Semantic Search and Visualization of Time-Series Data

Tatiana von Landesberger, Viktor Voss, and Jörn Kohlhammer

Abstract. In the economic and financial analysis domain, quick access to the right information plays a major role. Using current systems, the search for and presentation of data is very cumbersome. The data, mostly in form of time-series, is stored in various databases. In order to retrieve the searched data, the analysts need to know where to search and sometimes even the structure of the database and its coding. Then it is required to export the data, process the data and create a chart to view the data. This might take time from tens of minutes to hours. In our work we present a first prototype of an integrated search engine that takes as input a natural language query and offers graphic and text output depending on the user task. The system automatically identifies the resulting time-series and types of graphical data presentation, and shows the results in a web browser or in Excel. The knowledge-based expert system uses domain ontologies for extraction of economic terms in the search queries and specially built data type taxonomies with user task and chart type ontologies for the identification of graphical output.

1 Introduction

In the economic and financial analysis, rapid access to the right information plays a major role. Often large amounts of data have to be evaluated in short time. The data are usually stored in heterogeneous systems and in various databases. In their work, analysts combine numerical data from various sources with text (such as news, reports, etc.), and expertise that exists in the company.

Tatiana von Landesberger
Technische Universität Darmstadt, Darmstadt, Germany
e-mail: `tatiana.von.landesberger@gris.informatik.tu-darmstadt.de`

Viktor Voss and Jörn Kohlhammer
Fraunhofer IGD, Darmstadt, Germany
e-mail: `{viktor.voss,joern.kohlhammer}@igd.fraunhofer.de`

S. Schaffert et al. (Eds.): Networked Knowledge - Networked Media, SCI 221, pp. 205–216.
springerlink.com

In order to collect all the required information, analysts spend a lot of time every day with information search. Analysts usually simply browse the data sources or use data search options available by the data providers. These conventional systems for searching for financial and economic information, however, do not offer naturally formulated search queries such as ”GDP composition in Germany” with direct graphical output.

Finding the relevant data does not end the analysis task. In the analysis process, after the data is gathered, the data is copied by hand into another system for further processing (e.g. into Excel). Finally, appropriate graphics for data presentation are created. In addition to the domain expertise this graphics construction requires also skills in information visualization and graphics design from the analyst.

In this paper, we introduce a first prototype of a semantic-based search engine which takes as input a natural language query, searches over financial information across multiple financial data services and displays the results in a way that suits best the user tasks (e.g. as table, text or graph). A preliminary user study to examine the common analytic tasks with matching answers was conducted. In a later stage, selected end-users were asked for feedback. Results of both studies are presented.

The paper is structured as follows: Section 2 discusses relevant work about automatic graphical data representation. The third section describes current user search tasks and used systems for data search. In the Section 4, our framework is presented. Section 5 provides examples of search output visualization. After giving an overview of related visual analytics research initiatives, we conclude and discuss future work.

2 Related Work

Automatic generation of graphical data representation goes back to the work of Mackinlay [17]. The Automatic Presentation Tool (APT) defines graphical representation based on the description of visual attributes. Roth and Mattis based their SAGE system on Mackinlay’s approach for designing two-dimensional static presentation of relational data [18]. Casner introduced the BOZ system taking a step further toward user-centered design with a task-analytic approach [6]. BOZ concentrated on the design of graphics that optimize human performance in information processing tasks. The idea was to replace logical inferences that are cognitively demanding with faster perceptual inferences. However, all of these approaches and their successors today are still computationally too inefficient for interactive applications. According to the visualization design methodology, we can divide current systems into two categories - constructive bottom-up and template-based top-down method [14]. The comparison of the two types of techniques can be found in [16]. In this paper, we follow a top-down approach that is strongly supported by semantics. The development of data type and visualization taxonomies was based on the work by Shneiderman [19] and by Tory et al. [22] which match different data types with visualization driven by user tasks. A general overview of analytical tasks in visualization is provided by Amar et al. [3]. Amar et al. [2] earlier described a knowledge task-based framework for design and evaluation of information visualizations. Fujishiro et Al. [13] introduce a taxonomic approach to semi-automatic design of information visualization applications using modular visualization environment. An overview of visualization techniques specialized for time-series can be found in [1]; a taxonomy for temporal data visualizations can be found in [11]. Kohlhammer [15] shows the use of domain ontologies for effective visualization. We discuss this approach in more detail in Section 4.

