The Face of Text
Computer Assisted Text Analysis in the Humanities

CaSTA (Canadian Symposium on Text Analysis) conference
at McMaster University
November 19th to 21st 2004

Automatic Trend Detection and Visualization using the Trend Mining Framework (TMF)

Author:
Dipl.-Kfm. Tobias Kalledat,
School of Business and Economics of the Humboldt University in Berlin,
Spandauer Str. 1, D-10178 Berlin, Germany

Postal address:
Tobias Kalledat, Eddastr. 94, D-13127 Berlin, Germany, Email: Tobias@Kalledat.de

1. Introduction

The documentation of historical Domain Knowledge about fields of activity usually takes place in the form of unstructured text -, picture -, audio- and video documents that are produced over longer time periods. Structured (e.g. relational data) and semi structured (e.g. HTML pages) and unstructured documents (e.g. texts) become distinguished with regard to the degree of an internal structure.

The usage of structured languages such as XML for tagging of textual data stands only at the beginning of its development path and is for historical documents therefore not to be found. Thus in the relevant literature is assumed that up to 80-90% of the electronically stored Knowledge is hidden in such unstructured sources [Tan99, Dörr00]. There are three implications most relevant:

a) The production of such documents rises over the time due to the distribution of Information Systems and their shared use, e.g. over the Internet. The availability of large sources of potentially interesting Knowledge becomes ubiquitous.

b) The usage of actual Knowledge becomes more and more a critical factor for competition of market participants. To adjust their product portfolio quickly it is necessary to adapt new ideas in a short development period, because consumers asking for product live cycles getting shorter and services getting closer to their individual needs. For members of the underlying processes studying lifelong gets more and more important and the flexibility of adapting new Domain Knowledge in a short time turns to one of the most important tasks.

c) Today the use of implicit Knowledge that is hidden in the huge amount of unstructured data is an approach that can be used because of the rapid development of powerful hardware that can handle such large data sources and the methods that were developed under the terms “Knowledge Discovery in Databases” (KDD) and "Data Mining" since the early 1990ies [Fayy96]. The Data Mining process allows discovering formerly unknown, potentially useful Knowledge Patterns out of various input data using techniques and methods of statistics, linguistics and data base research [Düsi99], [Biss99].

These are challenges for the educational sector as well as for the designers of Information Systems that support these processes of KDD. Therefore there is a market demand for turning the Knowledge Discovery Process (KDP) itself from an individual approach of a small number of specialists to a process that supports a large amount of “Domain Knowledge Researcher” in firms and organizations. Aim of the current research of the author is to find a methodology, which is able to support the KDP on large Technical Domain Corpora.

The Data Mining Process should be performed continuously. From the efficiency point of view it must be made sure, that Knowledge that was discovered in past time periods is be used as a basis for an adjustment of current decisions, which have relevance for the future.

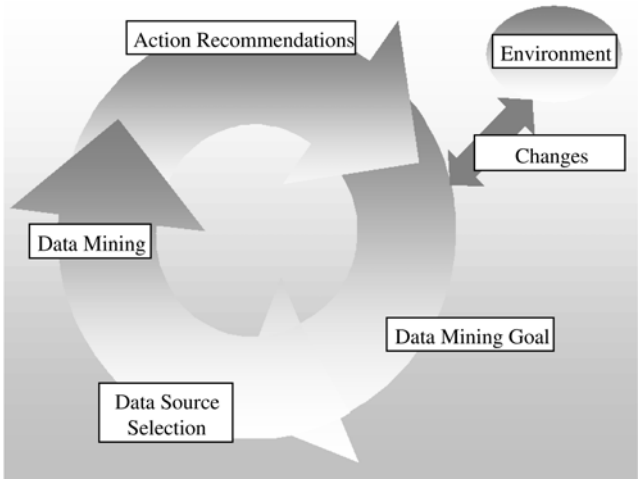

Fig. 1: Data Mining Management Cycle, translated from [Kall04]

Therefore a Data Mining Management Cycle was being proposed in [Kall04], see Fig. 1. While analyzing large amounts of textual data, for the Domain Knowledge Researcher two problems result: On the one hand the pure amount of unstructured historical and present data represents a substantial entrance barrier. On the other hand unstructured data themselves cannot be automatically processed easily. A substantial realization gain is to be expected, if methods are found to open and evaluating the mentioned unstructured sources of information.

