain ontology that describes, at an abstract level, the domain

for which the real-time data is extracted. The reasoner also ensures consistency of the knowledge base, both at a static and temporal level. An additional input to the temporal reasoner consists of the tOWL knowledge base itself, that is used for checking the consistency of the real-time data obtained through the Information Extraction Module with the current version of the knowledge base.

For the application used to illustrate the functionality of the proposed framework, i.e., the aggregation of market recommendations, the temporal reasoner is mostly used for determining the temporal validity of recommendations. The recommendations are assigned a default validity duration of six months, or until a new recommendation is issued. In the future, this can be made a parameter and be optimized for its best value. Thus, with the addition of of new recommendations to the tOWL knowledge base, the relevant instances are checked for determining the new interval, if applicable, for which they are valid.

6.4.3 The Temporal Ontology and Knowledge Base

The domain for which the temporal decision support system is designed is described in a temporal ontology. For the framework proposed in this chapter, we rely on the state-of-the-art tOWL language for the representation of the ontology. In the market recommendations aggregation application, we use two ontologies, namely a Financial Domain Ontology (FDO), as well as a Market Recommendations Ontology (MRO).

The FDO describes companies and is mostly focussed around their properties, such as the name, quoteSymbol, the stock exchange where the company is listed, the industry sector in which the company is active, etc. Additionally, this ontology describes analysts in terms of their unique code, name, and affiliation (where available). The recommendations in the MRO are matched to companies from the FDO. In the MRO we describe the basic properties of market recommendations, such as the analyst that issued the recommendation, the company for which the recommendation is issued, and the default duration

of recommendations. Here, we assume a default validity of a recommendation to be six months, unless a new recommendation is issued by the same analyst, for the same company, within this time interval. This is relevant for the aggregation of recommendations since we need to be able to determine, at any point in time, the recommendations that hold for a company in order to generate the aggregated recommendation.

The knowledge bases for FDO and MRO contain the concrete knowledge that is available about companies, analysts, and already issued recommendations. Especially in the case of the MRO knowledge base, the temporal dimension of the tOWL language provides added value in storing and reasoning with temporal knowledge. This is mostly due to the temporal nature of recommendations, i.e., they only hold for a limited, predefined period of time (6 months), or until a new recommendation is issued. The temporal reasoner is able in such cases to adjust the ending time of a recommendation based on information that becomes available.

6.4.4 Aggregation Module

The aggregation module is aimed at using existing information for the generation of a forecast/optimal decision. The term aggregation is used here in the broad sense of a model taking various inputs and generating an output that can be used in decision-making. The inputs are stored in the tOWL knowledge base(s).

For the market recommendations aggregation system, the aggregation module processes the recommendations that hold true at any point in time, and, based on the past performance of the analysts that have issued those recommendations, generates an aggregated advice in the form of buy, hold, or sell, for the asset being considered. The aggregation module for the market recommendations aggregation system is outlined in Section 6.5.

6.4.5 System Output

The general architecture that we propose outlines a time-enabled decision support system. Thus, the system we propose is aimed at decision support, where a decision is regarded as a selection between different available courses of action.

For the market recommendations aggregation system, the output consists of an aggregated recommendation, in the form of either a buy, hold, or sell recommendation. This aggregated recommendation is given at a certain point in time, and holds until the system generates an aggregated recommendation that is different from the current recommendation.

6.4.6 Instantiation of the Framework

Based on the general framework presented in this chapter, we can formulate an instantiation of this framework. This instance of our framework consists of a market recommendations aggregation system. For this system, the different components can be instantiated as presented in Figure 6.5. Here, the real-time data consists of market recommendations, i.e., advices of, roughly speaking, three different types: buy, hold, and sell. Historical data is data related to asset prices, as well as past recommendations. The tOWL ontology describes properties of recommendations as well as firms issuing the recommendations. Based on these sources, the temporal reasoner is used to update the tOWL knowledge base with new recommendations and performance data on firms. From the tOWL knowledge base, the advice aggregation module gathers and aggregates relevant data for the generation of an aggregated advice, which is also the output of the system.

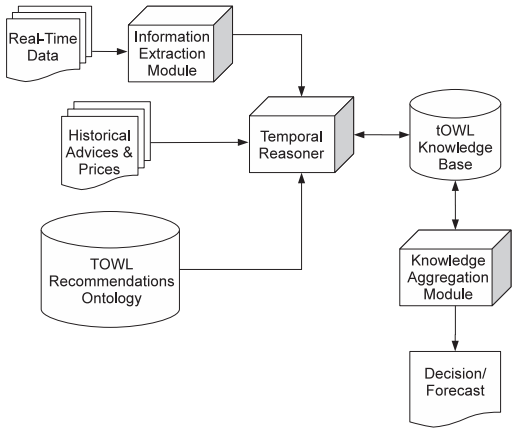

Figure 6.5: The market recommendations aggregation system

6.5 The Market Recommendations Aggregation Module

In this section we outline how individual recommendations issued by analysts can be aggregated into a single recommendation. This aggregated recommendation takes into account the past performance of the analysts being considered. In computing this performance, we correct the achieved average return of analysts by the standard deviation of those returns. The resulting measure, known as the Sharpe ratio (Sharpe, 1966), gives a quantification of the performance in terms of achieved return corrected for the risk taken. The Sharpe ratio is computed for each advice type, i.e., buy, hold, and sell.

We illustrate the computation of the aggregated recommendation by means of an example. At time t, the distribution of advices presented in Figure 6.6 is known for company C, given the analysts denoted as A_n. Given this distribution, the goal is to compute the aggregated recommendation by taking into account the past performance of the analysts. In the past, each of the analysts has issued the advices presented in Figure 6.7, denoted as a_n.

A_1 : `Buy`
A_2 : `Buy`
A_3 : `Buy`
A_4 : `Hold`
A_5 : `Hold`
A_6 : `Sell`

Figure 6.6: Advice distribution – example

A_1 : a_1, a_2
A_2 : a_3
A_3 : a_4
A_4 : a_5, a_6
A_5 : a_7
A_6 : a_8

Figure 6.7: Historical recommendations – example

In order to determine the expected performance of each advice type, the most obvious choice would be to look at the past returns of the analysts and aggregate these for each advice type. This expected performance for an advice type, say buy, can be obtained as follows in our example. Here, r_{a_n} represents the return generated by the advice a_n one day after the recommendation was issued.

$$E(R_B) = (r_{a_1} + r_{a_2} + r_{a_3} + r_{a_4})/4 \tag{6.1}$$

In similar manner we obtain the expected returns $E(R_H)$ and $E(R_S)$ for the other two advice types. However, judging the performance of an advice solely in terms of generated returns paints an incomplete image, since the risk taken for obtaining these returns is not considered. We choose to measure the risk in terms of the standard deviation of the returns

as follows. In the following example, we compute the risk for buy recommendations for our imaginary stock:

$$\begin{aligned}\sigma(R_B) &= 1/3((r_{a_1} - E(R_B))^2 + (r_{a_2} - E(R_B))^2) + \\ &\quad +(r_{a_3} - E(R_B))^2 + (r_{a_4} - E(R_B))^2)^{1/2}\end{aligned} \tag{6.2}$$

The measure of expected performance that takes into account risk, as given by the Sharpe ratio, is calculated as follows for the buy advices in our example:

$$S_B = E(R_B)/\sigma(R_B) \tag{6.3}$$

In similar fashion we obtain the Sharpe ratios for the other two advice types, which we denote as S_H and S_S.

Though tempting, only considering the maximum value of the Sharpe ratio in determining the aggregated advice would not take into account the number of analysts that issued recommendations for each advice type. A weighted measure of performance is thus desired that accounts for the number of analysts who issued an advice type, as well as the Sharpe ratio. We obtain this by multiplying each Sharpe ratio, for each advice type, with the number of analysts that issued advices of that advice type. This is computed as follows for the buy advices in our example, where n_B^t denotes the number of recommendations that hold at time t for the buy advice type:

$$P_B = n_B^t S_B \tag{6.4}$$

Having computed this performance measure for the other two advice types, denoted as P_H and P_S, respectively, the aggregated recommendation is computed as the maximum of the three individual P values.

Generalizing this approach, we begin by computing, for each advice type, the expected return $E(R_x)$ as follows:

$$E(R_x) = \frac{\sum_{i=1}^{m} r_{a_i}^h}{m} \tag{6.5}$$

Here x can be either buy, hold, or sell, m denotes the total number of advices issued in the past by the analysts who issued recommendation x, h denotes the time horizon being considered for the computation of the returns of the individual advices, and a_i are the individual advices being considered.

The standard deviation of the returns per advice type, $\sigma(R_x)$, is computed as follows, where the symbols preserve their previous meaning:

$$\sigma(R_x) = \texttt{left}(\frac{1}{m-1}\sum_{i=1}^{m}(r_{a_i}^h - E(R_x))^2\texttt{right})^{1/2} \tag{6.6}$$

The Sharpe ratio is computed as follows for each advice type x:

$$S_x = \frac{E(R_x)}{\sigma(R_x)} \tag{6.7}$$

The expected performance, P_x, of each advice type as a function of its associated Sharpe ratio, S_x, and the number of advices holding at time t for advice type x, denoted as n_x^t, is computed as follows:

$$P_x = n_x^t S_x \tag{6.8}$$

Finally, the aggregated recommendation is determined by choosing the advice type with the maximum expected performance, P_x.

6.6 Results and Discussion

In this section we present the results obtained by testing different market recommendations aggregation methods by using the time-enabled decision support system described in this chapter. In Section 6.6.1 we describe the experimental setup that stands at the basis of the experiments that we perform. The results we obtain are presented in Section 6.6.2. We conclude with a discussion of the results in Section 6.6.3.

6.6.1 Experimental Setup

For the experiments, we use data collected for the period January 1st, 2000 to December 31st, 2010. The data consists of all the recommendations issued for a company listed in the Dow Jones Industrial Average (DJIA) index, a set consisting of 30 companies, such as American Express, Boeing, and JPMorgan Chase.

We compute the performance of analysts that have issued a recommendation based on their past performance in the three years prior to the recommendation being considered. The performance is thus computed over a moving window comprising three years previous to the point in time when an aggregated recommendation is being computed. This computation takes into account all recommendations issued by the analyst for any company in any of the US markets, thus not being restricted to past performance in relation to the DJIA companies.

The aggregated recommendations are computed for every day in the dataset, but we only measure the performance of recommendations that, at time t, are different from the recommendation issued at $t - 1$. In other words, although our system computes daily aggregated recommendations, we consider an advice to be issued by the system only if that advice is different from the advice issued on the previous occasion. It is only for these recommendations that we compute the performance in terms of returns for different time horizons.

