A Visual Query Interface for a Very Large Newspaper Article Repository

Wolfgang Kienreich, Vedran Sabol, Michael Granitzer
Werner Klieber, Mattias Lux, Walter Sarka

Know-Center, Competence Center for Knowledge-Based Applications and Systems
{wkien|vsabol|mgrani|wklieber|mlux|wsarka}@know-center.at

Abstract

The archives of large national and international news agencies typically contain millions of articles featuring significant textual content and annotated metadata. Boolean queries and relevance ranked result lists have been traditional means of inquiry in such a context. This publication presents an interface for query formulation and visual query analysis for the news article repository of the Austrian Press Agency, combining traditional means of inquiry with several visualisation components. All parts of the interface are fully synchronised, allowing users to employ the power of visual metaphors for analysis and refinement while retaining the simplicity of conventional query formulation.

1. Introduction

The archives of large national and international news agencies typically contain millions of articles. Each article features significant textual content and annotated metadata like date of release, author or resort. In such a context, boolean queries and relevance ranked result lists have been traditional means of inquiry. However, in the face of radically expanding archives and increasingly complex user requirements, new forms of query formulation and result representation must be considered.

Several visualisation environments for newspaper archives have been presented. Systems like Galaxy of News [1] or, more recently, Lighthouse [2] and InfoSky [3] employ novel visual metaphors in interactive visualisations enabling users to quickly gain a broad understanding of a news base. However, most approaches are based on assumptions which do not easily generalize: For example, systems depend on articles being organized into a graph or on explicit links between articles, or interfaces are modelled on the assumption that users are expert analysts.

In this publication we present a visualisation environment which tries to combine traditional boolean query formulation and relevance ranked result lists with a variety of visualisation approaches into a single consistent system. In a project carried out in close cooperation with the Austrian Press Agency APA [4], a prototype has been developed which will serve as a baseline for APA's future search and retrieval systems.

2. Environment

The Austrian Press Agency APA is Austria's national news agency and one of the country's leading information providers. APA editors supply national and international political, economic and financial news as well as topical reports on science, culture, sports and local events to agency customers in near real-time. APA Defacto, APA's subsidiary specializing in knowledge technologies and services, provides customers with several millions of documents available for enquiry through a client interface based on boolean queries and relevance-ranked result lists.

Articles in the repository are available in plain text, with annotated general metadata like date of creation, author and source. In addition, editors attach classification tags like, for example, "domestic politics" or "life science", to articles manually. Roughly 18 million documents in a date range from 1950 until today can be queried, by full text or metadata attribute, through both an online (web based) and an offline client.

3. Repository

The system described has been based on a database technology called Powersearch specifically designed for newspaper article repositories. This technology has been developed by APA IT, the IT subsidiary of the Austrian Press Agency. It focuses on efficient storage and fast retrieval of short text entities. Basic functionality of PowerSearch has been extended with clustering mechanisms[5][6] applied to search results and a similarity search mechanism based on the "best" words of an article.

refinement terms are extracted from the corpus, which users can then add to query formulations to narrow the scope. Refinement terms are determined from a recent subset of the whole corpus, and clustering is restricted to that subset, too.

Once a result set has been determined, metadata is gathered. Metadata is present in an article due to either editorial annotation or extraction at indexing time. We used shallow text parsing techniques [7] to identify metadata entities not annotated in the article database. In our example implementation, we facilitated person