The advantage of an evaluation of historical text documents in the opposite to interviewing time witnesses exists in the reflection of the historical reality free of subjective information distortions. On this assumption, the provision of information problem reduces to the procurement of suitable text documents. Expenditures for the determination of time witnesses and for the execution and evaluation of interviews can be minimized thereby. The analysis of electronic sources gives also the opportunity to transform semantic structures of Knowledge Domains in an Information Technology (IT)-based representation that allows documenting and sharing Knowledge more easily.

To support individuals in the KDP is economically interesting. Potentially expensive manual work can be substituted by automatically working Information Technology driven solutions. Approaches in this support are basing mostly on methods that working on each text file itself for clustering, tagging or classifying purposes [Lend98, p.130], [Moen00]. The objective mostly is to support information retrieval or later querying against a mass of these text files that were proceeded the same way. Most of the used procedures do not consider the time dimension.

Technical Domains, e.g. Information Technology, do have special qualities, which make it necessary to configure the methods of research to the needs of the research aim. Such corpora must be handled different to "usual" corpora. For example, product names, programming languages and other proper names must be kept during all analysis steps. Pure linguistic approaches therefore not applicable. Therefore instead of "Word" the term "Phrase" is used from now which covers a wider range of alphanumeric sign combinations, e.g. product names and technical norms.

For a Domain Knowledge Researcher it is important to know: How does the semantic of Phrases change over the time? Which topics are growing, falling and what is the semantically basis of the Domain? For distinguishing between these clusters, rules for significant decisions are needed. An important challenge for research is to define methods, which can extract significant pattern and track time dependent changes.

The main specifics of the proposed methodology are:

1) Proposing a top down approach for pre-filtering such patterns, that are worth for further analysis using appropriate Meta Data Measures of corpora
2) Proposing the use of time dependent Ontology for tracking Trends and covering semantic changes
3) Suggesting a visualization concept, which visualizes the Domain Knowledge, that is represented by a corpus
4) Combining the top down Meta Data Measure approach with the Ontology based semantic modelling using an OLAP-based intuitive navigation concept

The objective of this paper is to propose and evaluate appropriate methods for Automatic Detection, Classification and Visualization of Trends in large technical focused Domain Corpora. Chapter 2 introduces common Approaches for pattern recognition in textual data; chapter 3 is dealing with the proposed methodical approach of the "Trend Mining Framework". It's components and the proposed Trend Mining Process are introduced in chapter 4. The paper closes with conclusions and outlook in chapter 5.

2. Approaches for pattern recognition in textual data

Since the 1990-ies under the term Data Mining methods were developed, which make it possible to recognize unknown structures in data and derive from it action-relevant and economi-

cal useful Knowledge [Codd93]. These methods are based on classical statistic procedures as well as methods of adjacent research fields and were adapted for the employment on appropriate data.

Methods for the investigation of unstructured data, e.g. large text corpora or speeches, usually subsumed under the headline Computer Linguistics or Content Analysis. As a special research field Text Mining was developed for the computer-based analysis of unstructured textual data. In the Mid-1990-ies the research activities get pushed by carrying out the *"Topic Detection and Tracking"* (TDT) task by a few research organizations and the U.S. government. Subject of this research is event-based organization of broadcast news [Alla02a], [Alla02b]. The main research task was divided into the following sub-tasks: *Story Segmentation, First Story Detection, Cluster Detection, Story or Topic Tracking* and *Story Link Detection.* Most of the used methods are bottom up approaches that analyzing text corpora word-by-word or sentence-by-sentence and using clustering and tagging techniques [Spil02]. Other methods are more statistically based and working with Features, e.g. corpus wide measures or Vector Space Models which represent sentences or whole stories. Applications based on these methods are realized, e.g. for *Automatic News Summarizing, Document Clustering and Patent Mining.*