3 User Tasks and Current State of Data Search

In this section we describe the current working environment of financial and economic analysts and the questionnaire that was sent out to these analysts. This dialog with the application end users was very important as a guidance for our research in this area. Only those visualization solutions that are embedded in the current work flow of users will be successful solutions.

3.1 Search Tasks

In order to best meet the analytic requirements, we have asked ten financial and economic analysts from various work domains, what kind of search engines they use to find the data and what systems they use for data presentation. They were asked to provide as many data queries as possible including their desired outcomes and estimation of the time needed to accomplish the analytical task (see Table 1).

More than two thirds of the searches are aimed at getting data in tables, single data observations or data charts (see Figure 1). Chart responses included line, column, stacked column and scatter plot charts. The respondents use financial and internal databases and copy this data to create charts in Excel. The time needed to accomplish such tasks (from data search to the production of relevant data presentation) varies strongly according to the desired output and the familiarity with the searched data sources. Searching for novel information that is not accessed by the user on a regular basis is much more time-consuming than a search in well-known data sources. On average, users spend 18 minutes when creating charts, 38 minutes when creating tables and only 3 minutes when looking for a single value. However,

Table 1 Example answers to the questionnaire

Question	Expected Output	Expected Duration
Show me EUR/SKK development in the last week	Intra-day line chart	5 mins
DJI index in the past year	Tick data, line chart	5 mins
Show me GDP composition in Germany and France?	2 pie charts showing GDP components	30 mins

Fig. 1 Required search query responses. Types of search responses.

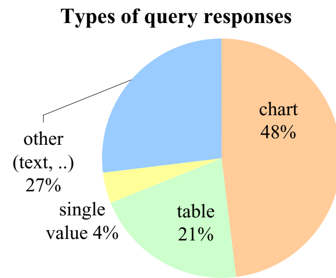
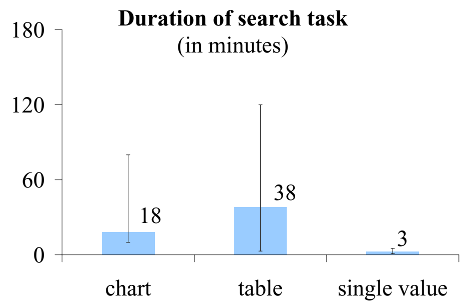


Fig. 2 Time needed for response by type (average, minimum and maximum)



the difference between minimum and maximum time is very high (see Figure 2) with a maximum of 4 hours for one task.

3.2 Data Sources and Currently Used Search Engines

The sources of information needed are spread across multiple databases from various data providers (e.g. Bloomberg, Reuters, DataStream, Eurostat, ECB or inter-

Fig. 3 Example of search results using the ECB search engine (http://sdw.ecb.int)

nal sources). Each data provider uses their own data classification system and data description system making the search for data across databases very complicated. Search in such databases is usually constrained to explorative overview of database entries by category. Some providers offer search engines which should facilitate the search for data. These search functions, in general, offer as outcome long lexicographic ranking (lists) of the names of the available time series (see Figure 3).

4 Semantics-Based Time-Series Search and Visualization

This section introduces our main concept of semantics-based search and visualization. Based on a general approach called decision-centered visualization, we designed a framework for query processing and chart generation. At the core of this approach is the semantics-based determination of a suitable chart type for the data and task at hand.

4.1 Decision-Centered Visualization

Our framework follows the decision-centered visualization (DCV) approach [15]. This approach uses knowledge representation, in particular domain ontologies and meta-databases, to filter and prioritize information and events dynamically depending on the current task type and user role. In contrast to a simple filtering algorithm, this approach takes at the same time the visualization requirements of the events

