Isaías and Nunes (Eds.)
4th International Conference on the WWW / Internet
Lisbon, Portugal, 19-22 October 2005
pp. 395-402, ISBN 972-8924-02-X
(extended version of conference paper)

AUTOMATING MARKET FORECAST SUMMARIZATION FROM INTERNET DATA

Benno Stein and Sven Meyer zu Eissen
Faculty of Media, Media Systems
Bauhaus University Weimar, Germany
{benno.stein, sven.meyer-zu-eissen}@medien.uni-weimar.de

Gernot Graefe
C-LAB Paderborn, Germany
gernot.graefe@c-lab.de

Frank Wissbrock
Computer Science Institute
Paderborn University, Germany
frankw@upb.de

ABSTRACT

The World Wide Web has been discovered by market researchers. Useful information about new markets, market volumes, or the development of existing markets can be found and compiled from Web documents. This paper focuses on the task of market forecast summarization and how it can be automated by the combination of Web-based information retrieval and information extraction techniques. Market forecast summarization is essential for investment strategists who base their investment decisions on predicted market developments. Against this background, we break down the complex analysis of market developments into three specific questions concerned with turnover estimation.

The contributions of the paper relate to technology for automatically collecting and analyzing documents to answer these questions: focused search, document type assessment and filtering, time extraction, shallow parsing, and plausibility analysis. We have developed a prototype that operationalizes the technology and conducted a case study from the field of new technology assessment in order to demonstrate the ideas.

KEYWORDS

information extraction, information summarization, market forecasting

1. INTRODUCTION

"Worldwide revenues from radio frequency identification (RFID) tags will jump from \$300 million in 2004 to \$2,8 billion in 2009." [1]

How is the future potential of a market estimated? Market research offers a bundle of methods to answer this question [Berekoven et al. 2001]. One of the most popular approaches in recent years is Web-based literature research. In this context the World Wide Web is a rich source of market forecast statements, like the one at the beginning of this text. An analyst who collects and interprets these statements may obtain a reasonable idea about the future market volume.

However, manually conducted literature research based on documents found in the World Wide Web is time-consuming and usually not exhaustive, since human abilities to manage the information flood on the

[1] In-Stat market research, `http://www.rfidjournal.com/article/articleview/1320/1/1`.

World Wide Web are limited. These facts motivated our research on automatic market forecast summarization. In particular our attention focuses on market forecasts that contain sufficient information to answer the following questions:

(a) What is the predicted turnover of a defined market?

(b) For which time or time period does the forecast apply? (called predicted time here)

(c) At what time was the forecast made? (called statement time here)

We are developing a tool suite that uses state-of-the-art information retrieval and information extraction technology to automatically find answers to these questions and to generate charts of the type shown in Figure 1.

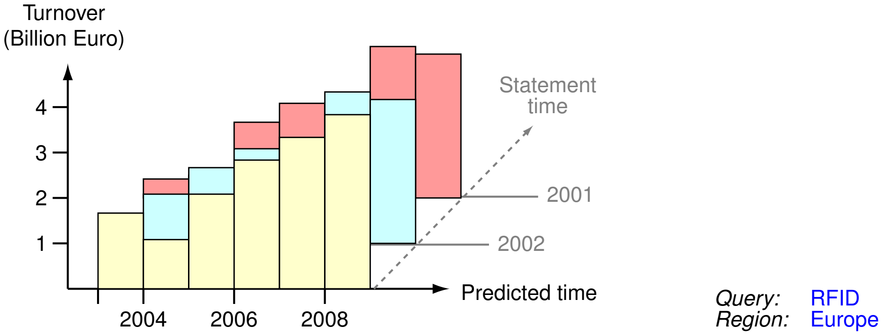


Figure 1. Market development of RFID tags in Europe. The extracted information is visualized as a bar chart along three dimensions: predicted turnover, predicted time, and statement time. The color of a bar indicates the information quality of the underlying information source.