The performance of recommendations is computed both in terms of returns as well as Sharpe ratios, for different time horizons. The horizons that we consider are, relative to the time when an aggregated recommendation is issued: one day, one week, one month, half-year, and one year, respectively. The strategies that we consider are *Index*, which consists of investing in the DJ30 index with no additional strategies, *Majority voting*, where we consider the advices issued by the highest number of analysts as the aggregated recommendation, *Analyst performance*, where the aggregated advice is computed by taking into account the past performance of the analysts (measured by the Sharpe ratio), and *Equally-weighted index*, which is a variant of the DJ30 index where all companies are given equal weights.

6.6.2 Results

In this section we provide an overview of the performance of the aggregated recommendations.

Table 6.2 presents the mean returns for the four investment strategies considered, for each of the five time horizons that we use. When the two aggregation methods are compared, the majority voting method is outperformed by the aggregation method that takes into consideration the past performance of analysts, for short time horizons (1-day, 1-week, and 1-month). However, this relationship reverses for longer time horizons (half-year and 1-year). When compared with the performance of the index, the latter outperforms both methods for very short time horizons (1-day, 1-week), but is outperformed by either one or the other aggregation method for the remaining time horizons, with one exception: for 1-month majority voting performs worse than the index. Additionally, for very short time horizons the equally weighted index outperforms all aggregation methods.

In Table 6.3 we present the standard deviations of the returns computed for the four different investment strategies, at different time horizons. We note that, when comparing the two aggregation methods, the standard deviations for short time horizons (1-day,

	1-day	1-week	1-month	Half-year	1-year
Index	0	-0.0002	0.0012	0.0054	0.0143
Majority voting	-0.0004	-0.0008	-0.0002	0.0141	0.0247
Analyst perf.	-0.0002	-0.0003	0.0018	0.0092	0.0171
Equally-weighted index	0.0001	0.0004	0.0014	0.0045	0.0098

Table 6.2: Mean returns

	1-day	1-week	1-month	Half-year	1-year
Index	0.0128	0.0267	0.0520	0.1151	0.1796
Majority voting	0.0270	0.0533	0.0967	0.1815	0.2644
Analyst perf.	0.0269	0.0529	0.0979	0.2201	0.3107
Equally-weighted index	0.0224	0.0490	0.0935	0.1953	0.2701

Table 6.3: Standard deviation of returns

1-week) are highly similar, with slighter lower values for the aggregation method based on the Sharpe ratio. At longer time horizons, the majority voting method attains lower standard deviations. However, both methods incur more risk when compared to the index, the latter displaying lower standard deviations for all time horizons.

Last, we present the performance of the different investment strategies in terms of the Sharpe ratio in Table 6.4. As in the case of returns, when comparing the two aggregation methods, the majority voting method is outperformed on the shorter term (1-day, 1-week, 1-month), but superior for longer time horizons (half-year and 1-year). For short time horizons, the index outperforms both aggregation methods in terms of Sharpe ratio, which is also the case for the time horizon of one year. For the half-year interval, the majority voting method is able to outperform both the index, as well as the equally-weighted index.

	1-day	1-week	1-month	Half-year	1-year
Index	0.0062	0.0112	0.0230	0.0473	0.0793
Majority voting	-0.0157	-0.0165	-0.0024	0.0778	0.0667
Analyst perf.	-0.0082	-0.0054	0.0188	0.0420	0.0551
Equally-weighted index	0.0071	0.0095	0.0149	0.0231	0.0362

Table 6.4: Sharpe ratios

6.6.3 Discussion

The previous section outlines the results obtained for the two aggregation methods that we consider, as well as the performance of the DJ30 index and a fictive, equally-weighted index. In terms of raw performance, measured as mean returns, the two aggregation methods are dependent in their performance on the time horizon being considered. An aggregation method that takes into account the past performance of analysts when performing the aggregation, delivers superior results for relatively short time horizons. However, for longer time horizons, a majority voting approach is superior. Despite this result, both aggregation methods are outperformed by the DJ30 index for very short time horizons (one day and one week), as well as by the equally-weighted index. However, both aggregation methods outperform the DJ30 index as well as the equally-weighted index for time horizons of half-year and one year, leading to the conclusion that the information available to analysts when issuing their recommendations has an abnormal effect only in the long term.

In terms of risk, measured as standard deviation, the index is the least risky investment strategy considered, followed by the equally-weighted index, a predictable result. Despite this, we note that the majority voting aggregation method outperforms the equally-weighted index, in terms of incurred risk, for longer time horizons (half-year and one year). The majority voting aggregation method outperforms both the index, as well as the equally weighted index, in terms of Sharpe ratio for the half-year time horizon. For shorter time horizons, the previous results based on standard deviation where the index was the less risky than the aggregation methods are also supported by the Sharpe ratio computations.

6.7 Conclusions and Future Work

This chapter describes a temporal framework for decision support systems. The framework we propose relies on the state-of-the-art tOWL language for the representation of temporal information. We illustrate how temporal information can be used in decision support systems in a systematic, consistent way. Additionally, by relying on such an approach, the process of data storage and retrieval, as well as consistency checks on the data are separated from the main application. The added value of the chapter consists of a temporal framework for decision support systems that relies on state-of-the-art Semantic Web technology for handling temporal knowledge.

We illustrate the applicability of our system in a practical context, by extracting, storing, and aggregating information related to market recommendations. By implementing different aggregation methods we demonstrate the flexibility of the proposed system. Additionally, we show how the systematic storing of knowledge by relying on the tOWL language enables different models to be deployed within the same application, models that (may) use different pieces of data from the information available within the knowledge base(s).

Although the system presented in this chapter is used for an application in the financial domain, the applications of the system are not restricted to this domain. For making use of the proposed framework, one needs to define a tOWL domain ontology, the input information, the historical data, and an application engine able to process all this temporal information into a valuable output. We envision applications in the automated processing of news by news agencies, accompanied by tagging and classification of these news items based on subject, geographic area, entities involved, i.e., persons and locations.

As future work we plan to extend the proposed framework to exploit tOWL expressions directly in the information extraction phase, based on the tOWL ontology and knowledge base(s) used in the application. This can be achieved in different ways, one of which is

the definition of patterns based on the the knowledge base that can directly be used by the information extraction module.

Chapter 7

Conclusions

"What we call the beginning is often the end. And to make an end is to make a beginning. The end is where we start from." – T.S. Elliot

This dissertation answers four research questions, all focused on the subject of News Analytics in Finance. Along the incremental structure on which the five main chapters are built, we answer these questions. Chapter 2 addresses the issue of how information from structured text can be used for the prediction of economic trends. This is an early inquiry into the use of news in financial trading, hence the choice for structured text having the property of being easier to process and compare across time. In Chapter 3 we broaden our definition of news to include different types of events, and answer the question of how information from news can be included in trading strategies. The framework we introduce is validated empirically, and shows that indeed the events that we consider generally lead to improved trading strategies.

However, besides underlining the measurable added value of using news in trading strategies, these first two chapters also identify a need for the representation of (temporal) knowledge in a more systematic way, while allowing for consistency checks and inference on the representation. We build a Temporal Web Ontology Language and extend it, in Chapters 4 and 5. Although built with the financial domain in mind, the expressiveness and computational properties of the tOWL language make it suitable for a variety of

other domains. Chapter 6 brings together the insights of chapters 2 and 3, as well as the technology of chapters 4 and 5. This final chapter presents a general framework for time-aware decision support systems. We validate this framework empirically by showing how market recommendations can be automatically collected, represented, and aggregated based on different methodologies.

7.1 Concluding Remarks

Our inquiry into the use of news from text in financial trading started with the computational analysis of European Central Bank statements. A desirable property of these statements is that they follow a similar structure over time, which makes them relatively comparable to each other and, to some extent, eases the analysis. This chapter introduces a framework for the computational analysis of such statements. The main components that identify this framework are: linguistic/semantic preprocessing, content fingerprinting, and the content modelling component. Next to the statements, we rely on historical price data. We validate this framework through two instantiations: an adjective frequency approach, and an approach based on fuzzy grammar fragments. The accuracy is measured in terms of how often the proposed models correctly predict the upward and downward movements of the MSCI EURO index. Though considering different levels of complexity in the analysis of text, the approaches are comparable in terms of results. The accuracy of both methods outperforms a random classifier. We thus conclude that quantifying the content of ECB statements based on a content dictionary and different levels of complexity in the depth of the analysis can be done in a meaningful way. This validates the framework that we propose, and shows that information from structured text can be used for predicting economic trends, following the methodology outlined in Chapter 2.

By relaxing the definition of news as used in algorithmic trading, we proceeded to combine technical trading indicators with events extracted from news messages. The framework that results from this analysis considers genetic programming at the core of

the trading system. This approach is highly flexible in that it can adapt to changing market conditions by learning new rules and adjusting existing ones. A comparison of the performance of some of the widely known technical trading indicators with event information extracted from news shows the latter's superiority in explaining price changes. When the variable pool includes both technical trading indicators and event information from news, we find that most trading rules discovered by the genetic program include the news variable. This finding supports the idea that by extracting event information, news can be included in trading strategies following the methodology outlined in Chapter 3.

During the inquiries of the first two chapters, another question surfaced regarding the representation of knowledge. What became apparent was that dealing with such large volumes of knowledge requires a systematic way of representing this knowledge. The knowledge that we seek to represent relates to both the structure of the domain being studied – in our case the financial domain – and more concrete, ephemeral knowledge emerging from the information sources used – in our case news messages. Additionally, the representation must enable automated inference on the knowledge base, and provide the ability to discover inconsistencies. The available technology failed at satisfying all these requirements to an acceptable extent, one of the main shortcomings being the lack of support for dealing with time and change. Starting from existing languages, we designed a temporal web ontology language that satisfies the stated requirements. The language design is based on a layered approach, where we incrementally introduce, as extensions to the existing Web Ontology Language, the following features: *concrete domains*, which enable the representation of restrictions using binary concrete domain predicates; *temporal representation*, through which we enable the representation of timepoints and intervals, as well as the relationships that may hold between them; *timeslices and fluents*, directly aimed at the representation of ephemeral knowledge. A discussion of more complex constructs is in tOWL is also provided. We support our theoretical contributions through different examples, such as the representation of a Leveraged Buy Out process in

tOWL. The language, the optimizations, the new constructs, and application examples are presented in Chapters 4 and 5.

With more expressive representation tools at our disposal, and the acquired knowledge on news-based algorithmic trading, we develop a more general framework that is able to deal with the dynamic nature of financial markets. Central to our framework is the tOWL language, used for representing the knowledge extracted from the information sources that are considered for the system. We show that both abstract knowledge, represented in a tOWL ontology, and concrete knowledge, represented the tOWL knowledge base, are required within such a system. This approach enables part of the application intelligence to be incorporated in the representation itself. The flexibility of systems based on the framework that we introduce is shown by implementing a market recommendations aggregation system. The design enables the representation of different aggregation methods, by relying on knowledge that can be reused across applications as well as systems. Although not further investigated in this dissertation, the framework based on the tOWL language that we introduce also seems suitable for decision support outside the financial domain. The framework for time-aware decision support systems and the proof-of-concept application are presented in Chapter 6.