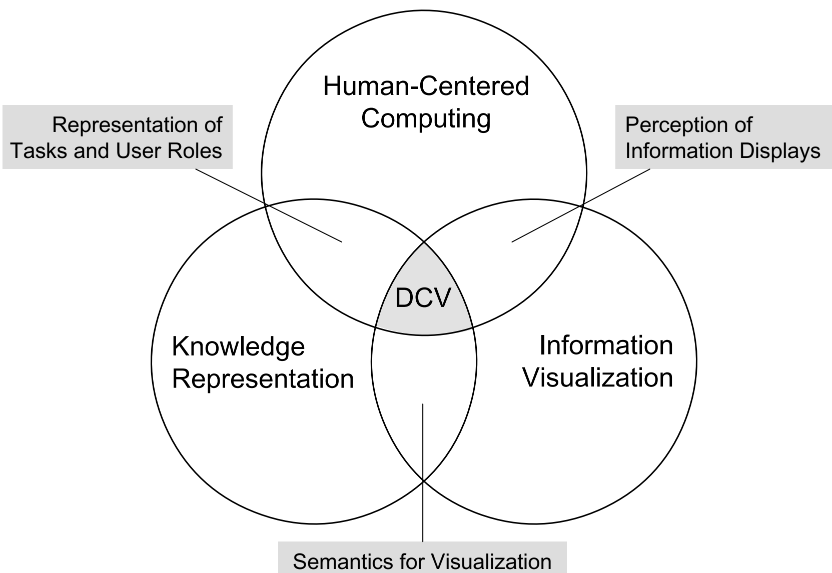


Fig. 4 Overview of the decision-centered visualization approach

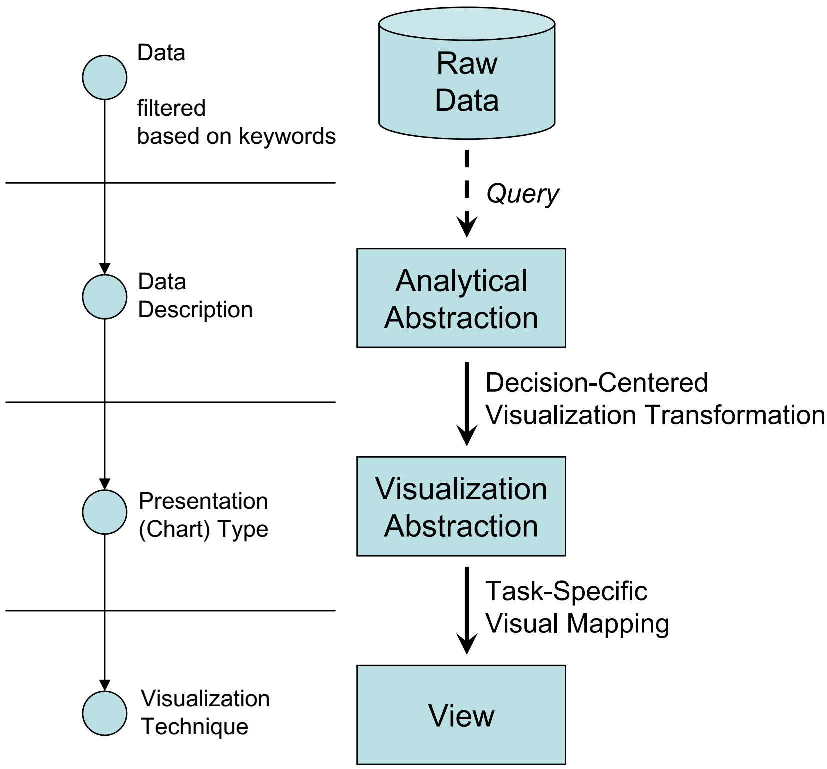


Fig. 5 DCV-specific adaptation of Chi's data state model

and information into account to be able to support effective and focused visualization (see Figure 4).

The connection between information visualization and knowledge representation is the support of determining what to visualize or how to visualize with the help of represented metadata and knowledge. The visualization of knowledge representation structures like ontologies, is not as relevant here, though these techniques are necessary on another level for creating and maintaining the ontologies. At the heart of this approach lies the mechanism to represent the presentation knowledge, i.e. the presentation requirements. These requirements are based on data types similar to those of Shneiderman, Card, and Mackinlay [19, 5] while their handling is based on an adaptation of Chi's data state model [7] (see Figure 5).

4.2 Search and Visualization Framework

Our knowledge-based system for integrated data search and visualization consists of three major parts (see Figure 6): in the first stage the input query is processed in order to extract relevant search criteria. Then a suitable set of responses is compiled and extracted from the database. Finally, the results are shown in a web browser. Our system makes use of the ConWeaver framework [9]. ConWeaver provides automated knowledge network construction, semantic integration and intelligent search in portals and intranets.