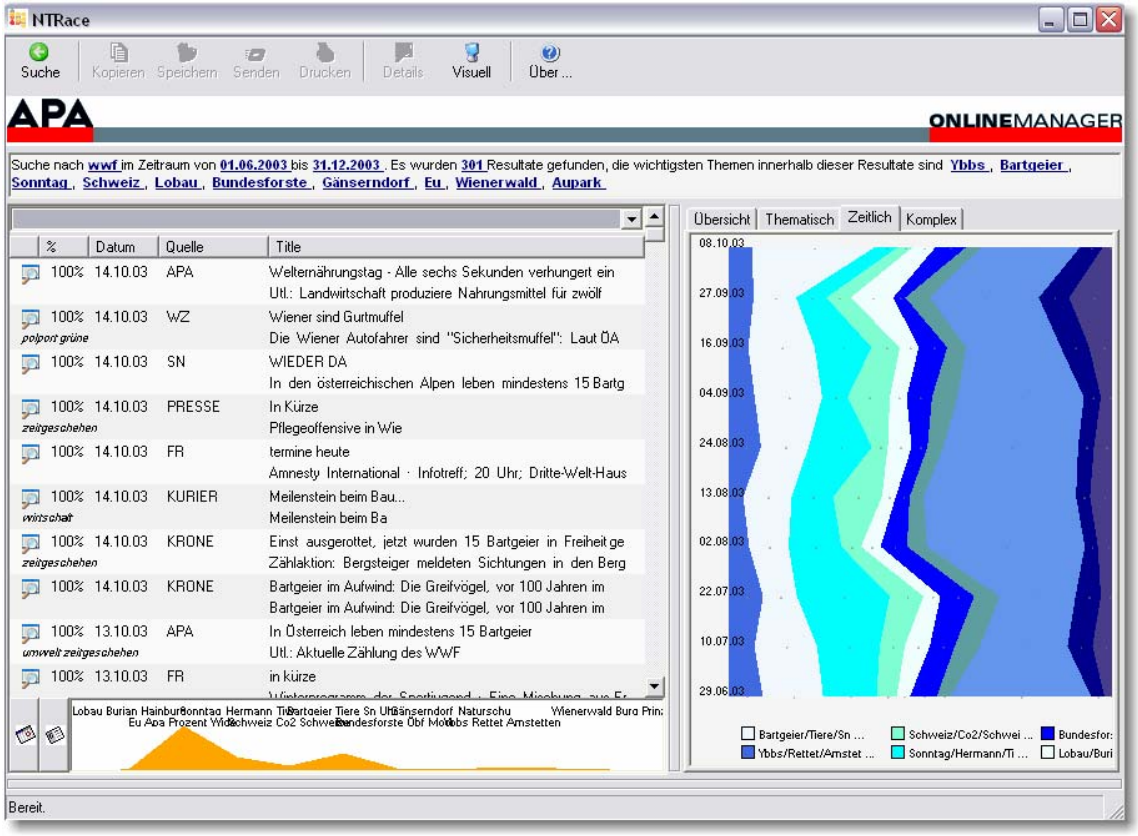

Figure 1: APA Visual Result Browser

Response time requirements for query and analysis were one to three seconds for a boolean query against the whole 18 million document corpus. These requirements were adressed by organizing documents into segments by month (for recent time ranges) and year (for ancient time ranges) and only using time ranges relevant to a given query. This kind of organization is justified by the fact that most queries cover small time intervals (a week to several months). The problem of queries too general in nature, which yield too many results, has been adressed by the introduction of an automated search result extension suggestion mechanism. For very general search terms,

detection based on a firstname thesaurus and simple syntactical rules, and identified geographic locations based on a geographic thesaurus and regular expression matches. Finally, search results are supplemented with visualisation data computed for all visual components, for example by applying multidimensional scaling techniques [8] to determine document locations.

Details on information retrieval aspects are beyond the scope of this discussion and will be provided in a separate publication. We will now focus on interface and visualisation aspects.

4. Interface

The visual interface consists of a visual component framework encapsuling two major steps, query formulation and result presentation. Query formulation has been kept intentionally conventional. Initial result display is in the form of a relevance-ranked list with only a simple visualisation. Additional visual components are available on demand to avoid overloading non-expert users with information.