7.2 The Future of News Analytics for Financial Decision Support

The process of answering the research questions stated in the Introduction, as well as the answers themselves, raise new issues, from several perspectives. The interdisciplinary approach taken in this dissertation leads us to believe that the future of news-based algorithmic trading relies on developments in several fields.

Extracting knowledge from free text is a daunting task. Dealing with complex constructs, ambiguity, and even simpler grammatical constructs, such as negation, constitute

some of the subjects of ongoing work. In addition, the automated extraction of opinions, sentiment, and content, although going through a period of rapid development, is based on techniques still far from the capabilities of the human brain (we consider here the accuracy of these methods rather than their ability to deal with very large volumes of unprocessed information).

The representation of knowledge while taking into consideration re-usability and automatic inference is being extensively studied within the Semantic Web community. This field has experienced developments at a rapid pace, with different applications proving its usability in several fields (life sciences, social sciences, engineering, etc.). However, the quest for the application that will elevate the Semantic Web to an indisputable, critical role in representing and processing information is far from over. Before such an application can emerge, further development is required of the underlying formalisms of Semantic Web languages and the available tools, as well as semantic data sources. Further work on the consolidation and standardization of ontologies should also be the focus of research in this field.

The systemic implications of automated approaches to financial trading require careful consideration. The "Flash Crash" example discussed in the Introduction illustrates a possible cost of *limited* context awareness, until now an inherent trait of machine-based trading. And although the upside of HFT in financial markets has been academically documented, deeper analysis is required in order to understand and weight the benefits and the costs of these new markets. A predictable conclusion, already outed by different market participants and overseeing bodies, is a requirement for significant updates of current regulation in financial markets.

The fact that a relationship exists between news and financial markets is unquestionable, as also shown in this dissertation. The content of news messages, the events contained therein, or recommendations issued by analysts all contribute to explaining economic trends and the value of individual financial instruments. We have shown that

the use of such information can be automated, and with this, a final thought requires consideration – the issue of trust, i.e., is it a good idea to deal with vast amounts of money based on algorithms, and assign such a role at the core of our economies to machines? Looking at other fields and industries, technology is a major contributor to the advancement of the manufacturing industry, aviation, and even medicine, to name a few examples. The role of technology in aviation has been increasing at accelerating pace throughout time; concurrently, airline safety has been more or less steadily increasing over the past two decades, with 2011 being the safest year in aviation (Ascend, 2012). The same development can be observed in medicine, through technology such as clinical decision support systems and computer-assisted surgery.

What is true of other fields and professions, also holds for the field studied in this dissertation. A certain amount of common sense in required, both in the design of automated systems that trade based on news, as well as in their general use in the markets. Continuous, extensive feedback and reflection in this developmental period are critical: "It is the obvious which is so difficult to see most of the time. People say 'It's as plain as the nose on your face.' But how much of the nose on your face can you see, unless someone holds a mirror up to you?" (Asimov, 1950)

Summary

One third of all share trades in the United States were initiated through algorithmic trading in 2006. Recent estimates of this share for US equity trading reach as high as 77%, while for European markets this figure ranges between 30% and 50%. Machines are the new breed of trader. The fast growth of the use of automated trading and the overall advancement of such technologies contribute to decreased profitability of such algorithms. This trend is set to continue in the same direction over the coming ten years. The competitive advantage in this market is most likely to come from developments in News Analytics - automated techniques for collecting, processing, and using the information contained in news messages. The value of using information from news in Automated Trading comes from enabling (re)action to new information before it is incorporated into market prices.

This PhD thesis contributes to the newly emerged, growing body of scientific work on the use of News Analytics in Finance. Regarded as the next significant development in Automated Trading, News Analytics extends trading algorithms to incorporate information extracted from textual messages, by translating it into actionable, valuable knowledge. The added-value of this dissertation can be divided roughly in two contributions. On one side, we contribute to the design and implementation of systems that use news for trading. On the other side we deliver more theoretical contributions related to the technology that is required for such systems.

The dissertation presents three systems that are able to use news for financial trading. The first system we present deals with structured text, in this case the monthly statements of the European Central Bank. Another system that we present introduces a framework for using information from news messages, in the form of events, for share trading. Finally, we introduce a more complex system, able to deal with time and time-related information as extracted from news. This system stands at the foundation of the time-aware framework for financial decision support that we introduce.

A considerable part of this dissertation deals with developing technology that we deem necessary for News Analytics in Finance. Here, we present a temporal web ontology

language. We envision this language as a building block of all systems that deal with temporal knowledge. The language is able to represent abstract domain knowledge as well as more concrete (temporal) facts, such as the information contained in news messages.

Finally, despite its high relevance for business and the "next big thing in automated trading" status, News Analytics, especially in a financial context, is still in its academic infancy. Next to the concrete contributions of this dissertation, our work also delivers a proposal. Our discussion identifies the disciplines that can contribute to this emerging field. Finally, we provide possible focus points for future research.

Samenvatting

Een derde van alle aandelenhandel in de VS werd gestart met behulp van algoritmen in 2006. Recente schattingen van deze omvang voor de aandelenhandel in de VS bereiken een waarde van 77%, terwijl bij Europese markten dit getal reikt tussen 30% en 50%. Machines zijn de moderne beurshandelaren geworden. De snelle groei van de geautomatiseerde financiële handel en de geavanceerde technologie dragen bij aan een dalende winstgevendheid van dergelijke algoritmen. Deze trend zal zich in de komende tien jaar voortzetten. De concurrentiepositie in deze markten zal, naar alle waarschijnlijkheid, voortkomen uit ontwikkelingen op het gebied van kwantitatieve nieuwsanalyse. Dit is een geautomatiseerde techniek voor het verzamelen, verwerken en gebruiken van informatie verkregen vanuit nieuwsberichten. Het voordeel van het gebruiken van informatie verkregen uit nieuwsberichten bij geautomatiseerde handel is het in staat stellen om sneller te reageren op nieuwe informatie, nog voordat het in de marktprijzen verwerkt is.

Dit proefschrift levert een bijdrage aan wetenschappelijk onderzoek met betrekking tot het gebruik van kwantitatieve nieuws analyse binnen financiële besluitvorming. Kwantitatieve nieuws analyse breidt handelsalgoritmen uit door nieuwe informatie, verkregen vanuit tekstberichten, om te zetten in doelgerichte en winstegevende kennis. De toegevoegde waarde van deze dissertatie kan ruwweg onderscheiden worden in twee onderdelen. Enerzijds dragen wij bij aan het design en de implementatie van systemen die nieuws gebruiken voor handel. Anderzijds leveren wij theoretische bijdragen gerelateerd aan de technologie die vereist is voor dergelijke systemen.

We presenteren drie systemen die in staat zijn om nieuwsberichten te gebruiken voor de financiële handel. Het eerste systeem maakt gebruik van gestructureerde tekst, namelijk de maandelijkse persberichten van de Europese Centrale Bank. Een ander systeem laat zien hoe informatie uit nieuwsberichten, in de vorm van events, gebruikt kan worden bij aandelenhandel. Het laatste systeem laat een complexere structuur zien, die de mogelijkheid biedt om tijd- en tijdsgerelateerde informatie die verkregen wordt vanuit nieuwsberichten representeren en te gebruiken. Hiermee wordt een algemene structuur

geïntroduceerd voor financiële besluitvormingssystemen waarbij de tijdsdimensie als relevant wordt beschouwd.

Een aanzienlijk deel van de dissertatie gaat over de ontwikkeling van de technologie die wij als noodzakelijk beschouwen voor kwantitatieve nieuwsanalyse met betrekking tot financiële besluitvorming. Hierin presenteren we een temporal web ontology language. Deze formele taal wordt gezien als bouwsteen van alle systemen die te maken hebben met kennis waarbij de tijdsdimensie een belangrijke rol speelt. Deze taal stelt ons in staat om zowel abstracte domeinkennis als meer concrete (tijdsgerelateerde) feiten, zoals de informatie die nieuwsberichten bevatten, te gebruiken.

Kwantitatieve nieuwsanalyse staat in de financiële wereld, ondanks zijn grote relevantie voor het bedrijfsleven en de status als "next big thing in automated trading", nog in de kinderschoenen. We leveren concrete bijdragen aan de ontwikkeling van kwantitatieve nieuwsanalyse. Daarnaast brengt onze discussie in kaart welke disciplines een bijdrage kunnen leveren aan dit vakgebied. Tot slot worden er suggesties voor toekomstig onderzoek aangedragen.

Bibliography

Acampora, G., Loia, V., 2005a. Fuzzy control interoperability and scalability for adaptive domotic framework. IEEE Transactions on Industrial Informatics 1 (2), 97–111.

Acampora, G., Loia, V., 2005b. Using FML and fuzzy technology in adaptive ambient intelligence environments. International Journal of Computational Intelligence Research 1 (2), 171–182.

Achelis, S., 2000. Technical analysis from A to Z. McGraw-Hill.

Allen, F., Karjalainen, R., 1999. Using genetic algorithms to find technical trading rules. Journal of Financial Economics 51 (2), 245 – 271.

Allen, J., 1983. Maintaining knowledge about temporal intervals. Communications of the ACM 26 (11), 832–843.

Andreevskaia, A., Bergler, S., 2006. Mining wordnet for fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In: The 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2006). Vol. 6. pp. 209–216.

Angel, J., Harris, L., Spatt, C., 2010. Equity trading in the 21st century (FBE 09-10). URL `http://dx.doi.org/10.2139/ssrn.1584026`

Aristotle, 1991. Aristotle on rhetoric: A theory of civic discourse. Oxford University Press.

Arroyo-Figueroa, G., Sucar, L., Villavicencio, A., 1998. Probabilistic temporal reasoning and its application to fossil power plant operation. Expert Systems with Applications 15 (3), 317–324.

Artale, A., Franconi, E., 1994. A computational account for a description logic of time and action. In: The 4th Conference on Principles of Knowledge Representation and Reasoning (KR 1994). Morgan Kaufmann, pp. 3–14.

Artale, A., Franconi, E., 1998. A temporal description logic for reasoning about actions and plans. Journal of Artificial Intelligence Research 9 (2), 463–506.

Artale, A., Franconi, E., 2000. A survey of temporal extensions of description logics. Annals of Mathematics and Artificial Intelligence 30 (1), 171–210.

Artale, A., Franconi, E., Mosurovic, M., Wolter, F., Zakharyaschev, M., 2001. The $\mathcal{DLR}_{\mathcal{US}}$ temporal description logic. In: The 2001 Description Logic Workshop (DL 2001). CEUR Workshop Proceedings, pp. 96–105.

Artale, A., Kontchakov, R., Ryzhikov, V., Zakharyaschev, M., 2010. Temporal conceptual modelling with DL-Lite. In: 23rd International Workshop on Description Logics (DL 2010).