The classical linguistic text analysis has a long developmental history, which reaches back up to time periods of the middle of the last century. An important influence on linguistics Chomsky had, who characterized the research in the middle of the last century. He especially criticized the use of the Corpus Linguistics approach for learning more about the language itself, because all analysed corpora can only be a subset of the whole variety of the language. With this "generative transformational grammar" he revolutionized linguistics and became the most important theorist of his branch. Modern research approaches are using statistical methods for quantitative analyses of text corpora, e.g. for semantic similarity checks of document sets. The linguistic methods therefore can be differed into two main directions of research activity:

(1) The predominant quantitative approach, that uses Meta Data Measures of text corpora, e.g. term or word frequency for evaluation and comparison of different text sources [Atte71]. Some researchers are assuming, that all the rules are inside the used language, that it is worth analysing real life data to learn more about languages. Purposes for this Corpus Linguistics approach e.g. comparisons between authors and their styles or detecting, whether a text belongs to author A or author B.

(2) The more qualitative approach of descriptive grammars [Lend98, p. 106], that makes use of interpretation techniques and is working with thesauri or special grammatical algorithms for word analysing purposes (e.g. stemma finding). Goal is describing languages with rules and building a computable basis, e.g. for Machine Translation solutions.

It can be observed that most of the methods for text mining are bottom up approaches, coming from the smallest unit of textual data, a word or n-gram (only a few letters) and generalizing the pattern, which were found. For the tracking of Trends in Technical Domain Corpora over time these known bottom up procedures are having limitations:

i. There are performance issues when real large corpora is analysed.

ii. The patterns found are based on generalized results of multi parametric algorithms, which means that there is to be expected a biased result due to the multiplication of error terms.

iii. Linguistic approaches are not appropriate, because technical information is not covered or is destroyed during stemma finding processes.

iv. The generation of action recommendations is not transparent to “normal” users.

To support the process of Automatic Detection, Classification and Visualization of Trends in large Technical Domain Corpora instead of using usual bottom up procedures a promising approach is to combine classical Meta Data Measure oriented analysis in the first step with a projection step of found patterns into the detail layer of Phrases, in order to overcome this lack of research. Analysing appropriate Meta Data in the first step should reduce the complexity for pattern recognition dramatically. After this, the relations between interesting Phrases can be analysed more focused. Appropriate methods for modelling hierarchical structures of elements, e.g. Phrases, are known under the term Ontology. These are semantic concepts modelling Knowledge Domains by the use of directed graphs, which can show relations between elements of Domain Corpora. Based on such concepts deeper analyses, e.g. the use of classical Data Mining techniques is possible in later steps. The proposed methodology is introduced in the following chapters.

3. Trend Mining Framework

The TMF is a proposed methodology and also a process of Text Mining, which makes it possible to extract time-related Domain Knowledge based on unstructured textual data semi automatically. To establish a framework that allows exploring and analysing large Technical Domain Corpora in an intuitively interactive way is the main goal of developing the TMF approach. For this, methods were evaluated, which allow measuring the quantitative characteristics of a time tracked Domain Corpus.

A basis assumption is, that the frequency of Phrases is positive correlated with their importance within a corpus (no articles et cetera, but nouns and names). Based upon this assumption possible Meta Data Measure candidates can be defined, e.g.:

I. In [Crou90] the “Term Frequency Inverse Document Frequency” (TF-IDF) is used, which is defined as the number of times a Phrase appears in a document multiplied by a monotone function of the inverse number of documents in which the Phrase appears. In difference to the original source here the former defined term “Phrase” was used instead of the term “word”.

II. “Type-Token Ratio” (TTR) is defined as the ratio of different Phrases to their number of occurrence in a corpus [Lend98].

III. The “Phrase Repetition Quota“ (PRQ) or Frequency is the ratio between all different Phrases to all Phrases occurring in a corpus.