4.3 Query Processing and Response Identification

In the first stage of our framework (see Figure 5 and Figure 6), the input query is processed in order to extract economic and financial terms in the query. In our work, we have created modules that identify economic and financial terms in the data descriptions, by looking for longest terms first. For example "GDP per capita" is a different term than "GDP". Furthermore synonyms (consumer prices and inflation) and abbreviations (e.g. GDP = gross domestic product) can be extracted.

For defining the suitable data presentation type, we have analyzed visualization types used in the financial and economic domain as well as the results of our user study. A template-based system for graph generation, which can be adjusted by the users, is applied. In the system, each presentation type is described by the type of data input, visualization parameters and the analytical purpose (keywords).

For this application, a description was created for each chart type identified in the user questionnaire (line, column, pie, scatter plot, etc.). This description was saved in the chart type ontology. It describes types of data that the specific chart can use as input and the user tasks for each chart type. Each chart parameter (each chart axis) is described via the type of data that it takes as input (quantitative, nominal, ordinal, etc.) and its cardinality. For example a line chart has 3 "axis": X axis that usually takes quantitative data of cardinality ≥ 5, Y Axis takes quantitative data of cardinality ≥ 1, and Legend takes nominal data of cardinality 1-10 (more than 10 series usually make the graph overcrowded). Chart is also characterized by "keywords" indicating user task connected to the specific chart type. The keywords were

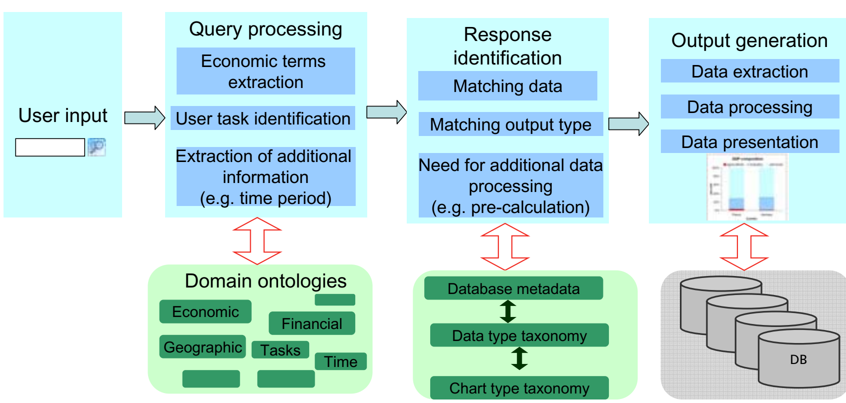


Fig. 6 Overview of the search and visualization framework

identified via matching search queries with expected responses (see in Table 1). For example, “composition” indicates pie chart, “development” indicates line chart.

4.4 Output Generation

In the output generation stage, the identified suitable response data are collected from the database. The output type is influenced by the rank of matching data, the fit of the data to specific chart types and the user task match with the chart keywords. The output data type and the identified user task are matched with the best possible chart type. The chart type match (see Figure 7) uses a weighted data type hierarchy for defining the best possible output type for a given data set to be shown. The output match is quantitatively expressed by the product of the data type hierarchy step weights. The weights are pre-set parameters, which can be adjusted. Additionally the user tasks influence the output type. Additionally, the weight of each chart type is influenced by matching the chart keywords with the search query. A ranking list of output types is thereby created. The top ones are used in the output visualization.

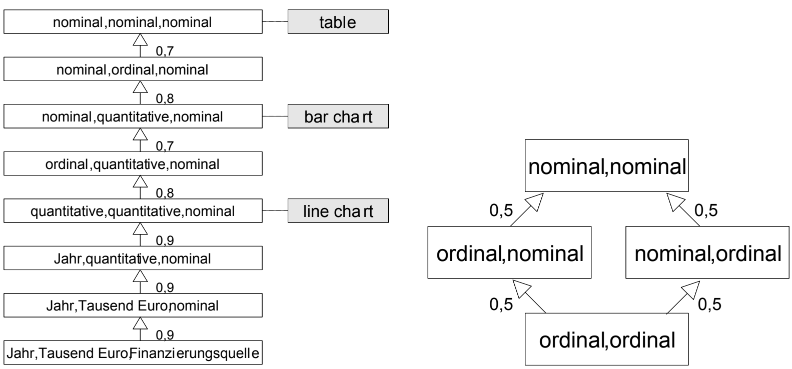


Fig. 7 Semantics-based determination of the presentation type (left) using a data set type hierarchy (right)