Information retrieval draws on many other disciplines and is generally understood as the art of searching for information in all kinds of document collections. While there is no clear distinction between the term "information retrieval" and the related terms "data retrieval", "document retrieval", or "text retrieval", there is common understanding that information *extraction* is much closer to computaional linguistics [Engels and Bremdal 2000]: Information extraction typically leads to the introduction of a semantic interpretation of meaning based on the narrative under consideration and also includes stages of natural language parsing.

The remainder of the paper is organized as follows. Section 2 outlines existing approaches for market forecast summarization with respect to our objectives. Section 3 introduces a four stage approach to market forecast summarization. Section 4 presents a prototypic analyst tool which uses Web documents to identify statements that correspond to a user defined market and extracts the answers related to turnover, predicted time, as well as statement time.

2. CLASSICAL APPROACHES TO MARKET ANALYSIS

Emerging technologies frequently change existing markets or even create new market segments. Market forecasting seeks to anticipate the future development of such technologies at an early stage. Market forecasting is vital for most companies in order to develop reasonable business strategies and to make appropriate corporate investments. It is important to distinguish between the creation of "original" forecasts (next subsection) and "recycling", i. e., the interpretation and summarization of existing market reports (Subsection 2.2). The focus of our research is on the latter.

2.1 Market Forecast Creation

Market research companies track the developments of new technologies and identify the industries that are influenced by those technologies to predict future market volumes. They define the relevant market and collect information about the present situation, including competing technologies, major suppliers and customers, and other influencing factors. Hereon statistical instruments like neural networks [Tchaban et al. 1998; Thiesing and Vornberger 1997] and Bayesian models [Neelamegham and Chintagunta 1999] are employed to analyze the gathered data and forecast future market volumes.

2.2 Market Forecast Summarization

For new technologies such as RFID, Web Services, or Augmented Reality, we observe large differences between the forecasts of market volumes that are published. This is not surprising, considering that different forecasting approaches are used for prediction. Moreover, the point of time when forecasts are made affects the prediction of market volumes: New forecasts consider recent information, whereas forecasts made earlier rely on different sources.

As a consequence, companies try taking all available market forecasts into consideration prior to making important decisions. It is general practice to search for and consolidate information found on the World Wide Web in order to collect a broad information basis.

So far, this process is carried out manually. The technology that is described in the following sections aims to support this process of collecting, assessing, and consolidating information from the World Wide Web to form a comprehensive presentation of the expected market volume.

3. AUTOMATING MARKET FORECAST SUMMARIZATION

This section introduces a four stage approach to automate the problem of market forecast summarization, which consists of the following steps: collecting candidate documents, report filtering, time and money identification, and a phrase analysis along with template filling (see Figure 2). The next subsections discuss these steps in greater detail.

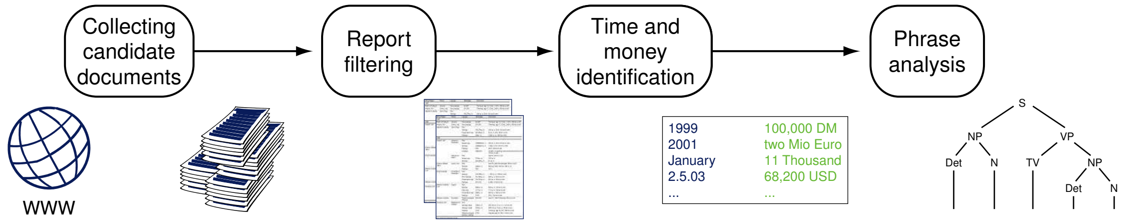


Figure 2. Illustration of the four stage approach to market forecast summarization.