Initially, the PRQ was used as Meta Data Measure in a prototypic realization of the Trend Mining Process on a large corpus. The TF-IDF, TTR and monotone transformations of them as well as further candidates seem to be worth for testing in later realizations. It has to be taken under consideration, that only a few (short) words or Phrases are occurring often in a corpus. More important words or Phrases are longer, but rare [Zipf49]. A usual method is to filter very frequent, but not meaningful “stop words” to focus the analysis on semantically more important Phrases. In the current investigation a short stop-word list of some very frequent terms was used for filtering.

As an example, the historical development of the Domain “Information Technology” was analysed using the TMF. By the use of the TMF methodology it was possible to identify temporal semantic developments and to differentiate these regarding its persistence characteristics in short living Hypes and long-term Trends and to describe their development paths while keeping the technical Peculiarities of the sources. Extracting corpus Meta Data and transforming it into an appropriate detail level using a Domain specific Ontology do this. The TMF consists of

1. **Format Parsers**: for separation of content and formatting as well as converters to ASCII
2. **Analytical Parsers**: for determination of the Meta Data of the corpus and Feature Extraction, e.g. Phrase frequency
3. **Data Warehouse**: for the storage, e.g. of intermediate results and for a semi-automatic assignment of Phrases to an Ontology out of the Domain Corpus
4. **Statistical Tools**: for analyses and tests
5. **Graphical component**: for visualization and navigation

In the next chapter the proposed Trend Mining Process using the TMF is described in detail.

4. Trend Mining Process

In the TMF process the main task of Automatic Detection, Classification and Visualization of Trends in large Technical Domain Corpora is divided into a few sub tasks. The tasks are performed by the different components, mentioned earlier.

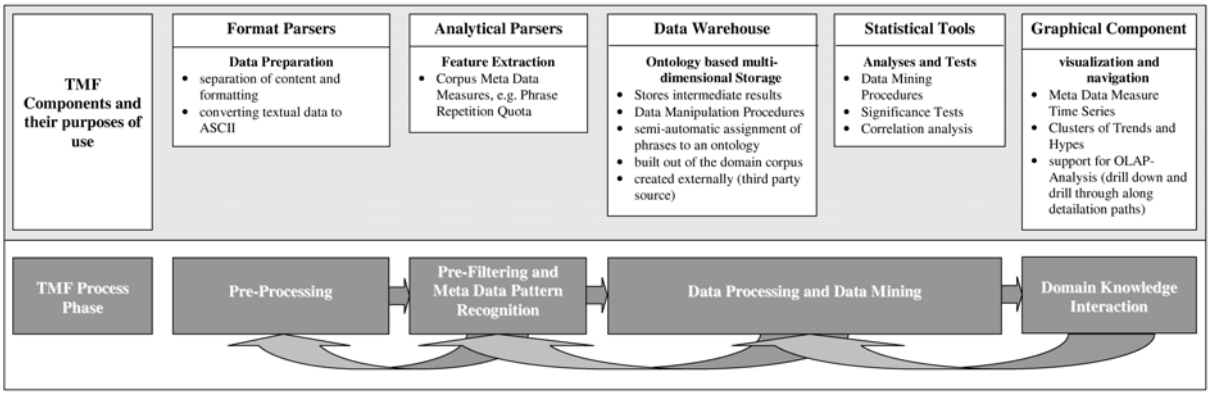

Fig. 2: Components of the TMF and their use in the Trend Mining Process

In Fig. 2 the Components of the TMF and their use in the Trend Mining Process are shown. The Trend Mining Process consists of the phases: *Pre-Processing, Pre-Filtering and Meta Data Pattern Recognition, Data Processing and Data Mining as well as Domain Knowledge Interaction*. Prototypically the TMF was used to analyse a large technical orientated corpus of the German issue of the weekly magazine "Computerwoche" (engl.: "Computer Week") of the last 29 years, starting in 1975 [Kall03]. For this paper the articles of the year 2003 added to the corpus and the results were updated.