5 Search System Output and Evaluation

For testing the search and visualization framework, real world data sets from the Federal Reserve Bank of St. Louis (http://research.stlouisfed.org/fred2/) were used. The FRED database contains about 14,000 economic and financial time series in flat text files. It provides raw data and time series descriptions.

The engine displays the search results on a Web front-end in the form of dynamically generated HTML pages based on the search results. The data that best match the user query are displayed in a graphical form determined by the response identification phase (see Figure 8). In our prototypic application, the user may click on each chart to see the underlying data table in Excel (see Figure 9). The users can then further process the data results or design further charts in this application. Other export formats can be employed on demand. In addition to numeric data, the search engine offers a list of documents searched on the Internet using common text search engines (e.g. Google). Using the ConWeaver system [9] we can also easily include the possibility to search in internal documents, however we did not have any such data at our disposal. Thus, our approach provides a composite of relevant information in one window.

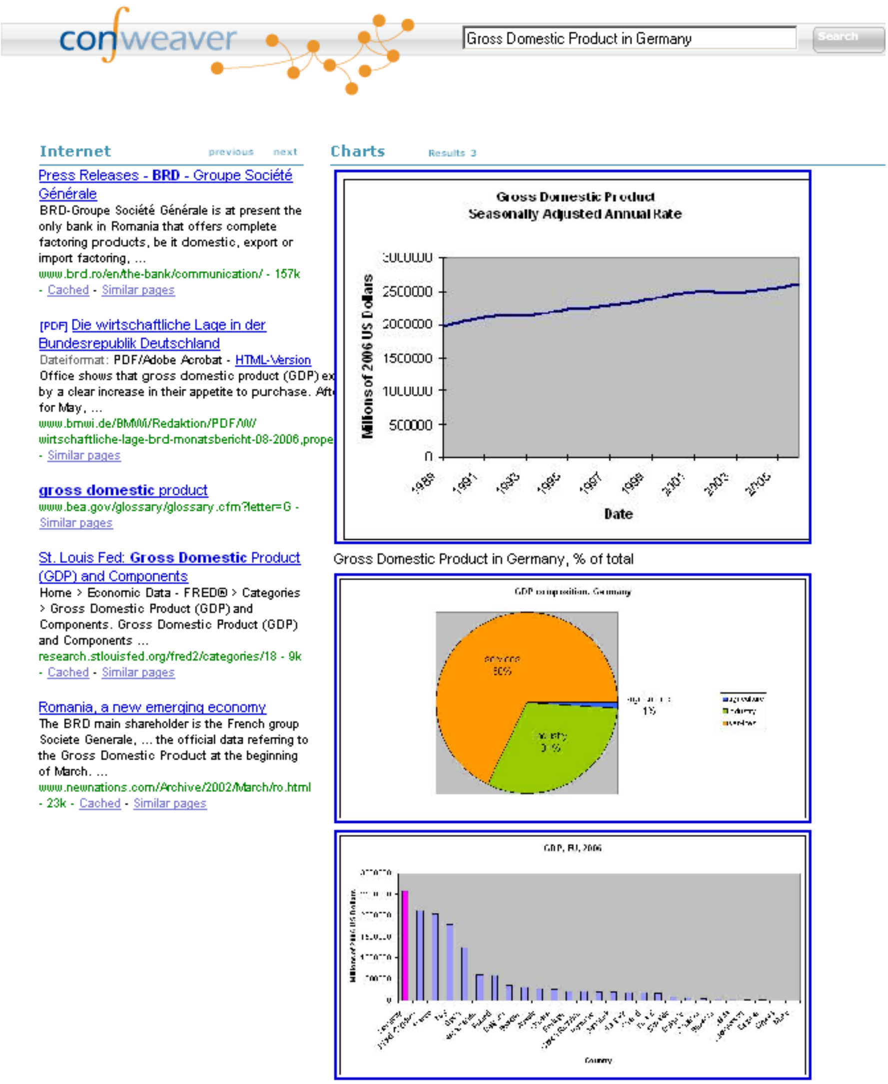


Fig. 8 Example search result using our prototype tool

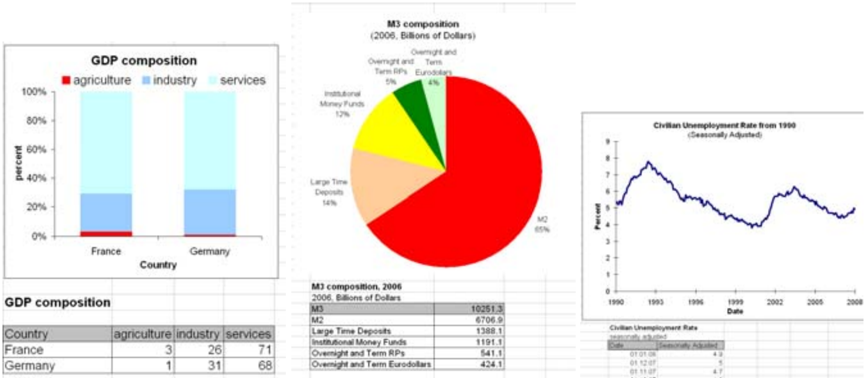


Fig. 9 Selected user responses exported to Excel

The prototype user queries were used to test the results of our search engine. The users liked very much the possibility to see an overview of the search responses in graphic/table format. It enables them to quickly identify major data developments without cumbersome data export and graph creation for each single data series. This quick and easy way of preliminary data analysis saves the analysts' time which they can use for more in-depth data analysis using e.g. data mining methods.

6 Relationship to Visual Analytics Research Initiatives

One larger extension of this work is currently implemented within the project THESEUS [21] – the development of a semantics-based Visual Analytics framework. This framework is intended to provide a visual connection between "high-level" (semantic) data and "low-level" data - the actual content. This connection works in both directions, each of which is necessary to support specific tasks. The primary idea of the Visual Analytics Framework is not to provide means to navigate and use semantic knowledge, but more importantly to create, develop and verify this knowledge, represented in the semantic structures. This bottom-up approach is realized using data analysis techniques like feature detection, similarity identification, cluster identification etc.