Ascend, 2012. 2011: safest year on record for air travel and for insurers, lowest claims for seven years. Tech. rep., Ascend.

Asimov, I., 1950. I, robot. Voyager.

Baader, F., Calvanese, D., McGuinness, D., Patel-Schneider, P., Nardi, D., 2003. The description logic handbook: theory, implementation, and applications. Cambridge Univ Press.

Baader, F., Hanschke, P., 1991. A scheme for integrating concrete domains into concept languages. In: The 12th International Joint Conference on Artificial Intelligence, (IJCAI 1991). Morgan Kaufmann, pp. 452–457.

Babuška, R., 1998. Fuzzy modeling for control. Vol. 12. Kluwer Academic Pub.

Banko, M., Moore, R. C., 2004. Part of speech tagging in context. In: The 20th International Conference on Computational Linguistics (COLING 2004). ACM, pp. 556–561.

Barber, B., Lehavy, R., McNichols, M., Trueman, B., 2001. Can investors profit from the prophets? Security analyst recommendations and stock returns. The Journal of Finance 56 (2), 531–563.

Batini, C., Ceri, S., Navathe, S., 1991. Conceptual database design: an entity-relationship approach. Addison Wesley.

Batsakis, S., Petrakis, E., 2010. SOWL: spatio-temporal representation, reasoning and querying over the semantic web. In: 6th International Conference on Semantic Systems. ACM, pp. 1–9.

Batsakis, S., Petrakis, E., 2011. Representing temporal knowledge in the semantic web: The extended 4d fluents approach. Combinations of Intelligent Methods and Applications, 55–69.

Bechhofer, S., Van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D., Patel-Schneider, P., Stein, L., 2004. OWL web ontology language reference. W3C Recommendation.

Berners-Lee, T., Hendler, J., Lassila, O., 2001. The semantic web. Scientific American 284 (5), 28–37.

Bezdek, J., 1981. Pattern recognition with fuzzy objective function algorithms. Kluwer Academic Publishers.

Bjerring, J., Lakonishok, J., Vermaelen, T., 1983. Stock prices and financial analysts recommendations. The Journal of Finance 38 (1), 187–204.

Black, F., 1971a. Towards a fully automated exchange, part i. Financial Analysts Journal 27 (4), 29–34.

Black, F., 1971b. Towards a fully automated exchange, part ii. Financial Analysts Journal 27 (4), 28–35.

Bohlen, M. H., Snodgrass, R. T., Soo, M. D., 1996. Coalescing in temporal databases. IEEE Computer 19, 35–42.

Bollen, J., Mao, H., Zeng, X., 2011. Twitter mood predicts the stock market. Journal of Computational Science 2, 1–8.

Bommel, J., 2003. Rumors. The Journal of Finance 58 (4), 1499–1520.

Bonnier, K., Bruner, R., 1989. An analysis of stock price reaction to management change in distressed firms. Journal of Accounting and Economics 11 (1), 95–106.

Brickley, D., Guha, R., 2004. RDF vocabulary description language 1.0: RDF schema. W3C Recommendation.

Brogaard, J., 2012. High frequency trading and volatility. SSRN eLibrary.

Calegari, S., Farina, F., 2007. Fuzzy ontologies and scale-free networks analysis. International Journal of Computer Science & Applications 4 (2), 125–144.

Cecchini, M., Aytug, H., Koehler, G., Pathak, P., 2010. Making words work: Using financial text as a predictor of financial events. Decision Support Systems 50 (1), 64–175.

Chan, W., 2003. Stock price reaction to news and no-news: drift and reversal after headlines. Journal of Financial Economics 70 (2), 223–260.

Chen, P., 1976. The entity-relationship model–toward a unified view of data. ACM Transactions on database systems 1 (1), 9–36.

Chinn, S., Madey, G., 1997. A framework for developing and evaluating expert systems for temporal business applications. Expert Systems with Applications 12 (3), 393–404.

Das, S., Chen, M., 2007. Yahoo! for Amazon: Sentiment extraction from small talk on the web. Management Science 53 (9), 1375–1388.

Dence, S., Latimore, D., White, J., 2006. The trader is dead, long live the trader! a financial markets renaissance. Tech. Rep. G510-6271, IBM Institute for Value.
URL `http://www-935.ibm.com/services/us/imc/pdf/g510-6271-trader.pdf`

Dyreson, C., 2003. Temporal coalescing with now granularity, and incomplete information. In: ACM SIGMOD International Conference on Management of Data (SIGMOD 2003). ACM, pp. 169–180.

ECB Press Conferences, 2011. European central bank. Last Accessed: August 2nd, 2011. URL http://www.ecb.int/press/pressconf

Esuli, A., Sebastiani, F., 2006. Determining term subjectivity and term orientation for opinion mining. In: The 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2006). pp. 193–200.

Ewalds, S., Schauten, M., Steenbeek, O., 2000. De informatiewaarde van kwartaalcijfers. Maandblad voor Accountancy en Bedrijfseconomie (7/8), 333–341.

Fan, W., Pathak, P., Wallace, L., 2006. Nonlinear ranking function representations in genetic programming-based ranking discovery for personalized search. Decision Support Systems 42 (3), 1338–1349.

Fenza, G., Loia, V., Senatore, S., 2008. A hybrid approach to semantic web services matchmaking. International journal of approximate reasoning 48 (3), 808–828.

Foresight, 2010a. The future of computer trading in financial markets. Tech. Rep. 11-1276, Foresight.
URL http://www.bis.gov.uk/assets/bispartners/foresight/docs/computer-trading/11-1276-the-future-of-computer-trading-in-financial-markets.pdf

Foresight, 2010b. Technology trends in the markets: A 2020 vision. Tech. Rep. 11-1222, Foresight.
URL http://www.bis.gov.uk/assets/foresight/docs/computer-trading/11-1222-dr3-technology-trends-in-financial-markets.pdf

Frasincar, F., Milea, V., Kaymak, U., 2010. tOWL: Integrating time in OWL. In: Virgilio, R. D., Giunchiglia, F., Tanca, L. (Eds.), Semantic Web Information Management: A Model-Based Perspective. Springer, pp. 225–246.

General Inquirer, 2011. Harvard university. Last Accessed: August 2nd, 2011.
URL http://www.wjh.harvard.edu/~inquirer/

Gomber, P., Arndt, B., Lutat, M., Uhle, T., 2011. High-frequency trading. Tech. rep.
URL http://dx.doi.org/10.2139/ssrn.1858626

Grandi, F., 2009. Multi-temporal rdf ontology versioning. In: International Workshop on Ontology Dynamics (IWOD 2009).

Gregersen, H., Jensen, C., 1999. Temporal entity-relationship models-a survey. IEEE Transactions on Knowledge and Data Engineering 11 (3), 464–497.

Gutierrez, C., Hurtado, C., Vaisman, A., 2007. Introducing time into RDF. IEEE Transactions on Knowledge and Data Engineering 19 (2), 207–218.

Haase, P., Stojanovic, L., 2005. Consistent evolution of owl ontologies. In: 2nd European Semantic Web Conference (ESWC 2005). Springer, pp. 182–197.

Harold D. Lasswell, N. L., 1949. Language of politics: Studies in quantitative semantics. MIT Press, out of press.

Hayes, P., McBride, B., 2004. RDF semantics, W3C Recommendation.

Hendershott, T., Jones, C., Menkveld, A., 2011. Does algorithmic trading improve liquidity? The Journal of Finance 66 (1), 1–33.

Hendershott, T. J., Riordan, R., 2011. Algorithmic trading and information. SSRN eLibrary.

Hobbs, J., Pan, F., 2004. An ontology of time for the semantic web. ACM Transactions on Asian Language Information Processing 3 (1), 66–85.

Horrocks, I., 1998. The FaCT system. Automated Reasoning with Analytic Tableaux and Related Methods, 307–312.

Horrocks, I., Patel-Schneider, P., 1998. DL systems comparison. In: The 1998 Description Logics Workshop (DL 1998). Vol. 11 of CEUR-WS. pp. 55–57.

Ikenberry, D., Ramnath, S., 2002. Underreaction to self-selected news events: The case of stock splits. Review of Financial Studies 15 (2), 489–526.

Jensen, C., Clifford, J., Elmasri, R., Gadia, S., Hayes, P., Jajodia, S., 1994. A consensus glossary of temporal database concepts. SIGMOD Record 23 (1), 52–64.

Juarez, J., Campos, M., Palma, J., Marin, R., 2008. Computing context-dependent temporal diagnosis in complex domains. Expert Systems with Applications 35 (3), 991–1010.

Kaymak, U., Babuska, R., 1995. Compatible cluster merging for fuzzy modelling. In: The 1995 IEEE International Conference on Fuzzy Systems. Vol. 2. IEEE, pp. 897–904.

Keown, A., Pinkerton, J., 1981. Merger announcements and insider trading activity: An empirical investigation. Journal of Finance 36 (4), 855–869.

Kim, S., Lin, J., Slovin, M., 1997. Market structure, informed trading, and analystsï¿½ recommendations. Journal of Financial and Quantitative Analysis 32 (4), 507–524.

Kirilenko, A. A., Kyle, A. P. S., Samadi, M., Tuzun, T., 2011. The flash crash: The impact of high frequency trading on an electronic market. SSRN eLibrary.

Klingemann, H., Mohler, P., Weber, R., 1982. Das reichtumsthema in den thronreden des kaisers und die okonomische entwicklung in deutschland 1871-1914. Computerunterstutzte Inhaltsanalyse in der empirischen Sozialforschung.

Klyne, G., Carroll, J., 2004. Resource description framework (RDF): Concepts and abstract syntax. W3C Recommendation.

Kohane, I., 1987. Temporal reasoning in medical expert systems. Tech. rep.

Koza, J., 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. Cambridge, MA: MIT Press.

Lee, C., Jian, Z., Huang, L., 2005. A fuzzy ontology and its application to news summarization. IEEE Transactions on Systems, Man, and Cybernetics, Part B 35 (5), 859–880.

Lee, C., Wang, M., 2009. Ontology-based computational intelligent multi-agent and its application to CMMI assessment. Applied Intelligence 30 (3), 203–219.

Lee, C., Wang, M., Chen, J., 2008. Ontology-based intelligent decision support agent for CMMI project monitoring and control. International Journal of Approximate Reasoning 48 (1), 62–76.

Lee, T., Hendler, J., Lassila, O., et al., 2001. The semantic web. Scientific American 284 (5), 34–43.

Leibniz, G., 1969. Philosophical papers and letters. D. Reidel.

Leigh, W., Purvis, R., Ragusa, J., 2002. Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support. Decision Support Systems 32 (4), 361–377.

Liu, B., 2010. Sentiment analysis and subjectivity. Handbook of Natural Language Processing, 2nd ed., 627–666.

Lutz, C., 2000. Interval-based temporal reasoning with general tboxes. LTCS-Report LTCS-00-06, LuFG Theoretical Computer Science, RWTH Aachen.