The purpose of the visual components provided is to allow users to answer questions about the result set quickly, to analyse the result set, without having to reformulate the initial query. Analysis could, for example, determine which results include a specific person or geographic location, or fall into a certain topical range. Reformulating the query to answer such questions is problematic in several aspects: Terms may be ambiguous, requiring elaborate formulations in the query language; Questions may only be answerable by applying multidimensional restrictions (i.e. time and topic) which are hard to formulate in a query language at all; Finally, some forms of analysis, for example regarding result distribution along metadata dimensions, cannot be executed at all by refining the query.

4.1. Query Formulation

Initial query formulation is done using a standard textual query formulation language and some user interface components for selecting query properties like sources (in news archive context, meaning newspapers, magazines and the like) or date range. No explicit visual query formulation (like, for example, MetaCrystal [9]) is facilitated because there was reason to assume that the target user group would not accept such a radical departure from traditional boolean query formulation. Once query results are available, users can employ the visual components present to quickly restrict (and expand) the query without changing its formulation manually.

4.2. Result presentation

Query results are initially presented as shown on the left side of Figure 1: A conventional, ranked result list is combined with a simple graph visualisation which depicts either a representation of clusters found in the result set, and document count per cluster, or one of document count per time unit (users decide which variant to display). Both views are fully interactive, allowing users to select ranges using the mouse. Range selection immediately filters the result list to documents within that temporal or topical range. It was felt that such a simple visualisation would pose no problem even to inexperienced users, which is why it is visible initially. As shown on the right side of Figure 1, additional visualisations can be opened for a given result set. These visualisations have been designed to serve as analysis tools for advanced users: A metadata tree view allows quick narrowing of a result set by selection of metadata instances, a timeline enables temporal analysis of results and a 2D and 3D cluster visualisation provide a topical perspective.

4.3. Article Display and Query Expansion

When a user detects an article of interest within the result set, he can open this article in a separate window. Article display shows metadata directly within the article text: For example, after each recognized person name, a small “human” icon is displayed, and can be clicked to display information about that person if external databases provide such, or simply other documents also talking about the person in question. The same is true for geographical references, clusters and other types of metadata.

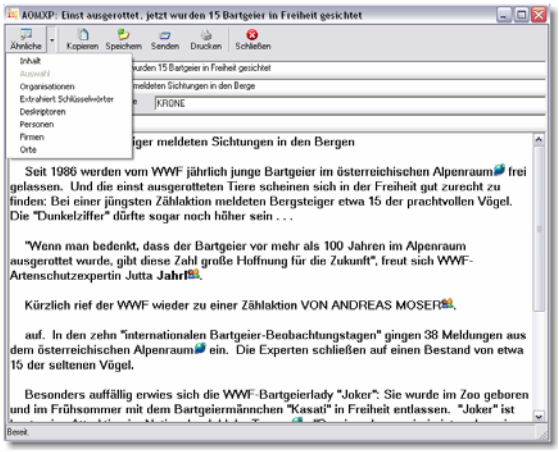

Figure 2: Article Visualisation

Queries can be expanded from the article display by several methods: A general “find similar articles” option determines the 20 to 30 most important terms of an article and generates a query searching for that terms. More special expansion options based on selected text passages or metadata present in an article (e.g. “find articles dealing with the same people as the current one” or “find articles referring to similar geographic locations”) are also available.

4.5. Metadata Visualisation

For each article in a result set, a number of metadata attributes is available. Such attributes may be annotated by editors when articles are added to the archive, or they may be extracted when articles are returned in a result set. In the prototype system, person references were extracted using shallow text processing techniques to identify names, and geographic references were identified by means of geographical thesauri. Such discrete metadata items were then provided to users as means of quickly filtering result sets.