3.1 Collecting Candidate Documents on the Web

Web search engines provide an obvious way to retrieve documents about a certain topic, and we use them in a meta-search-manner to compile a first candidate document collection. However, the vocabulary of potentially useful documents is unknown in advance, and, given a search task, there is the question of the "optimum" query. This problem relates to query expansion, local context analysis, and co-occurrence analysis [Voorhees 1994; Xu and Croft 1996, 2000]. We combine several techniques from these fields within a knowledge-based query construction process.

Starting point is a set T_{M_1} of keywords that characterize the market of interest.[2] By adding synonyms, coordinate terms, and derivationally related forms, the set T_{M_1} is extended towards a set T_{M_2} [Fellbaum 1998]. Within T_{M_2} groups of words are identified by exploiting statistical knowledge about significant left and right neighbors, as well as adequate co-occurring words, yielding the set T_{M_3} [University of Leipzig 1995]. Then, a sequence of queries is generated (and passed to search engines) by selecting elements from the Cartesian

[2] $T_{M_1} = \{\text{RFID}\}$ in the example at the outset.

Table 1. Important features for report filtering.

Feature type	Feature instance
Concentration measures	maximum density of turnover symbols avg. word frequency class
Closed-class word sets	avg. # of turnover symbols avg. # of currency symbols avg. # of shop symbols avg. # of date symbols avg. # of words not in Webster's
Text statistics	avg. # of numerals avg. # of alphanumeric words avg. # of digits

product $T_{M_3} \times T_F$, where T_F is a hand-crafted vocabulary with typical market analysis terms.[3] This selection step is controlled by *quantitative* relevance feedback: Depending on the number of found documents more or less "esoteric" queries are generated. Note that such a control can be realized by a heuristic ordering of the sets T_{M_3} and T_F, which considers word group sizes and word frequency classes [Voorhees 1994]. The result of this step is a candidate document collection $\mathcal{C}$.

3.2 Report Filtering

Report filtering is the process of dividing a candidate document collection $\mathcal{C}$ into two sets, $\mathcal{C}^+$ and $\mathcal{C}^-$. The former set shall contain those documents—called reports here—that are likely to contribute valuable information for market analysis, the latter shall not. Our hypothesis is that report filtering relates to *Web genre identification*, which differentiates between documents with respect to their form, style, or targeted audience.[4]

Each Web genre comes with its own characteristics. Previous work has shown that these characteristics can be captured by features that quantify presentation aspects, specific word sets, or text statistics. The challenge here is the identification of features that provide the discriminative power to divide $\mathcal{C}$ into $\mathcal{C}^+$ and $\mathcal{C}^-$. Note that this kind of genre classification automates one of the most time-consuming tasks in manually conducted forecast summarization.

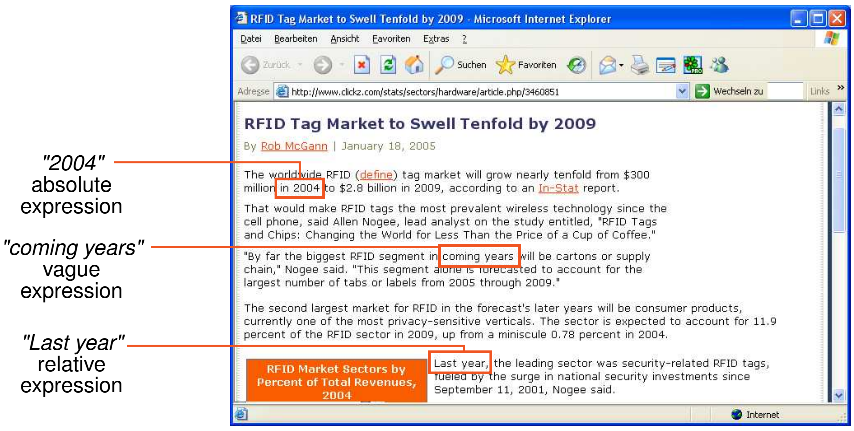
Aside from applying existing genre identification technology to report filtering, we introduce a new class of concentration measures. These measures are based on the observation that in relevant documents the use of certain closed-class word sets varies considerably. Such a variation can be quantified by computing the maximum "concentration", γ_R^*, of a word set, R, with respect to a small text window in the document. Given the sequence $t_1, \ldots, t_n$ of terms in a document d, a text window $W_i = \{t_{i_1}, \ldots, t_{i_m}\}$ of length m is defined as a subset of m consecutive terms in d. For a document d, a window size m and a word set R, γ_R^* is defined as follows:

$$\gamma_R^* = \max_{W_i \subset d}(\gamma_R(W_i)), \quad \text{with } \gamma_R(W_i) = \frac{|W_i \cap R|}{m}$$

In our experiments, the set R is compiled from the closed-class word sets for time-, currency-, and business-related terms. Table 1 gives an overview of the most important features that we employ for report filtering.

[3] T_F contains words like "forecast", "market", "figures", "estimation", etc.

[4] Examples for Web genres are online shops, help pages, discussion forums, technical articles, or private portrayals [Meyer zu Eißen and Stein 2004]. Observe in this respect that Web search engines provide a topic-centered service: It is hardly ever possible to instruct them to deliver documents of a specific genre with a high precision.



http://www.clickz.com/stats/sectors/hardware/article.php/3460851

RFID Tag Market to Swell Tenfold by 2009

By Rob McGann | January 18, 2005

The worldwide RFID (define) tag market will grow nearly tenfold from $300 million in 2004 to $2.8 billion in 2009, according to an In-Stat report.

That would make RFID tags the most prevalent wireless technology since the cell phone, said Allen Nogee, lead analyst on the study entitled, "RFID Tags and Chips: Changing the World for Less Than the Price of a Cup of Coffee."

"By far the biggest RFID segment in coming years will be cartons or supply chain," Nogee said. "This segment alone is forecasted to account for the largest number of tabs or labels from 2005 through 2009."

The second largest market for RFID in the forecast's later years will be consumer products, currently one of the most privacy-sensitive verticals. The sector is expected to account for 11.9 percent of the RFID sector in 2009, up from a miniscule 0.78 percent in 2004.

RFID Market Sectors by Percent of Total Revenues, 2004

Last year, the leading sector was security-related RFID tags, fueled by the surge in national security investments since September 11, 2001, Nogee said.

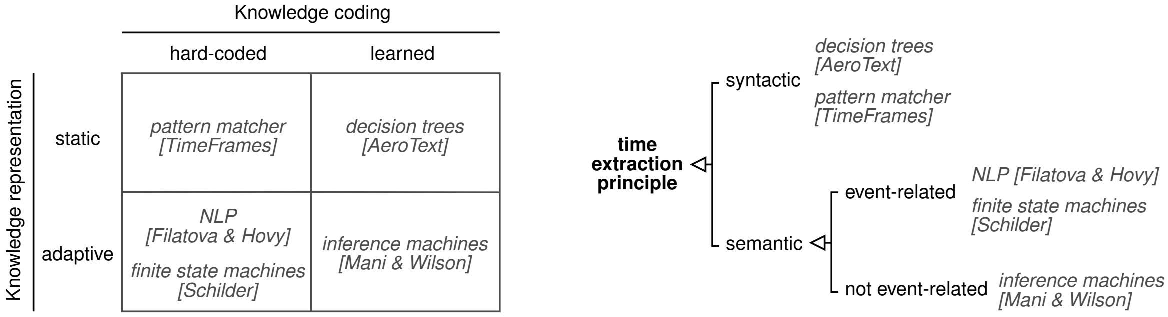
Figure 3. A website with market forecasts. The different types of time expressions are highlighted.

3.3 Time and Money Identification

This step is carried out on the set of candidate reports, $\mathcal{C}^+$, and pertains to the identification and unified representation of time and money information in order to simplify the subsequent phrase analysis.