Lutz, C., 2002. Adding numbers to the SHIQ description logic: First results. The Eighth International Conference on Principles of Knowledge Representation and Reasoning (KR 2002), 191–202.

Lutz, C., 2003. Description logics with concrete domains–a survey. In: Advances in Modal Logics. Vol. 4. King's College Publications, pp. 265–296.

Lutz, C., Milicic, M., 2007. A tableau algorithm for description logics with concrete domains and general tboxes. Journal of Automated Reasoning 38 (1–3), 227–259.

Lutz, C., Wolter, F., Zakharyashev, M., 2008. Temporal description logics: A survey. In: 15th International Symposium on Temporal Representation and Reasoning (TIME 2008). IEEE, pp. 3–14.

Maggiore, D., 2008. Deliverable 4.3: Design of a TOWL temporal reasoner. Towl project deliverable.

Marek, R., Ly, C., 2009. Ontological Approach To Development Of Computing With Words Based Systems. International Journal of Approximate Reasoning 50 (1), 72–91.

Martin, T., Shen, Y., Azvine, B., 2008. Incremental evolution of fuzzy grammar fragments to enhance instance matching and text mining. IEEE Transactions on Fuzzy Systems 16 (6), 1425–1438.

McCarthy, J., Hayes, P., 1969. Some philosophical problems from the standpoint of artificial intelligence. Machine Intelligence 4, 463–502.

McCroskey, J. C., 2001. An introduction to Rhetorical Analysis. Allyn and Bacon.

McGuinness, D., Van Harmelen, F., 2004. OWL web ontology language overview. W3C recommendation.

Mehta, K., Bhattacharyya, S., 2004. Adequacy of training data for evolutionary mining of trading rules. Decision support systems 37 (4), 461–474.

Menkveld, A. J., 2011. High frequency trading and the new-market makers. SSRN eLibrary.

Michaely, R., Thaler, R., Womack, K., 1995. Price reactions to dividend initiations and omissions: Overreaction or drift. Journal of Finance 50 (2), 573–608.

Milea, V., Almeida, R., Kaymak, U., Frasincar, F., 2010a. A fuzzy model of the msci euro index based on content analysis of european central bank statements. In: The 2010 IEEE World Congress on Computational Intelligence (WCCI 2010). IEEE, pp. 154–160.

Milea, V., Frasincar, F., Kaymak, U., 2008a. Knowledge engineering in a temporal semantic web context. In: The Eighth International Conference on Web Engineering (ICWE 2008). IEEE Computer Society Press, pp. 65–74.

Milea, V., Frasincar, F., Kaymak, U., 2008b. The tOWL web ontology language. In: The 20th Belgian-Dutch Conference on Artificial Intelligence (BNAIC 2008).

Milea, V., Frasincar, F., Kaymak, U., 2012a. tOWL: A temporal web ontology language. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 42 (1).

Milea, V., Frasincar, F., Kaymak, U., di Noia, T., 2007. An OWL-based approach towards representing time in web information systems. In: The 4th International Workshop of Web Information Systems Modeling Workshop (WISM 2007). Tapir Academic Press, pp. 791–802.

Milea, V., Frasincar, F., Kaymak, U., Houben, G., 2012b. Temporal optimisations and temporal cardinality in the tOWL language. International Journal of Web Engineering and Technology, to appear.

Milea, V., Mrissa, M., van der Sluijs, K., Kaymak, U., 2008c. On temporal cardinality in the context of the TOWL language. In: The 5th International Workshop of Web Information Systems Modeling Workshop (WISM 2008). Springer, pp. 457–466.

Milea, V., Sharef, N., Almeida, R., Kaymak, U., Frasincar, F., 2010b. Prediction of the MSCI EURO index based on fuzzy grammar fragments extracted from european central bank statements. In: The 2010 International Conference of Soft Computing and Pattern Recognition (SoCPaR 2010). IEEE, pp. 231–236.

Mitchell, M., Mulherin, J., 1994. The impact of public information on the stock market. Journal of Finance 49 (3), 923–950.

Mitra, G., Mitra, L., 2011. The Handbook of News Analytics in Finance. Vol. 596. Wiley.

Mittermayer, M., Knolmayer, G., 2006. Text mining systems for market response to news: A survey. Institute of Information Systems University of Bern. http://www. ie. iwi. unibe. ch/publikationen/berichte/resource/WP-184. pdf.

Morgan, J., Stocken, P., 2003. An analysis of stock recommendations. RAND Journal of Economics 34 (1), 183–203.

Motik, B., 2010. Representing and querying validity time in rdf and owl: a logic-based approach. International Semantic Web Conference (ISWC 2010), 550–565.

Motik, B., Patel-Schneider, P., Parsia, B., Bock, C., Fokoue, A., Haase, P., Hoekstra, R., Horrocks, I., Ruttenberg, A., Sattler, U., et al., 2009. OWL 2 web ontology language: Structural specification and functional-style syntax. W3C Recommendation.

Neuendorf, K. A., 2002. The content analysis guidebook. Sage Publications.

Noy, N., Klein, M., 2004. Ontology evolution: Not the same as schema evolution. Knowledge and Information Systems 6 (4), 428–440.

Noy, N., Rector, A., 2005. Defining n-ary relations on the semantic web. Working Draft for the W3C Semantic Web best practices group.

Ozsoyoglu, G., Snodgrass, R., 1995. Temporal and real-time databases: A survey. IEEE Transactions on Knowledge and Data Engineering 7 (4), 513–532.

Patel-Schneider, P., Hayes, P., Horrocks, I., 2004. Web ontology language (OWL) abstract syntax and semantics. W3C Recommendation.

Perkins, W., Austin, A., 1990. Adding temporal reasoning to expert-system-building environments. IEEE Expert 5 (1), 23–30.

Porter, M., 1980. An algorithm for suffix stripping. Program 14 (3), 130–137.

Ramnath, S., Rock, S., Shane, P., 2006. A review of research related to financial analysts forecasts and stock recommendations. Unpublished working paper, Georgetown University.

Roger D. Wimmer, J. R. D., 2010. Wadsworth Publishing.

Rosen, R., 2006. Merger momentum and investor sentiment: The stock market reaction to merger announcements. The Journal of Business 79 (2), 987–1017.

Sakaji, H., Sakai, H., Masuyama, S., 2008. Automatic extraction of basis expressions that indicate economic trends. Advances in Knowledge Discovery and Data Mining, 977–984.

Schmid, H., 1995. Treetagger – a language independent part-of-speech tagger. Tech. rep.

Schmid, H., 1997. Probabilistic part-of-speech tagging using decision trees. In: Jones, D., Somers, H. (Eds.), New Methods in Language Processing. Studies in Computational Linguistics. UCL Press, London, GB, pp. 154–164.

SEC, 2010. Concept release on equity market structure. Tech. Rep. 34-61358, Securities and Exchange Commission.
URL http://www.sec.gov/rules/concept/2010/34-61358.pdf

Setnes, M., Babuska, R., Kaymak, U., van Nauta Lemke, H., 1998. Similarity measures in fuzzy rule base simplification. IEEE Transactions on Systems, Man, and Cybernetics, Part B 28 (3), 376–386.

Shadbolt, N., Hall, W., Berners-Lee, T., 2006. The semantic web revisited. Intelligent Systems, IEEE 21 (3), 96–101.

Sharef, N., Martin, T., Shen, Y., 2009. Minimal combination for incremental grammar fragment learning. In: The Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference (IFSA-EUSFLAT 2009). pp. 909–914.

Sharef, N., Shen, Y., 2010. Text fragment extraction using incremental evolving fuzzy grammar fragments learner. In: The 2010 IEEE World Congress on Computational Intelligence (WCCI 2010). pp. 3026–3033.

Sharpe, W., 1966. Mutual fund performance. The Journal of Business 39 (1), 119–138.

Sornette, D., von der Becke, S., 2011. An evaluation of risks posed by high-speed algorithmic trading. Tech. Rep. 11-1226-dr7, Foresight.
URL http://www.bis.gov.uk/assets/bispartners/foresight/docs/computer-trading/11-1226-dr7-crashes-and-high-frequency-trading.pdf

Spaccapietra, S., Parent, C., Zimanyi, E., 1998. Modeling time from a conceptual perspective. In: Seventh International Conference on Information and Knowledge Management (CIKM 1998). ACM, pp. 432–440.

Steiner, A., Norrie, M., 1997. Implementing temporal databases in object-oriented systems. In: Fifth International Conference on Database Systems for Advanced Applications (DASFAA 1997). World Scientific Press, pp. 381–390.

Stone, P., Dunphy, D., Smith, M., 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press Cambridge.

STW Thesaurus for Economics, 2011. Leibniz information centre for economics. Last Accessed: August 2nd, 2011.
URL http://zbw.eu/stw/versions/latest/about

Takagi, T., Sugeno, M., 1985. Fuzzy identification of system and its applications to modelling and control. IEEE Transactions on Systems, Man, and Cybernetics 15 (1), 116–132.

Tauzovich, B., 1991. Toward temporal extensions to the entity-relationship model. In: 10th International Conference on the Entity Relationship Approach (ER 1991). ER Institute, pp. 163–179.

Tetlock, P., 2007. Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance 62 (3), 1139–1168.

Tetlock, P., Saar-Tsechansky, M., Macskassy, S., 2008. More than words: quantifying language to measure firms' fundamentals. The Journal of Finance 63 (3), 1437–1467.

The Stanford Encyclopedia of Philosophy, 2003. Zalta, e.n. Last Accessed: January 10th, 2012.
URL `http://plato.stanford.edu/`

Thomson One Banker, 2011. Thomson. Last Accessed: August 2nd, 2011.
URL `http://banker.thomsonib.com/`

Toutanova, K., Klein, D., Manning, C., Singer, Y., 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In: The 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL 2003). Association for Computational Linguistics, pp. 252–259.

Toutanova, K., Manning, C., 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: The Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000. Association for Computational Linguistics, pp. 63–70.

Wang, F., Zaniolo, C., 2004. XBiT: an XML-based bitemporal data model. In: 23rd International Conference on Conceptual Modeling (ER 2004). Springer, pp. 810–824.

Warner, J., Watts, R., Wruck, K., 1988. Stock prices and top management changes. Journal of Financial Economics 20 (1), 461–92.

Welty, C., Fikes, R., 2006. A reusable ontology for fluents in OWL. In: The Fourth International Conference on Formal Ontology in Information Systems (FOIS 2006). IOS Press, pp. 226–336.

Widagdo, T., 2007. Managing referential integrity in bitemporal databases. In: International Conference on Electrical Engineering and Informatics (ICEEI 2007). pp. 305–308.

Wijsen, J., 1999. Temporal fds on complex objects. ACM Transactions on Database Systems 24 (1), 127–176.

Zhang, J., Kawai, Y., Kumamoto, T., Tanaka, K., 2009. A novel visualization method for distinction of web news sentiment. Springer, pp. 181–194.