In a certain sense, the work on this framework extends the research by [7] to Visual Analytics. Chi proposes his data state model for the analysis of visualization techniques to provide a clearer understanding of the interactions between data and operators. He describes 36 different analyzed techniques with to have a classification and a selection of how to implement different operators in a large visualization system. It can be seen that several techniques share different operating steps that could be standardized or modularized for reuse in other systems.

Visual Analytics is a field, where analysis techniques are used in conjunction with graphical displays that integrate the user and his specific abilities and interests into the process. Visual Analytics as a field requires standards, which allow a broad interchange of data of very different types between the different components of the system for the data integration, analysis and interaction. The Visual Analytics Framework will serve this purpose and will revise existing techniques in order to asses their value for a specific use-case.

7 Conclusions

A prototype system for search in economic time-series databases which takes as input a natural language query, and graphically visualizes the search results has been developed. The choice of the proper visualization type is task-dependent and is determined using a data taxonomy and a visualization type ontology. Domain ontology for economic data is used for finding suitable search responses. The system offers time savings to the potential users when looking for and presenting economic and financial data allowing them to concentrate on the data analysis part of the analytic process. Our approach overcomes the computationally time-intensive approaches for automatic graphics design by following a semantics-based top-down method, which is combined with a decision-centered visualization approach for generating visualization familiar to and expected by the user for the task at hand.

8 Future Work

In the near future we will explore the effect of different parameter setups for the hierarchy type match. It will also be necessary to expand the used financial and economic ontologies using new data sources (e.g. further classifications) and newly developed publicly available ontologies. This would allow for more accurate search responses. We would like to include automatic data processing using financial functions (e.g. calculation of indices) and widen the spectrum of used visualization techniques. It would be interesting to combine search for numeric data with search in news feeds. For example, particular patterns (peaks, strong decreases) could be matched with the relevant stock market interpretation.