Time extraction is a challenging task since time information appears in different forms—take for example absolute time information like dates, relative time information like "next year", or vague time information like "within the past years". Even absolute dates can take various forms, depending on the geographic location. An automated time extraction agent must identify all forms of time information and, as the last two examples show, track a relative or vague time information back to an absolute time or time span. Figure 3 shows an example.

Time extraction from text is a relatively new discipline and has been gaining increasing attention since the end of the 1990s. To this day only few approaches have been proposed, which are surveyed and classified in the left matrix of Figure 4 [Mani and Wilson 2000; Filatova and Hovy 2001; Koen and Bender 2000; Schilder and Habel 2001]. The approaches divide into static and adaptive approaches. The former use a list of fixed patterns that contain indicator words like "two days" to identify a starting point for time identification; specialized rules refine the kind of the found time information, e. g. they decide whether the time is absolute ("in two days") or designates a period ("for two days"). The latter approaches use more comprehensive context information like regular expressions or grammars to determine the kind of the time information. Aside from the knowledge representation, the approaches differ in the methods to fill their pattern bases. This information can either be coded manually, or it can be learned using a supervised learning procedure that takes documents with labeled time information as input.



Knowledge representation		Knowledge coding: hard-coded	Knowledge coding: learned
	static	pattern matcher [TimeFrames]	decision trees [AeroText]
	adaptive	NLP [Filatova & Hovy] finite state machines [Schilder]	inference machines [Mani & Wilson]

time extraction principle
syntactic
decision trees [AeroText]
pattern matcher [TimeFrames]
semantic
event-related
NLP [Filatova & Hovy]
finite state machines [Schilder]
not event-related
inference machines [Mani & Wilson]

Figure 4. Classification scheme for time extraction approaches. The approaches are classified along two dimensions: knowledge representation and knowledge coding; the text in the brackets contains the name of the author or product.

Table 2. Patterns used in the analysis of year expressions in English texts. The phrase "creation date" refers to the creation date of the respective document; the phrase "previously encountered year" refers to the last year appearing before the current pattern.

Pattern	Resolved to
four-digit number	unchanged
last year	creation date - 1
next year	creation date + 1
this year	creation date
in [n] years	creation date + n
[n] years later	previously encountered year + n
[n] years before, [n] years ago	previously encountered year - n

The tree on the right-hand side of Figure 4 organizes the same approaches by their linguistic scale. Syntactic approaches resolve the corresponding time value of a time expression without considering its context or its position in a text. Semantic approaches resolve time expression values from the text's context, if necessary. At the second level they further divide into event-related approaches, which tie time expressions to corresponding event expressions in the text, and event-unrelated approaches.

In the context of market forecast summarization, year expressions are of special interest, because a market forecast refers to the potential turnover on a specific market during a year. Our year extraction approach is based on a pattern collection $P_Y = \{(p, f(p), c)\}$, where p is a predefined year pattern, $f(p)$ is a function that resolves the value of pattern p, and c is a manually assigned confidence value. Table 2 summarizes important patterns from P_Y and indicates the resolved time values. After all year expressions are identified, a set of rules is applied to determine the creation year of the document. Our strategy include several heuristics, among others: (1) Search of single lines that contain a year expression that co-occurs with one of the terms "copyright", "published", etc. (2) Search of short paragraphs that contain a year and at least another time expression and that are located near a headline or the end of the text. (3) Search of combinations of a location along with a time expression, such as "New York, June 2004".

Table 3. Examples for money expressions and their unified representation.