Zhang, X., 2006. Information uncertainty and stock returns. The Journal of Finance 61 (1), 105–137.

Zhao, H., 2007. A multi-objective genetic programming approach to developing pareto optimal decision trees. Decision Support Systems 43 (3), 809–826.

Zimanyi, E., Parent, C., Spaccapietra, S., Pirotte, A., 1997. TERC+: a temporal conceptual model. In: International Symposium on Digital Media Information Base (DMIB 1997).

About the author

Viorel Milea was born on April 27th, 1982 in Bucharest, Romania. After finishing high-school in Romania, he moved to the Netherlands for his undergraduate studies. He obtained the MSc degree in Informatics & Economics from Erasmus University Rotterdam, the Netherlands, in 2006. Afterwards, he joined the Econometric Institute for his PhD research within the TOWL project, a project co-financed by the European Union. Currently, he is an assistant professor at the Econometric Institute of the Erasmus School of Economics at the Erasmus University Rotterdam.

His research interests cover the use of Semantic Web technologies for enhancing the current state-of-the-art in automated trading, Semantic Web theory, management information systems, content analysis, and nature-inspired classification and optimization techniques.

ERASMUS RESEARCH INSTITUTE OF MANAGEMENT (ERIM)

ERIM PH.D. SERIES RESEARCH IN MANAGEMENT

The ERIM PhD Series contains PhD dissertations in the field of Research in Management defended at Erasmus University Rotterdam and supervised by senior researchers affiliated to the Erasmus Research Institute of Management (ERIM). All dissertations in the ERIM PhD Series are available in full text through the ERIM Electronic Series Portal: http://hdl.handle.net/1765/1 ERIM is the joint research institute of the Rotterdam School of Management (RSM) and the Erasmus School of Economics at the Erasmus University Rotterdam (EUR).

DISSERTATIONS LAST FIVE YEARS

Acciaro, M., *Bundling Strategies in Global Supply Chains.* Promoter(s): Prof.dr. H.E. Haralambides, EPS-2010-197-LIS, http://hdl.handle.net/1765/19742

Agatz, N.A.H., *Demand Management in E-Fulfillment*. Promoter(s): Prof.dr.ir. J.A.E.E. van Nunen, EPS-2009-163-LIS, http://hdl.handle.net/1765/15425

Alexiev, A., *Exploratory Innovation: The Role of Organizational and Top Management Team Social Capital.* Promoter(s): Prof.dr. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2010-208-STR, http://hdl.handle.net/1765/20632

Asperen, E. van, *Essays on Port, Container, and Bulk Chemical Logistics Optimization*. Promoter(s): Prof.dr.ir. R. Dekker, EPS-2009-181-LIS, http://hdl.handle.net/1765/17626

Benning, T.M., *A Consumer Perspective on Flexibility in Health Care: Priority Access Pricing and Customized Care*, Promoter(s): Prof.dr.ir. B.G.C. Dellaert, EPS-2011-241-MKT, http://hdl.handle.net/1765/23670

Betancourt, N.E., *Typical Atypicality: Formal and Informal Institutional Conformity, Deviance, and Dynamics*, Promoter(s): Prof.dr. B. Krug, EPS-2012-262-ORG, http://hdl.handle.net/1765/32345

Bezemer, P.J., *Diffusion of Corporate Governance Beliefs: Board Independence and the Emergence of a Shareholder Value Orientation in the Netherlands.* Promoter(s): Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2009-192-STR, http://hdl.handle.net/1765/18458

Binken, J.L.G., *System Markets: Indirect Network Effects in Action, or Inaction,* Promoter(s): Prof.dr. S. Stremersch, EPS-2010-213-MKT, http://hdl.handle.net/1765/21186

Blitz, D.C., *Benchmarking Benchmarks,* Promoter(s): Prof.dr. A.G.Z. Kemna & Prof.dr. W.F.C. Verschoor, EPS-2011-225-F&A, http://hdl.handle.net/1765/22624

Borst, W.A.M., *Understanding Crowdsourcing: Effects of Motivation and Rewards on Participation and Performance in Voluntary Online Activities,* Promoter(s): Prof.dr.ir. J.C.M. van den Ende & Prof.dr.ir. H.W.G.M. van Heck, EPS-2010-221-LIS, http://hdl.handle.net/1765/ 21914

Budiono, D.P., *The Analysis of Mutual Fund Performance: Evidence from U.S. Equity Mutual Funds,* Promoter(s): Prof.dr. M.J.C.M. Verbeek, EPS-2010-185-F&A, http://hdl.handle.net/1765/18126

Burger, M.J., *Structure and Cooptition in Urban Networks*, Promoter(s): Prof.dr. G.A. van der Knaap & Prof.dr. H.R. Commandeur, EPS-2011-243-ORG, http://hdl.handle.net/1765/26178

Camacho, N.M., *Health and Marketing; Essays on Physician and Patient Decision-making*, Promoter(s): Prof.dr. S. Stremersch, EPS-2011-237-MKT, http://hdl.handle.net/1765/23604

Carvalho de Mesquita Ferreira, L., *Attention Mosaics: Studies of Organizational Attention*, Promoter(s): Prof.dr. P.M.A.R. Heugens & Prof.dr. J. van Oosterhout, EPS-2010-205-ORG, http://hdl.handle.net/1765/19882

Chen, C.-M., *Evaluation and Design of Supply Chain Operations Using DEA*, Promoter(s): Prof.dr. J.A.E.E. van Nunen, EPS-2009-172-LIS, http://hdl.handle.net/1765/16181

Defilippi Angeldonis, E.F., *Access Regulation for Naturally Monopolistic Port Terminals: Lessons from Regulated Network Industries,* Promoter(s): Prof.dr. H.E. Haralambides, EPS-2010-204-LIS, http://hdl.handle.net/1765/19881

Deichmann, D., *Idea Management: Perspectives from Leadership, Learning, and Network Theory*, Promoter(s): Prof.dr.ir. J.C.M. van den Ende, EPS-2012-255-ORG, http://hdl.handle.net/1765/ 31174

Desmet, P.T.M., *In Money we Trust? Trust Repair and the Psychology of Financial Compensations*, Promoter(s): Prof.dr. D. De Cremer & Prof.dr. E. van Dijk, EPS-2011-232-ORG, http://hdl.handle.net/1765/23268

Diepen, M. van, *Dynamics and Competition in Charitable Giving*, Promoter(s): Prof.dr. Ph.H.B.F. Franses, EPS-2009-159-MKT, http://hdl.handle.net/1765/14526

Dietvorst, R.C., *Neural Mechanisms Underlying Social Intelligence and Their Relationship with the Performance of Sales Managers,* Promoter(s): Prof.dr. W.J.M.I. Verbeke, EPS-2010-215-MKT, http://hdl.handle.net/1765/21188

Dietz, H.M.S., *Managing (Sales)People towards Performance: HR Strategy, Leadership & Teamwork*, Promoter(s): Prof.dr. G.W.J. Hendrikse, EPS-2009-168-ORG, http://hdl.handle.net/1765/16081

Dollevoet, T.A.B., *Delay Management and Dispatching in Railways*, Promoter(s): Prof.dr. A.P.M. Wagelmans, EPS-2013-272-LIS, http://hdl.handle.net/1765/1

Doorn, S. van, *Managing Entrepreneurial Orientation*, Promoter(s): Prof.dr. J.J.P. Jansen, Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2012-258-STR, http://hdl.handle.net/1765/32166

Douwens-Zonneveld, M.G., *Animal Spirits and Extreme Confidence: No Guts, No Glory*, Promoter(s): Prof.dr. W.F.C. Verschoor, EPS-2012-257-F&A, http://hdl.handle.net/1765/31914

Duca, E., *The Impact of Investor Demand on Security Offerings*, Promoter(s): Prof.dr. A. de Jong, EPS-2011-240-F&A, http://hdl.handle.net/1765/26041

Eck, N.J. van, *Methodological Advances in Bibliometric Mapping of Science*, Promoter(s): Prof.dr.ir. R. Dekker, EPS-2011-247-LIS, http://hdl.handle.net/1765/26509

Eijk, A.R. van der, *Behind Networks: Knowledge Transfer, Favor Exchange and Performance*, Promoter(s): Prof.dr. S.L. van de Velde & Prof.dr.drs. W.A. Dolfsma, EPS-2009-161-LIS, http://hdl.handle.net/1765/14613

Essen, M. van, *An Institution-Based View of Ownership*, Promoter(s): Prof.dr. J. van Oosterhout & Prof.dr. G.M.H. Mertens, EPS-2011-226-ORG, http://hdl.handle.net/1765/22643

Feng, L., *Motivation, Coordination and Cognition in Cooperatives,* Promoter(s): Prof.dr. G.W.J. Hendrikse, EPS-2010-220-ORG, http://hdl.handle.net/1765/21680

Gertsen, H.F.M., *Riding a Tiger without Being Eaten: How Companies and Analysts Tame Financial Restatements and Influence Corporate Reputation*, Promoter(s): Prof.dr. C.B.M. van Riel, EPS-2009-171-ORG, http://hdl.handle.net/1765/16098

Gharehgozli, A.H., *Developing New Methods for Efficient Container Stacking Operations*, Promoter(s): Prof.dr.ir. M.B.M. de Koster, EPS-2012-269-LIS, http://hdl.handle.net/1765/ 37779

Gijsbers, G.W., *Agricultural Innovation in Asia: Drivers, Paradigms and Performance,* Promoter(s): Prof.dr. R.J.M. van Tulder, EPS-2009-156-ORG, http://hdl.handle.net/1765/14524

Gils, S. van, *Morality in Interactions: On the Display of Moral Behavior by Leaders and Employees*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2012-270-ORG, http://hdl.handle.net/1765/ 38028

Ginkel-Bieshaar, M.N.G. van, *The Impact of Abstract versus Concrete Product Communications on Consumer Decision-making Processes*, Promoter(s): Prof.dr.ir. B.G.C. Dellaert, EPS-2012-256-MKT, http://hdl.handle.net/1765/31913

Gkougkousi, X., *Empirical Studies in Financial Accounting*, Promoter(s): Prof.dr. G.M.H. Mertens & Prof.dr. E. Peek, EPS-2012-264-F&A, http://hdl.handle.net/1765/37170

Gong, Y., *Stochastic Modelling and Analysis of Warehouse Operations*, Promoter(s): Prof.dr. M.B.M. de Koster & Prof.dr. S.L. van de Velde, EPS-2009-180-LIS, http://hdl.handle.net/1765/16724

Greeven, M.J., *Innovation in an Uncertain Institutional Environment: Private Software Entrepreneurs in Hangzhou, China*, Promoter(s): Prof.dr. B. Krug, EPS-2009-164-ORG, http://hdl.handle.net/1765/15426

Hakimi, N.A, *Leader Empowering Behaviour: The Leader's Perspective: Understanding the Motivation behind Leader Empowering Behaviour*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2010-184-ORG, http://hdl.handle.net/1765/17701