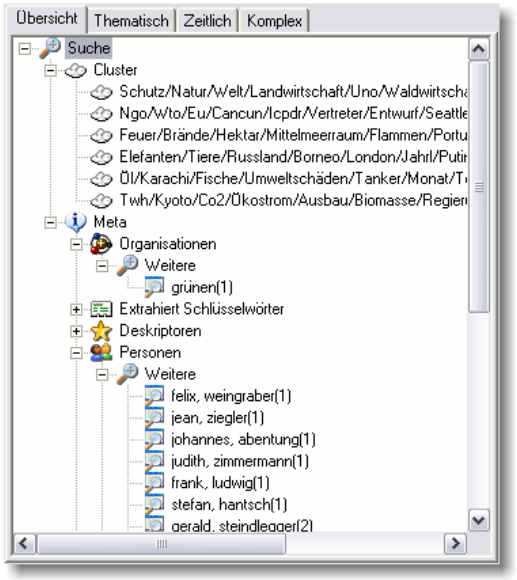

Figure 3: Metadata visualisation

The metadata visualisation consists of a tree of metadata type nodes. Each type node (e.g. Geographic, people) has all instances of the type it represents (e.g. Europe, Vienna) attached as sub-nodes. Each sub-node features an instance count which announces how many articles contain that specific instance. Clicking on a sub-node instantly filters the result list to articles containing the instance selected. The tree of topical clusters found in the result set is displayed close to the metadata tree, because it offers similar functionality.

4.6. Cluster Visualisation

Clusters identified within the result article set of a query are organized into a hierarchy which can be multiple levels deep. The cluster visualisation displays this hierarchy: Each cluster is shown as a bubble surrounded by its sub cluster which are connected to it with linking lines.

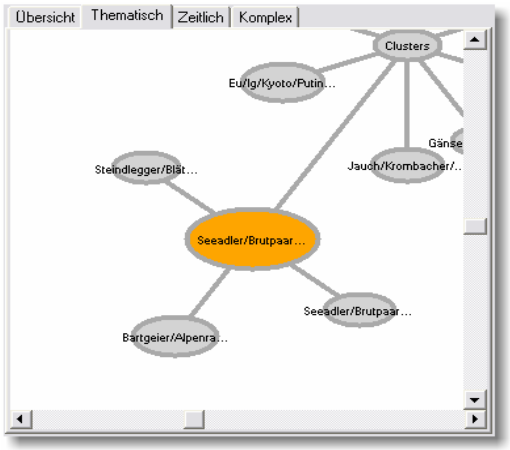

Figure 4: Cluster Visualisation

Users can expand a cluster, and simultaneously filter the result set to articles contained in that cluster, by clicking on it. Further interactions include free panning within the structure of clusters using the mouse. Visual appearance is similar to the DendroMap visualisation [10], but allows for more than two branches.

4.7. Timeline Visualisation

Timeline visualisation displays the relative size of article clusters in discrete time intervals. By stacking the area graphs of all top-level clusters involved into a single rectangle, the visual appearance of a river featuring multiple torrents, running from top to bottom (or from left to right) in time, can be achieved.

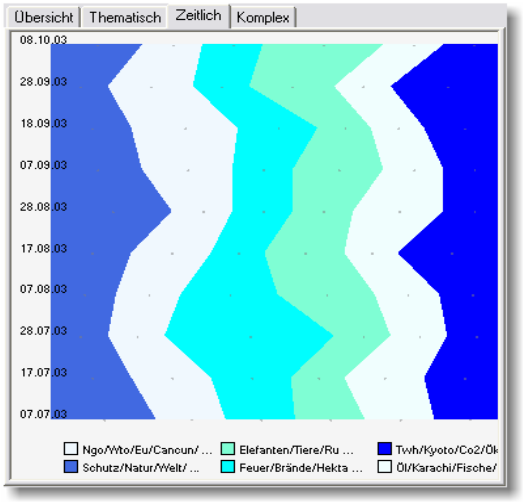

Figure 5: Timeline visualisation

Interactivity in the timeline visualisation includes selection of cluster and time range via mouse: Users can mark a rectangle of interest within the timeline, instantly filtering the range of documents in the result list. Visual appearance and interactivity of the timeline are similar, but less sophisticated, than that of ThemeRiver[11].