In the **Pre-Processing Phase** the 132.406 articles were converted to pure ASCII files by the use of Format Parsers. In the phase of Pre-Filtering and Meta Data Pattern Recognition, Corpus Meta Data Measures were extracted for all available articles of each year of publication. All articles of one year were analysed together at the same time. It was counted 29,777,462 Phrases for the whole corpus and 2,941,680 distinct Phrases, which means an average PRQ of 10.12. The PRQ was used for the Top Down Approach of Pattern Recognition. The Graph of the PRQ is shown in Fig. 3:

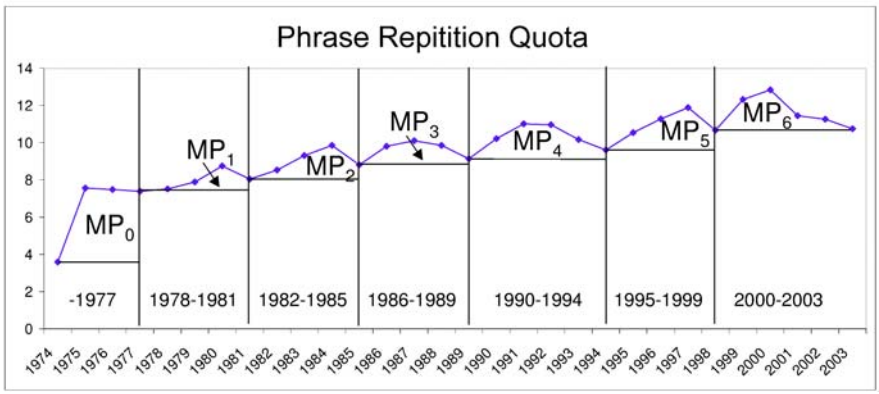

Fig. 3: Graph of Phrase Repetition Quota for each year of publication

The PRQ is rising on the average and periods are visible, which are divided by turn points of the PRQ (rising → falling → rising). The areas, marked as M_{P0}… M_{P6} are patterns that are worth to be analysed, because they are representing Phrases, which do have a Frequency that is higher than the average PRQ. The Frequencies where normalized that the specific corpus length of each year issues does not influence their comparability. After this **Phase of Meta Data Pattern Recognition** the articles of the corpus were time based segmented according to the periods, which were recognized. All further analyses are based on the segmented corpus. The phenomenon of an extensive PRQ being carried out via a restricted temporal period is understood by a "Hype" in the Information Technology in the following. Unlike such a thematic excess concepts, which learned, an extensive mentioning in several periods can exist. Such concepts are described below as Trends. Hypes are rather "nine days' wonders" after this interpretation, without considerable durable meaning, "Trends", on the other hand, characterize the Information Technology in the long run. Phrases, which have a Frequency higher than the lowest PRQ period limit, defining periods of "Trend" or "Hype" Phrases, which occurring more often in the corpus than other Phrases in the same period. Using this filter criterion, the number of Phrases was reduced by 16%. Thus, new time based corpora is analysed by considering results from former corpora also a basic learning capability is present.

From the viewpoint of the Set Theory, during a first step in the Phase of **Data Processing and Data Mining** it is possible to subsume the whole corpus as $MV_{CW1974\_2003} = M_{CW1974} \cup M_{CW1975} \cup ... M_{CW2003}$, which is equal with the Union Set of all articles. In the following steps, the previously built periodical segments of the corpus are clustered in a few Sub Sets as a preparation for further analysis. The average set of the Phrases of all years is $MK_{CW1974\_2003} = M_{CW1974} \cap M_{CW1975} \cap ... M_{CW2003}$ (approx. 50% of the corpus Phrases). $MK_{CW1974\_2003} = MK_L \cup MK_D$, where MK_L, contains Phrases that are typical for the language of which the corpus consists and MK_D, which consists of Phrases that are representing the constant basis of the Knowledge Domain. For the last period found (1999-2003), after an additional subtracting of Phrases, which belonging to MK_{CW1974_2003}, the number of Phrases was reduced by 82% until this step. In order to differ between the been left Phrases the usage of Ontology based semantically concepts are needed. For this analysis a prototypical Ontology was build manually out of the corpus. The Phrases were assigned to dimensions, e.g. *Vendor (German: Anbieter)*, *Programming Language*, et cetera. The built Ontology is a simple directed graph representation of the semantic of the Phrases of the corpus. In this step, also existing externally defined and more complex Ontology's can be integrated instead.