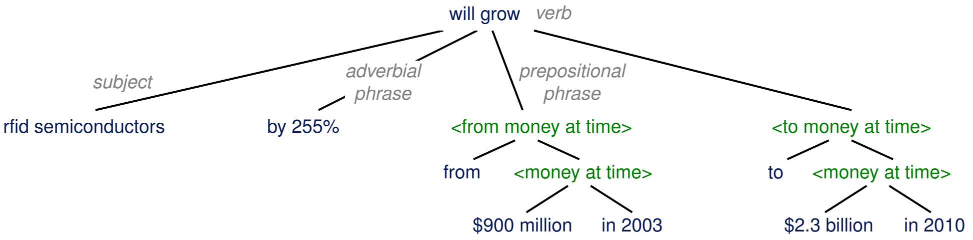
Original expression	Characteristics	Resolved to
$2.2million	unigram, prefix currency, mixed number-numeral	(2200000.0 USD)
100 million US dollars	mixed number-numeral bigram, postfix currency bigram	(100000000.0 USD)
EUR 2.2million	bigram, prefix currency, mixed number-numeral	(2200000.0 EUR)

Information about money can be represented in very different forms, Table 3 gives a few examples. Within our application the strategy for money value extraction and preparation is based on regular expression matching coupled with heuristic preference rules; it addresses the following issues: (1) identification of money information written as unigram, bigram, trigram, and 4-gram, (2) recoding of numerals and combined number-numeral representations, (3) unified representation as bigrams, consisting of a number a standardized currency symbol.

3.4 Phrase Analysis

The last analysis step, the identification of significant associations between corresponding time and money expressions is carried out with natural language parsing technology. Such a constituency analysis happens at the sentence level, and we need (1) a grammar that defines a theory on sentences that are allowed, and (2) a parser to match a particular sentence against such a grammar. Though this is a very hard problem in general, it is tractable here since we are in a narrow domain, and we may consider the phrases as a rather small number of possible templates that must be filled. For both the definition and the filling of such templates the so-called dependency grammars have proven a useful tool [Melcúk 1988; Schacht et al. 1994; Debusmann et al. 2004].

Various grammar formalisms have been developed, where the more well known include context free grammars, tree adjunction grammars, generalized phrase structure grammars, and dependency grammars. According to [Staab 1999; Engels and Bremdal 2000] dependency grammars have several advantages that make them preferable for information extraction tasks: syntactic simplicity, semantic correspondence, or discontinuity handling. Dependency grammars provide a means to identify and fill a terminological system consisting of concepts and roles. Technically speaking, they put the dependency between the words of a sentence in the center of their analysis. Typically the verb is of primary importance while the other words are subordinated, leading to a hierarchical structure. An example of the application of Step 3, time and money identification, and Step 4, phrase analysis, is shown in Figure 5.



"In a recent report future horizons predicts that rfid semiconductors will grow by 255 per cent from $900 million in 2003 to $2.3 billion in 2010."

```
((:START
  (:MONEY-RAISE (:FROM FROM)
    (:MONEY-AT-TIME (:MONEY 0.9E9 USD) (:IN IN) (:YEAR 2003))
    (:TO TO)
    (:MONEY-AT-TIME (:MONEY 2.3E9 USD) (:IN IN) (:YEAR 2010)))))
```

Figure 5. A dependency tree generated according to the money-raise-rule of the dependency grammar. Input to the parser is the left sentence, while the parser output is shown right.

4. CASE STUDY

Automatic market forecast summarization is a new discipline and no benchmark corpus is currently available. So we decided to build our own collections: We downloaded Web documents for two sample markets, "RFID" and "Outsourcing", and let human experts label the documents depending on their turnover forecast information. Table 4 gives an overview of the compiled corpora.

Table 4. Composition of the compiled corpora.

Market	# Relevant	# Irrelevant	Total
RFID	83	217	300
Outsourcing	107	216	323
Total	190	433	623

The computation of the features for the report filtering step is straightforward and can be done in linear time, during the parsing of the documents—assuming a constant time for dictionary look-ups. With respect to our new concentration measure, γ_R^*, a window size of $m = 40$ and term-wise sliding (for each document) was chosen; to quantify the feature set's classification performance a discriminant analysis was applied. Table 5 (left-hand side) shows the resulting confusion matrix for a 10-fold cross-validated classification of 600 documents with $|\mathcal{C}^+| = 180$ and $|\mathcal{C}^-| = 420$; Table 5 (right-hand side) ranks the features according to their discriminative power, based on Wilks' Lambda.