5. Evaluation

Formative user testing accompanied the development process of the system presented. For example, labelling of the small graph visualisation instantly visible for each result set has been modelled in several variants and then evaluated by a peer group of ten users by means of a questionnaire. Results of this evaluation have already been integrated into the system. However, summativ and comparative user testing remains to be done: An extensive user test comparing the traditional query interface to the new visual one is scheduled. Test setup includes multiple task assignments like locating specific documents or gathering information on specific topics. User behaviour will be recorded, and noted by an observer. Post-session interviews will supplement the material thus gathered. Statistic analysis of collected data will provide quantitative information on the performance of the new visual interface. We will report on the results in a separate publication.

6. Future Work

Future work will include the integration and evaluation of 3D visualisation components into the system. A prototype 3D cluster visualisation based on the visualisation islands approach [12] has already been added. Evaluation and comparison to existing solutions has a high priority in upcoming project phases, too. Finally, deriving a variant of the interface suited for web-based deployment is envisioned.

7. Concluding Remarks

This publication presented a visual interface for query formulation, refinement and result analysis in a newspaper article archive environment. Topical, temporal and metadata dimensions are presented in a single consistent environment based on multiple synchronized visual components, enabling users to answer complex questions without explicitly reformulating the query. A significant enhancement in terms of speed and user satisfaction is expected, and will be evaluated in user tests currently in progress.

8. Acknowledgements

The Know-Center is a Competence Center funded within the Austrian Competence Centers Programme K plus under the auspices of the Austrian Ministry of Transport, Innovation and Technology (www.kplus.at).

We also want to thank Waltraud Wiedermann, Martina Rathbauer, Manfred Mitterholzer, and Christian Piccardi, all of the Austrain Press Agency, for their feedback and efforts.

9. References

[1] E. Rennison, “Galaxy of news: An approach to visualizing and understanding expansive news landscapes”, *Proceedings of UIST 1994*, ACM, California, USA, 1994.

[2] A. Leuski and J. Allan, “Lighthouse: Showing the way to relevant information”, *Proceedings of the IEEE Symposium on Information Vizualization 2002*, IEEE, USA, 2002.

[3] M. Granitzer, W.Kienreich, V.Sabol, K.Andrews, W.Klieber, “Evaluating a System for Interactive Exploration of Large, Hierarchically Structured Document Repositories”, *Proceedings of the IEEE Symposium on Information Vizualization 2004*, IEEE,Washington, DC, USA, 2004.

[4] Austria Presse Agentur, Vienna. http://www.apa.at

[5] D.R.Cutting, D.R.Karger, J.O. Pedersen, J.W. Tukey, Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Copenhagen, Denmark, 1992

[6] G. Cao, D. Song and P.D. Bruza, “Suffix Tree Clustering on Post-retrieval Documents”, *DSTC Technical Report*, Distributed Systems Technology Centre, Australia, 2003.

[7] W. Daelemans, “Memory-Based Shallow Parsing for Text Mining”, *Learning Methods for Text Understanding and Mining*, Grenoble, France, 2004.

[8] M. Chalmers, “A linear iteration time layout algorithm for visualising high-dimensional data”, *Proceedings of Visualization 96*, IEEE, California, USA, 1996.

[9] A. Spoerri, “Coordinated Views and Tight Coupling to Support Meta Searching”, *Proceedings of CMV, Second International Conference on Coordinated and Multiple Views in Exploratory Visualization*, London, England, 2004

[10] M. Carey, D.C. Heesch and S.M.Rueger, “Info Navigator: A Visualization Tool for Document Searching and Browsing”, *Proceedings of International Conference on Distributed Multimedia Systems*, Florida, USA, 2003.

[11] S. Havre, B. Hetzler, and L.Nowell, “ThemeRiver: Visualizing Theme Changes over Time”, *Proceedings of the IEEE Symposium on Information Vizualization 2000*, IEEE Computer Society, Washington, DC, USA, 2000.

[12] V. Sabol, “Visualisation islands: Interactive visualisation and clustering of search results sets”, *Master’s Thesis at IICM*, Graz University of Technology, Graz, Austria, 2001.