Acknowledgements. This research was partly funded by the German Federal Ministry of Economics and Technology within the THESEUS project [21].

The authors thank Richard Stenzel and Thomas Kamps of ConWeaver GmbH [9]for their support and the possibility to use their engine for semantic search. We are thankful to the user study participants for their activity and helpful input. We thank Tobias Schreck for his helpful comments.

References

1. Aigner, W., Miksch, S., Muller, W., Schumann, H., Tominski, C.: Visualizing time-oriented data – A systematic view. Computers & Graphics 31(3), 401–409 (2007)
2. Amar, R., Stasko, J.: A Knowledge Task-Based Framework for Design and Evaluation of Information Visualizations. In: Proc. IEEE Symposium on Information Visualization, pp. 143–150 (2004)

3. Amar, R., Eagan, J., Stasko, J.: Low-Level Components of Analytic Activity in Information Visualization. Georgia Institute of Technology (2006)
4. Audersch, S., Flach, G.: Semantic Web Technologien fr die visuelle Exploration und Fusion multivariater Datenbestnden. In: Proceedings of Berliner XMLTage (2005)
5. Card, S.K., Mackinlay, J.D., Shneiderman, B.: Readings in Information Visualization. Using Vision to Think. Morgan Kaufmann, San Francisco (1999)
6. Casner, S.M.: A Task Analytic Approach to the Automated Design of Graphic Presentations. ACM Transactions on Graphics 10(5), 111–151 (1991)
7. Chi, E.H.: A Taxonomy of Visualization Techniques using the Data State Reference Model. In: Proceedings of the 2000 IEEE Symposium on Information Visualization, Salt Lake City, UT, USA (2000)
8. Codd, E.F.: A Relational Model of Data for Large Shared Data Banks. Communications of the ACM 13(6), 377–387 (1970)
9. ConWeaver, `http://www.conweaver.de` (cited December 1, 2008)
10. Cox, K., Grinter, R.E., Hibino, S.L., Jategaonkar Jagadeesan, J., Mantilla, D.: A Multi-Modal Natural Language Interface to an Information Visualization Environment. International Journal of Speech Technology (2001)
11. Daassi, C., Nigay, L., Fauvet, M.C.: A taxonomy of temporal data visualization techniques. Revue en Sciences du traitement de l'Information. Cepadues Editions 5(2), 41–63 (2005)
12. Date, C.J.: An Introduction to Database Systems, 7th edn. Addison- Wesley, Reading (2000)
13. Fujishiro, I., Ichikawa, Y., Furuhata, R., Takeshima, Y.: GADGET/IV: a taxonomic approach to semi-automatic design of information visualization applications using modular visualization environment. In: Proceedings IEEE Symposium on Information Visualization, pp. 77–83 (2000)
14. Kamps, T.M.: A Constructive Theory for Diagram Design and its Agorithmic Implementation. PhD Thesis, Technische Universitat Darmstadt (1999)
15. Kohlhammer, J.: Knowledge Representation for Decision-Centered Visualization. Dissertation, Technical University of Darmstadt. GCA-Verlag (2005)
16. Lange, S., Nocke, T., Schumann, H.: Visualisierungsdesign - ein systematischer Überblick, German, Universität Rostok (2006)
17. Mackinlay, J.: Automating the Design of Graphical Presentations of Relational Information. Transactions on Graphics 5(2) (1986)
18. Roth, S.F., Mattis, J.: Automating the Presentation of Information. In: Proceedings of the IEEE Conference on Artificial Intelligence Applications (1991)
19. Shneiderman, B.: The Eyes Have It: A Task by Data Type Taxonomy for Information Visualization. In: Proceedings IEEE Conference on Visual Languages (1996)
20. Sowa, J.F.: Principles of semantic networks. Explorations in the representation of knowledge. Morgan Kaufmann, San Francisco (1991)
21. Theseus, `www.theseus-programm.de` (cited on December 1, 2008)
22. Tory, M., Moller, T.: Rethinking Visualization: A High-Level Taxonomy. In: Proceedings IEEE Symposium on Information Visualization, pp. 151–158 (2004)
23. Ware, C.: Information Visualization: Perception for Design. Morgan Kaufmann, San Francisco