Table 5. The left table shows the confusion matrix for Step 2, report filtering. The values result from a 10-fold cross-validated classification of a sample with 600 documents drawn from both the RFID-corpus and the Outsourcing-Corpus. The table on the right-hand side shows the features according to their discriminative power.

	Relevant ($\mathcal{C}^+$)	Irrelevant ($\mathcal{C}^-$)	Σ
relevant ($\mathcal{C}^+$)	73,6%	26,4%	100%
irrelevant ($\mathcal{C}^-$)	25,6%	74,4%	100%

Rank	Feature
1	γ_R^*
2	avg. # of currency symbols
3	avg. # of turnover symbols
4	avg. # of numerals
5	avg. # of shop symbols
6	avg. # of alphanumeric terms
7	avg. # of digits
8	avg. # of date symbols
9	avg. # of words not in Webster's
10	avg. word frequency class

Figure 6 shows an automatically generated forecast summarization: Each bar in the chart represents three data dimensions, which are the statement time, the predicted time, and a turnover value. The triples are extracted from the set of relevant documents, $\mathcal{C}^+$, by means of Step 3 and Step 4, time and money identification and phrase analysis, whereas $\mathcal{C}^+$ is the result of the report filtering step.

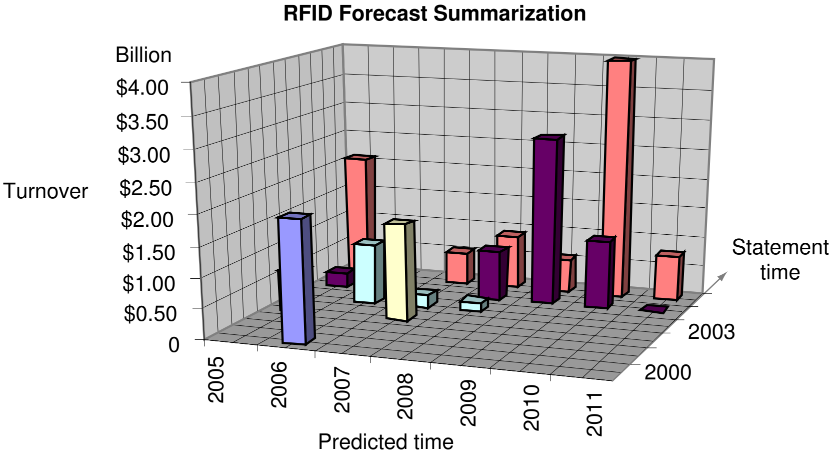


Figure 6. Generated bar chart of a forecast summarization for the RFID-market.

5. SUMMARY AND OUTLOOK

This paper introduces a new application in the field of high-level information retrieval tasks: The automatic creation of market forecasts from meta-searched Web documents. We presented an approach to this challenging problem as well as a prototypic implementation that is comprised of four consecutive steps: collecting candidate documents, report filtering, time and money identification, and phrase analysis. Among these steps, report filtering is presumably the most time consuming task for human experts, and we extended ideas from the field of document genre analysis to tackle this problem. In particular we introduced a new class of concentration measures to discriminate relevant reports from topically similar but irrelevant documents. The data extraction step is accomplished by a rule-based identification for time expressions that is able to differentiate between predicted time and statement time, the identification and unified representation of money expressions, and a phrase analysis that is based on the concept of dependency grammars.

