Automatic Detection, Classification and Visualization of Trends in large technical domain corpora

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

Attendant:
PD Dr. Bernd Viehweger,
School of Business and Economics of the Humboldt University in Berlin,
Spandauer Str. 1, D-10178 Berlin, Germany, Email: bv@wiwi.hu-berlin.de

Exposé

Motivation.

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languages such as XML for tagging of textual data stands only at the beginning of its development path and is for historical documents therefore not to be found.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. There are three implications most relevant:

- **First,** 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.
- **Second,** 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 production 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.
- **Third,** 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 term "Data Mining" since the early 1990ies. The Data Mining process allows to discover formerly unknown, potentially useful knowledge patterns out of various input data using techniques and methods of statistics, linguistics and data base research.

These are challenges for the education sector as well as for the information systems that support these processes of Knowledge Discovery (KD). Therefore there is a market demand for turning the knowledge discovery process itself from an individual approach of a small number of specialists to a process that supports a large amount of “Knowledge Worker” in firms and organizations.

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, relevant for the future.

Problem Definition.

For the domain knowledge researcher thereby 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 can not 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.

To support individuals in the knowledge discovery process 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. 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.

For a domain researcher it is important to know, how specific domain semantics is changed over the time, which topics are growing, falling an which are the domain specific basics. For distinguishing between these clusters rules for significant decisions are needed. An important challenge for research is to define methods, which can track changes if a significant pattern is found.

Objective of the work.

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 economical useful knowledge. 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 Content Analysis. The following main research fields can be found: *Information Retrieval, Early Trend Detection or Topic Detection and Tracking, Clustering and Classification.* As a special research field Text Mining was developed for the computer-based analysis of unstructured textual data. 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. Applications based on the underlying methods are realized e.g. in the fields of *Automatic News Summarizing 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. The linguistic methods can be differed into two main directions of research activity: *a) The predominant quantitative approach, that uses general measures of text corpora, e.g. term frequency for evaluation an comparison of differ-*

ent text sources. b) The more qualitative approach, that makes use of interpretation techniques and is working with thesauri for word analysing purposes (e.g. stemma finding).

Instead of working with “Words” the term “Phrase” is used from now which covers a wider range of alphanumeric combinations like product names and technical norms.

The objective of the current dissertation is to propose and evaluate appropriate methods for Automatic Detection, Classification and Visualization of Trends in large technical focused domain corpora.

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 found. For the tracking of Trends in technical Domain corpora over time these known bottom up procedures are having limitations:

1) There are performance issues when real large corpora is analysed.
2) The patterns found are based on generalizing results of multi parametric algorithms, which means that there is to be expected a biased result due to the multiplication of error terms.
3) Linguistic approaches are not appropriate, because technical information is not covered or is destroyed during stemma finding processes.
4) The generation of action recommendations is not transparent to “normal” users.

To reach the goal of Automatic Detection, Classification and Visualization of Trends in large technical domain corpora a promising approach is to combine classical Meta Data oriented approaches in the first step with a projection step of found pattern into the detail layer of phrases, in order to overcome this lack of research. Therefore a top down procedure is proposed, which uses appropriate Meta data measures of the corpus for a first rough pattern recognition process. After this a dramatically reduction of the amount of data is possible by filtering only such data that is found to be worth to be analyzed in further Data Mining steps.

Appropriate methods for building hierarchical structures of elements, e.g. phrases, are known under the term Ontology. These are concepts of directed graphs, which can model relations between elements of domain corpora. Based on such concepts deeper analyses are possible in later steps.

Methodical Approach.

In this context, the Trend Mining Framework (TMF) was developed to make the cognitive access to large technical oriented text sources more easily. The TMF is a proposed methodology and also a process of Text Mining, which makes it possible to evaluate in text form available historical sources of information and to extract extensive time-related domain knowledge semi automatically. By analysing appropriate Meta Data in the first step, the complexity for pattern recognition is reduced dramatically. As an example, the historical development of the domain “Information Technology” was analysed using the TMF. The characteristic of this methodology lies in the ability to identify temporal developments to differentiate these regarding its persistence characteristics in short living Hypes and long-term Trends and 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) Automatic Parsers for separation from contents and formatting as well as converters to ASCII
2) Analytical components for the determination of the Meta Data of the corpus
3) Parsers for Feature Extraction, e.g. phrase frequency
4) A 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 corpora
5) Tools for statistical analyses and tests
6) A visualization component for the representation and navigation

The philosophy of the visualization component as a core part of the TMF is based on the metaphor of a “Trend Landscape”. Here a concept is introduced, which uses a conversion of central quantitative measures into a 3-dimensional graphically representation. Fig. 1 shows schematically, how the Meta Data (M) of the text corpus (e.g. Frequency), 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 rises as cluster, the so-called Dimensional Mountains over the quantity of the phrases that occurring in all periods (MK_L u. MK_D).

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.

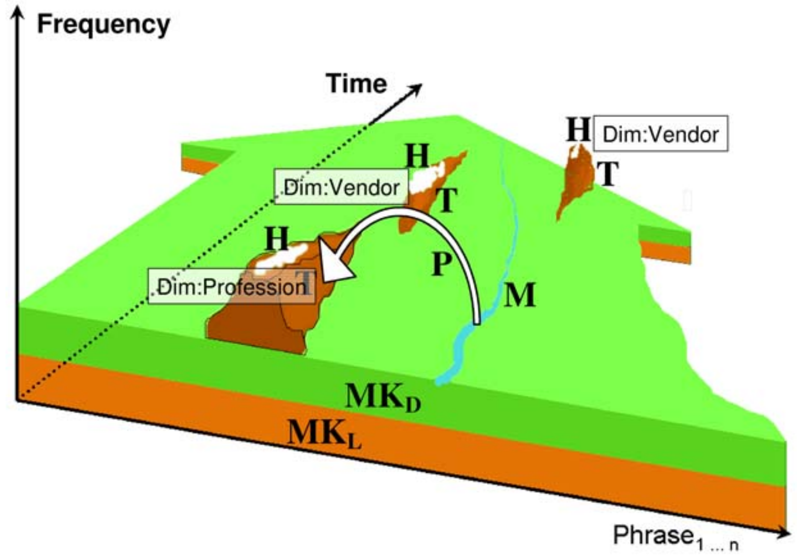

Fig. 1: "Trend Landscape” for the visualization of Trends and Hypes

The concept of the Trend Landscape also supports the idea of the OLAP analysis methods “Slice and the dice” as well as ”Drill down and Drill through” and navigating along detailing paths (e.g. “Market participant” → “Vendor” → „IBM“) using the ontology which was built semi-automatic out of the domain corpus.

Main Sub-Tasks that have to be covered in the dissertation for building the TMF are:

- Proposing and evaluating quantitative Meta Data Measures as entry points for the top down analysis approach
- Building a time dependent ontology for tracking trends and covering semantic changes
- Measuring the confidence of the proposed approach regarding the quantity and quality of the needed text corpus

This dissertation advances the research on Trend detection and Tracking by:

- Providing a Framework for Automatic Detection, Classification and Visualization of Trends in large technical domain corpora
- Evaluating a Top Down Meta Data measure driven approach for focusing on relevant pattern
- Suggesting a visualization concept to make the process of Trend tracking more intuitive