Hensmans, M., *A Republican Settlement Theory of the Firm: Applied to Retail Banks in England and the Netherlands (1830-2007)*, Promoter(s): Prof.dr. A. Jolink & Prof.dr. S.J. Magala, EPS-2010-193-ORG, http://hdl.handle.net/1765/19494

Hernandez Mireles, C., *Marketing Modeling for New Products,* Promoter(s): Prof.dr. P.H. Franses, EPS-2010-202-MKT, http://hdl.handle.net/1765/19878

Heyden, M.L.M., *Essays on Upper Echelons & Strategic Renewal: A Multilevel Contingency Approach*, Promoter(s): Prof.dr. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2012-259-STR, http://hdl.handle.net/1765/32167

Hoever, I.J., *Diversity and Creativity: In Search of Synergy*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2012-267-ORG, http://hdl.handle.net/1765/37392

Hoogendoorn, B., *Social Entrepreneurship in the Modern Economy: Warm Glow, Cold Feet*, Promoter(s): Prof.dr. H.P.G. Pennings & Prof.dr. A.R. Thurik, EPS-2011-246-STR, http://hdl.handle.net/1765/26447

Hoogervorst, N., *On The Psychology of Displaying Ethical Leadership: A Behavioral Ethics Approach*, Promoter(s): Prof.dr. D. De Cremer & Dr. M. van Dijke, EPS-2011-244-ORG, http://hdl.handle.net/1765/26228

Huang, X., *An Analysis of Occupational Pension Provision: From Evaluation to Redesign,* Promoter(s): Prof.dr. M.J.C.M. Verbeek & Prof.dr. R.J. Mahieu, EPS-2010-196-F&A, http://hdl.handle.net/1765/19674

Hytönen, K.A. *Context Effects in Valuation, Judgment and Choice*, Promoter(s): Prof.dr.ir. A. Smidts, EPS-2011-252-MKT, http://hdl.handle.net/1765/30668

Jalil, M.N., *Customer Information Driven After Sales Service Management: Lessons from Spare Parts Logistics,* Promoter(s): Prof.dr. L.G. Kroon, EPS-2011-222-LIS, http://hdl.handle.net/1765/22156

Jaspers, F.P.H., *Organizing Systemic Innovation*, Promoter(s): Prof.dr.ir. J.C.M. van den Ende, EPS-2009-160-ORG, http://hdl.handle.net/1765/14974

Jiang, T., *Capital Structure Determinants and Governance Structure Variety in Franchising*, Promoter(s): Prof.dr. G. Hendrikse & Prof.dr. A. de Jong, EPS-2009-158-F&A, http://hdl.handle.net/1765/14975

Jiao, T., *Essays in Financial Accounting*, Promoter(s): Prof.dr. G.M.H. Mertens, EPS-2009-176-F&A, http://hdl.handle.net/1765/16097

Kaa, G. van, *Standard Battles for Complex Systems: Empirical Research on the Home Network,* Promoter(s): Prof.dr.ir. J. van den Ende & Prof.dr.ir. H.W.G.M. van Heck, EPS-2009-166-ORG, http://hdl.handle.net/1765/16011

Kagie, M., *Advances in Online Shopping Interfaces: Product Catalog Maps and Recommender Systems,* Promoter(s): Prof.dr. P.J.F. Groenen, EPS-2010-195-MKT, http://hdl.handle.net/1765/19532

Kappe, E.R., *The Effectiveness of Pharmaceutical Marketing*, Promoter(s): Prof.dr. S. Stremersch, EPS-2011-239-MKT, http://hdl.handle.net/1765/23610

Karreman, B., *Financial Services and Emerging Markets,* Promoter(s): Prof.dr. G.A. van der Knaap & Prof.dr. H.P.G. Pennings, EPS-2011-223-ORG, http://hdl.handle.net/1765/ 22280

Kwee, Z., *Investigating Three Key Principles of Sustained Strategic Renewal: A Longitudinal Study of Long-Lived Firms,* Promoter(s): Prof.dr.ir. F.A.J. Van den Bosch & Prof.dr. H.W. Volberda, EPS-2009-174-STR, http://hdl.handle.net/1765/16207

Lam, K.Y., *Reliability and Rankings*, Promoter(s): Prof.dr. P.H.B.F. Franses, EPS-2011-230-MKT, http://hdl.handle.net/1765/22977

Lander, M.W., *Profits or Professionalism? On Designing Professional Service Firms*, Promoter(s): Prof.dr. J. van Oosterhout & Prof.dr. P.P.M.A.R. Heugens, EPS-2012-253-ORG, http://hdl.handle.net/1765/30682

Langhe, B. de, *Contingencies: Learning Numerical and Emotional Associations in an Uncertain World*, Promoter(s): Prof.dr.ir. B. Wierenga & Prof.dr. S.M.J. van Osselaer, EPS-2011-236-MKT, http://hdl.handle.net/1765/23504

Larco Martinelli, J.A., *Incorporating Worker-Specific Factors in Operations Management Models,* Promoter(s): Prof.dr.ir. J. Dul & Prof.dr. M.B.M. de Koster, EPS-2010-217-LIS, http://hdl.handle.net/1765/21527

Li, T., *Informedness and Customer-Centric Revenue Management,* Promoter(s): Prof.dr. P.H.M. Vervest & Prof.dr.ir. H.W.G.M. van Heck, EPS-2009-146-LIS, http://hdl.handle.net/1765/14525

Lovric, M., *Behavioral Finance and Agent-Based Artificial Markets*, Promoter(s): Prof.dr. J. Spronk & Prof.dr.ir. U. Kaymak, EPS-2011-229-F&A, http://hdl.handle.net/1765/ 22814

Maas, K.E.G., *Corporate Social Performance: From Output Measurement to Impact Measurement,* Promoter(s): Prof.dr. H.R. Commandeur, EPS-2009-182-STR, http://hdl.handle.net/1765/17627

Markwat, T.D., *Extreme Dependence in Asset Markets Around the Globe*, Promoter(s): Prof.dr. D.J.C. van Dijk, EPS-2011-227-F&A, http://hdl.handle.net/1765/22744

Mees, H., *Changing Fortunes: How China's Boom Caused the Financial Crisis*, Promoter(s): Prof.dr. Ph.H.B.F. Franses, EPS-2012-266-MKT, http://hdl.handle.net/1765/34930

Meuer, J., *Configurations of Inter-Firm Relations in Management Innovation: A Study in China's Biopharmaceutical Industry*, Promoter(s): Prof.dr. B. Krug, EPS-2011-228-ORG, http://hdl.handle.net/1765/22745

Mihalache, O.R., *Stimulating Firm Innovativeness: Probing the Interrelations between Managerial and Organizational Determinants*, Promoter(s): Prof.dr. J.J.P. Jansen, Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2012-260-S&E, http://hdl.handle.net/1765/32343

Moonen, J.M., *Multi-Agent Systems for Transportation Planning and Coordination*, Promoter(s): Prof.dr. J. van Hillegersberg & Prof.dr. S.L. van de Velde, EPS-2009-177-LIS, http://hdl.handle.net/1765/16208

Nederveen Pieterse, A., *Goal Orientation in Teams: The Role of Diversity*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2009-162-ORG, http://hdl.handle.net/1765/15240

Nielsen, L.K., *Rolling Stock Rescheduling in Passenger Railways: Applications in Short-term Planning and in Disruption Management*, Promoter(s): Prof.dr. L.G. Kroon, EPS-2011-224-LIS, http://hdl.handle.net/1765/22444

Niesten, E.M.M.I., *Regulation, Governance and Adaptation: Governance Transformations in the Dutch and French Liberalizing Electricity Industries*, Promoter(s): Prof.dr. A. Jolink & Prof.dr. J.P.M. Groenewegen, EPS-2009-170-ORG, http://hdl.handle.net/1765/16096

Nijdam, M.H., *Leader Firms: The Value of Companies for the Competitiveness of the Rotterdam Seaport Cluster,* Promoter(s): Prof.dr. R.J.M. van Tulder, EPS-2010-216-ORG, http://hdl.handle.net/1765/21405

Noordegraaf-Eelens, L.H.J., *Contested Communication: A Critical Analysis of Central Bank Speech,* Promoter(s): Prof.dr. Ph.H.B.F. Franses, EPS-2010-209-MKT, http://hdl.handle.net/1765/21061

Nuijten, A.L.P., *Deaf Effect for Risk Warnings: A Causal Examination applied to Information Systems Projects*, Promoter(s): Prof.dr. G. van der Pijl & Prof.dr. H. Commandeur & Prof.dr. M. Keil, EPS-2012-263-S&E, http://hdl.handle.net/1765/34928

Nuijten, I., *Servant Leadership: Paradox or Diamond in the Rough? A Multidimensional Measure and Empirical Evidence*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2009-183-ORG, http://hdl.handle.net/1765/21405

Oosterhout, M., van, *Business Agility and Information Technology in Service Organizations,* Promoter(s): Prof,dr.ir. H.W.G.M. van Heck, EPS-2010-198-LIS, http://hdl.handle.net/1765/19805

Oostrum, J.M., van, *Applying Mathematical Models to Surgical Patient Planning*, Promoter(s): Prof.dr. A.P.M. Wagelmans, EPS-2009-179-LIS, http://hdl.handle.net/1765/16728

Osadchiy, S.E., *The Dynamics of Formal Organization: Essays on Bureaucracy and Formal Rules*, Promoter(s): Prof.dr. P.P.M.A.R. Heugens, EPS-2011-231-ORG, http://hdl.handle.net/1765/23250

Otgaar, A.H.J., *Industrial Tourism: Where the Public Meets the Private,* Promoter(s): Prof.dr. L. van den Berg, EPS-2010-219-ORG, http://hdl.handle.net/1765/21585

Ozdemir, M.N., *Project-level Governance, Monetary Incentives and Performance in Strategic R&D Alliances*, Promoter(s): Prof.dr.ir. J.C.M. van den Ende, EPS-2011-235-LIS, http://hdl.handle.net/1765/23550

Peers, Y., *Econometric Advances in Diffusion Models*, Promoter(s): Prof.dr. Ph.H.B.F. Franses, EPS-2011-251-MKT, http://hdl.handle.net/1765/ 30586

Pince, C., *Advances in Inventory Management: Dynamic Models,* Promoter(s): Prof.dr.ir. R. Dekker, EPS-2010-199-LIS, http://hdl.handle.net/1765/19867

Porras Prado, M., *The Long and Short Side of Real Estate, Real Estate Stocks, and Equity*, Promoter(s): Prof.dr. M.J.C.M. Verbeek, EPS-2012-254-F&A, http://hdl.handle.net/1765/30848

Potthoff, D., *Railway Crew Rescheduling: Novel Approaches and Extensions,* Promoter(s): Prof.dr. A.P.M. Wagelmans & Prof.dr. L.G. Kroon, EPS-2010-210-LIS, http://hdl.handle.net/1765/21084