To evaluate our ideas we compiled for two interesting markets two sample corpora with Web documents and let human experts label the documents with respect to the contained forecasting information. Using these corpora we set up experiments to test our technology. The results look quite promising: for the report filtering step as well as for the data extraction tasks high precision values were achieved. We see potential to further improve our approach, and current work concentrates on the refinement and an extensive evaluation of the presented technology. Future work shall focus on the automatic summarization of sub-markets for a given market, which may be characterized by both geographical and structural segmentation.

REFERENCES

Ludwig Berekoven, Werner Eckert, and Peter Ellenrieder. *Marktforschung: Methodische Grundlagen und praktische Anwendung*. Gabler, 9 edition, 2001.

Ralph Debusmann, Denys Duchier, and Geert-Jan M. Kruijff. Extensible Dependency Grammar: A New Methodology. In *Proceedings of the COLING 2004 Workshop on Recent Advances in Dependency Grammar*, Geneva, Suisse, 2004.

Robert Engels and Bernt Bremdal. Information Extraction: State-of-the-Art Report. Technical report, CognIT a.s, Asker, Norway, 2000.

Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

E. Filatova and E. Hovy. Assigning time-stamps to event clauses. In *Proceedings of the ACL 2001 Workshop on Temporal and Spatial Information Processing*, 2001.

D. B. Koen and W. Bender. Time Frames: Temporal augmentation of the news. *IBM Systems Journal*, 39 (3+4):597–616, 2000.

I. Mani and G. Wilson. Robust Temporal Processing of News. In *Proceedings of the Association for Computational Linguistics (ACL-2000)*, pages 69–76, 2000.

I. Melcúk. *Dependency Syntax: Theory and Practice*. State University Press of New York, Albany, New York, 1988.

Sven Meyer zu Eißen and Benno Stein. Genre Classification of Web Pages: User Study and Feasibility Analysis. In Susanne Biundo, Thom Frühwirth, and Günther Palm, editors, *KI 2004: Advances in Artificial Intelligence*, volume 3228 LNAI of *Lecture Notes in Artificial Intelligence*, pages 256–269, Berlin Heidelberg New York, September 2004. Springer. ISBN 0302-9743.

Ramya Neelamegham and Pradeep Chintagunta. A Bayesian Model to Forecast Product Performance in Domestic and International Markets. *Marketing Science*, 18(2):115–136, 1999.

S. Schacht, U. Hahn, and N. Broker. Concurrent lexicalized dependency parsing: A behavioral view on ParseTalk events. In *Proceedings of the 15th International Conference on Computational Linguistics, Kyoto, Japan*, pages 489–493, 1994.

F. Schilder and C. Habel. From Temporal Expressions to Temporal Information: Semantic Tagging of News Messages. In *Proceedings of the ACL 2001 Workshop on Temporal and Spatial Information Processing*, pages 65–72, 2001.

S. Staab. Grading Knowledge: Extracting Degree Information from Texts. In *Lecture Notes in Computer Science (1744)*, Berlin, 1999. Springer.

T. Tchaban, J. P. Griffin, and M. J. Taylor. A comparison between single and combined backpropagation neural networks in the prediction of turnover. *Engineering Applications of Artificial Intelligence*, 11(1): 41–47, February 1998.

Frank M. Thiesing and Oliver Vornberger. Sales Forecasting Using Neural Networks. *International Conference on Neural Networks Proceedings*, pages 2125–2128, 1997.

University of Leipzig. Wortschatz. `http://wortschatz.uni-leipzig.de`, 1995.

Ellen M. Voorhees. Query expansion using lexical-semantic relations. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61–69, New York, NY, USA, 1994. Springer-Verlag New York, Inc. ISBN 0-387-19889-X.

Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11, New York, NY, USA, 1996. ACM Press. ISBN 0-89791-792-8. doi: http://doi.acm.org/10.1145/243199.243202.

Jinxi Xu and W. Bruce Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112, 2000. ISSN 1046-8188. doi: http://doi.acm.org/10.1145/333135.333138.