For the Phase of **Domain Knowledge Interaction** the previously prepared and clustered corpus data is presented to the user in a way, which allows not only visualization, but also an interactive analysis of the data. The philosophy of the visualization component as a core part of the TMF is based on the metaphor of a "Trend Landscape". Two-dimensional concepts,

e.g. "ThemeRiver" [Havr02] at which selected Frequencies are represented in a kind of topic flow in the course of time, are limited concepts conditionally in their representation and interaction ability.

A: Schematic View in Fig. 4 shows, how the Meta Data (M) of the text corpus (e.g. Frequency or monotone Transformations of it), which is represented in the two-dimensional Phrase/time layer, in the projection step (P) into the three-dimensional detail dimension is to be projected. According the geographical metaphor the Meta Data become represented as "River" in the Trend Landscape. The Hypes (H) and Trends (T) defined before are rising as clusters, the so-called Dimensional Mountains over the set of the Phrases that occurring in all periods (MK_L and MK_D). The methodology of the Trend Landscape also supports the idea of the OLAP analysis methods "Slice and dice" as well as "Drill down and Drill through" and navigating along detailing paths (e.g. "Market participant" → "Vendor" → „Netscape") using the Ontology, which was built, based on the Domain Corpus. In every step of interaction it is possible to disaggregate or aggregate the view on the data. This capability is similar to usual Data Warehouse Tools. In B: Clustered Segment View (Dimension Level) Dimensions that are part of MK_D are shown clustered in the periodical patterns, which were introduced earlier. The Rank of Phrases was transformed that a higher value represents a higher Frequency of the Phrases.