Poruthiyil, P.V., *Steering Through: How Organizations Negotiate Permanent Uncertainty and Unresolvable Choices*, Promoter(s): Prof.dr. P.P.M.A.R. Heugens & Prof.dr. S. Magala, EPS-2011-245-ORG, http://hdl.handle.net/1765/26392

Pourakbar, M. *End-of-Life Inventory Decisions of Service Parts*, Promoter(s): Prof.dr.ir. R. Dekker, EPS-2011-249-LIS, http://hdl.handle.net/1765/30584

Rijsenbilt, J.A., *CEO Narcissism; Measurement and Impact*, Promoter(s): Prof.dr. A.G.Z. Kemna & Prof.dr. H.R. Commandeur, EPS-2011-238-STR, http://hdl.handle.net/1765/ 23554

Roelofsen, E.M., *The Role of Analyst Conference Calls in Capital Markets*, Promoter(s): Prof.dr. G.M.H. Mertens & Prof.dr. L.G. van der Tas RA, EPS-2010-190-F&A, http://hdl.handle.net/1765/18013

Rosmalen, J. van, *Segmentation and Dimension Reduction: Exploratory and Model-Based Approaches*, Promoter(s): Prof.dr. P.J.F. Groenen, EPS-2009-165-MKT, http://hdl.handle.net/1765/15536

Roza, M.W., *The Relationship between Offshoring Strategies and Firm Performance: Impact of Innovation, Absorptive Capacity and Firm Size,* Promoter(s): Prof.dr. H.W. Volberda & Prof.dr.ing. F.A.J. van den Bosch, EPS-2011-214-STR, http://hdl.handle.net/1765/22155

Rus, D., *The Dark Side of Leadership: Exploring the Psychology of Leader Self-serving Behavior*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2009-178-ORG, http://hdl.handle.net/1765/16726

Schellekens, G.A.C., *Language Abstraction in Word of Mouth,* Promoter(s): Prof.dr.ir. A. Smidts, EPS-2010-218-MKT, ISBN: 978-90-5892-252-6, http://hdl.handle.net/1765/21580

Sotgiu, F., *Not All Promotions are Made Equal: From the Effects of a Price War to Cross-chain Cannibalization,* Promoter(s): Prof.dr. M.G. Dekimpe & Prof.dr.ir. B. Wierenga, EPS-2010-203-MKT, http://hdl.handle.net/1765/19714

Srour, F.J., *Dissecting Drayage: An Examination of Structure, Information, and Control in Drayage Operations*, Promoter(s): Prof.dr. S.L. van de Velde, EPS-2010-186-LIS, http://hdl.handle.net/1765/18231

Sweldens, S.T.L.R., *Evaluative Conditioning 2.0: Direct versus Associative Transfer of Affect to Brands,* Promoter(s): Prof.dr. S.M.J. van Osselaer, EPS-2009-167-MKT, http://hdl.handle.net/1765/16012

Teixeira de Vasconcelos, M., *Agency Costs, Firm Value, and Corporate Investment*, Promoter(s): Prof.dr. P.G.J. Roosenboom, EPS-2012-265-F&A, http://hdl.handle.net/1765/37265

Tempelaar, M.P., *Organizing for Ambidexterity: Studies on the Pursuit of Exploration and Exploitation through Differentiation, Integration, Contextual and Individual Attributes,* Promoter(s): Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2010-191-STR, http://hdl.handle.net/1765/18457

Tiwari, V., *Transition Process and Performance in IT Outsourcing: Evidence from a Field Study and Laboratory Experiments,* Promoter(s): Prof.dr.ir. H.W.G.M. van Heck & Prof.dr. P.H.M. Vervest, EPS-2010-201-LIS, http://hdl.handle.net/1765/19868

Tröster, C., *Nationality Heterogeneity and Interpersonal Relationships at Work*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2011-233-ORG, http://hdl.handle.net/1765/23298

Tsekouras, D., *No Pain No Gain: The Beneficial Role of Consumer Effort in Decision Making*, Promoter(s): Prof.dr.ir. B.G.C. Dellaert, EPS-2012-268-MKT, http://hdl.handle.net/1765/ 37542

Tzioti, S., *Let Me Give You a Piece of Advice: Empirical Papers about Advice Taking in Marketing,* Promoter(s): Prof.dr. S.M.J. van Osselaer & Prof.dr.ir. B. Wierenga, EPS-2010-211-MKT, hdl.handle.net/1765/21149

Vaccaro, I.G., *Management Innovation: Studies on the Role of Internal Change Agents,* Promoter(s): Prof.dr. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2010-212-STR, hdl.handle.net/1765/21150

Verheijen, H.J.J., *Vendor-Buyer Coordination in Supply Chains,* Promoter(s): Prof.dr.ir. J.A.E.E. van Nunen, EPS-2010-194-LIS, http://hdl.handle.net/1765/19594

Verwijmeren, P., *Empirical Essays on Debt, Equity, and Convertible Securities*, Promoter(s): Prof.dr. A. de Jong & Prof.dr. M.J.C.M. Verbeek, EPS-2009-154-F&A, http://hdl.handle.net/1765/14312

Vlam, A.J., *Customer First? The Relationship between Advisors and Consumers of Financial Products*, Promoter(s): Prof.dr. Ph.H.B.F. Franses, EPS-2011-250-MKT, http://hdl.handle.net/1765/30585

Waard, E.J. de, *Engaging Environmental Turbulence: Organizational Determinants for Repetitive Quick and Adequate Responses*, Promoter(s): Prof.dr. H.W. Volberda & Prof.dr. J. Soeters, EPS-2010-189-STR, http://hdl.handle.net/1765/18012

Wall, R.S., *Netscape: Cities and Global Corporate Networks,* Promoter(s): Prof.dr. G.A. van der Knaap, EPS-2009-169-ORG, http://hdl.handle.net/1765/16013

Waltman, L., *Computational and Game-Theoretic Approaches for Modeling Bounded Rationality*, Promoter(s): Prof.dr.ir. R. Dekker & Prof.dr.ir. U. Kaymak, EPS-2011-248-LIS, http://hdl.handle.net/1765/26564

Wang, Y., *Information Content of Mutual Fund Portfolio Disclosure*, Promoter(s): Prof.dr. M.J.C.M. Verbeek, EPS-2011-242-F&A, http://hdl.handle.net/1765/26066

Weerdt, N.P. van der, *Organizational Flexibility for Hypercompetitive Markets: Empirical Evidence of the Composition and Context Specificity of Dynamic Capabilities and Organization Design Parameters,* Promoter(s): Prof.dr. H.W. Volberda, EPS-2009-173-STR, http://hdl.handle.net/1765/16182

Wubben, M.J.J., *Social Functions of Emotions in Social Dilemmas*, Promoter(s): Prof.dr. D. De Cremer & Prof.dr. E. van Dijk, EPS-2009-187-ORG, http://hdl.handle.net/1765/18228

Xu, Y., *Empirical Essays on the Stock Returns, Risk Management, and Liquidity Creation of Banks*, Promoter(s): Prof.dr. M.J.C.M. Verbeek, EPS-2010-188-F&A, http://hdl.handle.net/1765/18125

Yang, J., *Towards the Restructuring and Co-ordination Mechanisms for the Architecture of Chinese Transport Logistics*, Promoter(s): Prof.dr. H.E. Harlambides, EPS-2009-157-LIS, http://hdl.handle.net/1765/14527

Zhang, D., *Essays in Executive Compensation*, Promoter(s): Prof.dr. I. Dittmann, EPS-2012-261-F&A, http://hdl.handle.net/1765/32344

Zhang, X., *Scheduling with Time Lags,* Promoter(s): Prof.dr. S.L. van de Velde, EPS-2010-206-LIS, http://hdl.handle.net/1765/19928

Zhou, H., *Knowledge, Entrepreneurship and Performance: Evidence from Country-level and Firm-level Studies,* Promoter(s): Prof.dr. A.R. Thurik & Prof.dr. L.M. Uhlaner, EPS-2010-207-ORG, http://hdl.handle.net/1765/20634

Zwan, P.W. van der, *The Entrepreneurial Process: An International Analysis of Entry and Exit*, Promoter(s): Prof.dr. A.R. Thurik & Prof.dr. P.J.F. Groenen, EPS-2011-234-ORG, http://hdl.handle.net/1765/23422

Erasmus Research Institute of Management - ERiM

NEWS ANALYTICS FOR FINANCIAL DECISION SUPPORT

This PhD thesis contributes to the newly emerged, growing body of scientific work on the use of News Analytics in Finance. Regarded as the next significant development in Automated Trading, News Analytics extends trading algorithms to incorporate information extracted from textual messages, by translating it into actionable, valuable knowledge.

The thesis addresses one main theme: the incorporation of news into trading algorithms. This relates to three main tasks: i) the extraction of the information contained in news, ii) the representation of the information contained in news, and iii) the aggregation of this information into actionable knowledge. We validate our approach by designing and implementing three semantic systems: a system for the computational content analysis of European Central Bank statements, a system for incorporating news in stock trading strategies, and a time-aware system for trading based on analyst recommendations.

The approach we choose for addressing these tasks is an interdisciplinary one. For the extraction of information from news we rely on approaches borrowed from Computer Science and Linguistics. The representation of the information contained in news is realized by using, and extending, the state-of-the-art in Semantic Web technology. We do this by bringing together insights from Logics, Metaphysics, and Computational Semantics. The aggregation of information is done by using techniques and results from Computational Intelligence and Finance.

ERiM

The Erasmus Research Institute of Management (ERIM) is the Research School (Onderzoekschool) in the field of management of the Erasmus University Rotterdam. The founding participants of ERIM are the Rotterdam School of Management (RSM), and the Erasmus School of Economics (ESE). ERIM was founded in 1999 and is officially accredited by the Royal Netherlands Academy of Arts and Sciences (KNAW). The research undertaken by ERIM is focused on the management of the firm in its environment, its intra- and interfirm relations, and its business processes in their interdependent connections.

The objective of ERIM is to carry out first rate research in management, and to offer an advanced doctoral programme in Research in Management. Within ERIM, over three hundred senior researchers and PhD candidates are active in the different research programmes. From a variety of academic backgrounds and expertises, the ERIM community is united in striving for excellence and working at the forefront of creating new business knowledge.

Design & layout: B&T Ontwerp en advies (www.b-en-t.nl) Print: Haveka (www.haveka.nl)

ERIM PhD Series
Research in Management

Erasmus Research Institute of Management - ERiM
Rotterdam School of Management (RSM)
Erasmus School of Economics (ESE)
Erasmus University Rotterdam (EUR)
P.O. Box 1738, 3000 DR Rotterdam,
The Netherlands

Tel. +31 10 408 11 82
Fax +31 10 408 96 40
E-mail info@erim.eur.nl
Internet www.erim.eur.nl