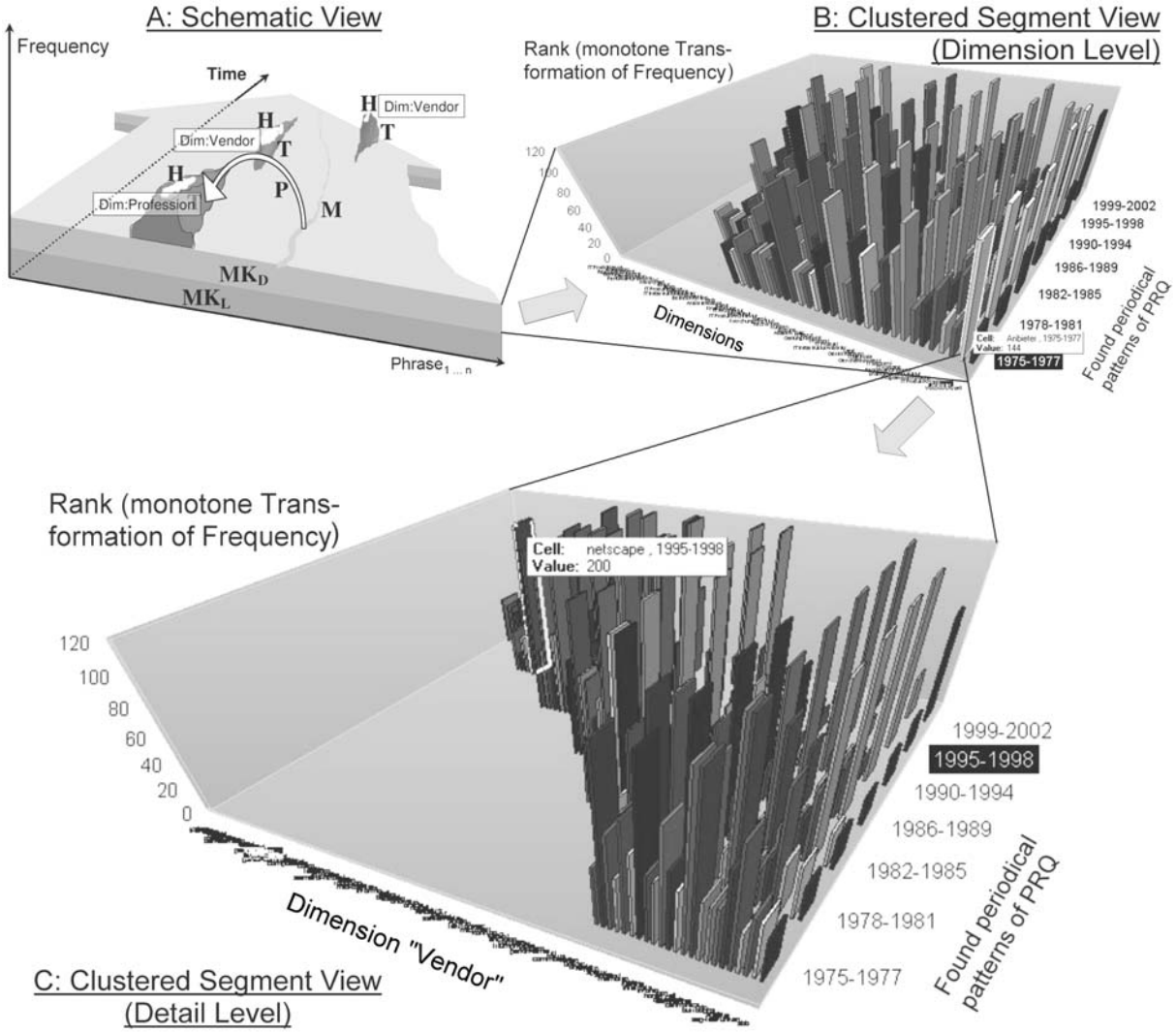

Fig. 4: "Trend Landscape" Views in the process of Domain Knowledge Interaction

The “Drill down” of the Dimension “Vendor” is shown in C: Clustered Segment View (Detail Level). It can be seen that some of the Phrases do not occurring in every periodical patterns, e.g. “Netscape”, which is present only from 1995-1998 as a Phrase with a higher Frequency than the average PRQ. Here it is possible to distinguish between a Hype Phrases (single occurrence) and a Trend Phrases (multiple occurrence).

5. Conclusions and Outlook

The analysis of a large Technical Domain Corpus was properly supported by the TMF. Meta Data Patterns of PRQ were used to segment the Phrases of the corpus due to their persistence character over time. The built segments can analysed separately, according to the aim that the Domain Knowledge Researcher wants to reach. Using Ontology’s, it is possible to assign each Phrase to a dimensional structure, which allows navigating through different views of the prepared Domain Knowledge representation by intuitive OLAP navigational concepts. Thus, the TMF concept is open, there are a few degrees of freedom for adapting this procedure regarding individual research needs by: a) Using different Meta Data Measures. b) Integrating various mathematical or statistical procedures into the Projection step (P). c) The rules for building clusters (in the actual example: Trends and Hypes) can be free defined. To track semantically changes, e.g. the changing of the semantics of the Phrase “Mailbox” over the time, instead of the current used manually built semantically Ontology the use of similarity measures is appropriate. In [Senel04] the Term Document Frequency (TDF) as a candidate discriminator measure is proposed instead of a hard computable cosine of all combinations within a term vector space model. They found, that terms, which appearing in 1 to 10% of the documents are good discriminators. Based on the assumption that Phrases, which are good discriminators for documents, also good discriminators for Semantic Dimensions, in further realizations additional measures for similarity are planned to be used in future.

Sources

[Alla02a] Allen, James: Introduction to Topic Detection and Tracking. Hrsg.: Allen, James: Topic Detection and Tracking: Event-based Information Organization. Massachusetts, Kluwer Academic Publishers, 2002

[Alla02b] Allen, James; Lavrenko, Victor; Swan, Russel: Explorations within Topic Tracking and Detection. Hrsg.: Allen, James: Topic Detection and Tracking: Event-based Information Organization. Massachusetts, Kluwer Academic Publishers, 2002

[Atte71] Atteslander, P.: Methoden der empirischen Sozialforschung (Methods of empirical social research). 2. Auflage, Berlin, de Gruyter, 1971

[Biss99] Bissantz, Nicolas: Aktive Managementinformation und Data Mining: Neuere Methoden und Ansätze. Hrsg.: Chamoni, Peter; Gluchowski, Peter: Analytische Informationssysteme: Data Warehouse, On-Line Analytical Processing, Data Mining. 2. Auflage, Berlin, Heidelberg, New York, Barcelona, Hongkong, London, Mailand, Paris, Singapur, Tokio, Springer, 1999 ISBN 3-540-65843-2.

[Codd93] Codd, E.F.; Codd, S.B.; Sally, C.T.: Providing OLAP (on-line analytical processing) to user-analysts – an IT mandat. White Paper. E.F. Codd & Associates, 1993

[Crou90] Crouch, C. J.: An approach to the automatic construction of global thesauri. Information Processing and Management. 1990, 26, p. 629-640

[Dörr00] Dörre, J.; Gerstl, P.; Seiffert, R.: Text Mining. Hrsg.: Hippner, H.; Küsters, U.; Meyer, M.; Wilde, K.D.: Handbuch Data Mining im Marketing. Braunschweig, Vieweg, 2000

[Düsi99] Düsing, Joachim: Knowledge Discovery in Databases und Data Mining. Hrsg.: Chamoni, Peter; Gluchowski, Peter: Analytische Informationssysteme: Data Warehouse, On-Line Analytical Processing, Data Mining. 2. Auflage, Berlin, Heidelberg, New York, Barcelona, Hongkong, London, Mailand, Paris, Singapur, Tokio, Springer, 1999 ISBN 3-540-65843-2. .

[Fayy96] Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.: From data mining to knowledge discovery: an overview. Hrsg.: Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R.: Advances in knowledge discovery and data mining. Menlo Park (California), 1996 S.p. 1-34,

[Havr02] Havre, S.; Hetzler, E.; Whitney, P.; Nowell, L.: ThemeRiver: Visualizing thematic changes in large document collections. IEEE Transactions on Visualization and Computer Graphics. 2002, *8(1)*, Jan – Mar 2002

[Kall03] Kalledat, T.: Separation of long-term constant elements in the field of information technology from short existing trends based on unstructured data. Hrsg.: Viehweger, B.: Perspectives in Business Informatics Research, Proceedings of the BIR-2003-Conference. Aachen, Shaker Verlag, 2003 S. 167-183,

[Kall04] Kalledat, T.: Perspektiven der Nutzung von Data-Mining-Technologien in der Energieversorgungswirtschaft. ew-Elektrizitätswirtschaft – Das Magazin für die Energiewirtschaft. 2004, *7*, S. 48-53, 1619-5795-D9785D.

[Lend98] Lenders, Winfried; Willée, Gerd: Linguistische Datenverarbeitung. 2. Auflage, Opladen/ Wiesbaden, Westdeutscher Verlag, 1998

[Moen00] Moens, Marie-Francine: Automatic Indexing and Abstracting of Document Texts. Massachusetts, Kluwer Academic Publishers, 2000

[Senel04] Senellart, Pierre P.; Blondel, Vincent, D.: Automatic Discovery of Similar Words. Hrsg.: Berry, M.: Survey of Text Mining: Clustering, Classification and Retrieval. New York, Springer, 2004

[Spil02] Spiliopoulou, M.; Winkler, K.: Text Mining auf Handelsregistereinträgen: Der SAS Enterprise Miner im Einsatz. Hrsg.: Klaus D. Wilde, Hajo Hippner, Melanie Merzenich: Data Mining: Mehr Gewinn aus Ihren Kundendaten. Düsseldorf, Verlagsgruppe Handelsblatt, S.S. 117-124,

[Tan99] Tan, A. -H.: Text Mining: The State of the Art and the Challenges. Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases. Peking, 1999 S.S. 65-70,

[Zipf49] Zipf, G. K.: Human Behavior and The Principle of Least Effort. Cambridge, Mass, Addison-Wesley, 1949