"'*Andreas, Rauber*'? Conference Pages Are over There, German Documents on the Lower Left,.." - An "Old-Fashioned" Approach to Web Search Results Visualization

Andreas Rauber and Harald Bina
Department of Software Technology, Vienna University of Technology
Favoritenstr. 9 - 11 /188, A-1040 Vienna, Austria
www.ifs.tuwien.ac.at/~andi www.ifs.tuwien.ac.at/~harry

Abstract

With the massive advance of electronic document repositories, usable interfaces to these repositories gain importance. While sophisticated information retrieval techniques provide acceptable result selection, interaction with a repository as such and with the documents retrieved leaves a lot to be desired, as users commonly are restricted to reading through one-dimensionally sorted lists of document descriptions.

In this paper we present the combination of a topical organization of documents via the SOMLib *system with the intuitive representation of document metadata provided by the* libViewer *visualization. It allows users to approach query results in a way comparable to how conventional libraries are being used. We demonstrate the benefits of this system by using it as a front-end to the AltaVista search engine.*

Keywords: *Library Representation, Information Space Metaphors, Document Classification*

1 Introduction

With the increasing amount of information available electronically, methods for searching these vast information repositories gain importance. While information that is available in a highly structured form can be searched quite efficiently by querying databases, retrieving information from rather unstructured full-text databases requires entirly different approaches to query processing and result representation. Research in information retrieval (IR) is providing us with ever better methods for retrieving the most relevant documents for a specific query. However, contrary to highly specific, structured queries against databases, full text queries cannot rely on a specific relevance ranking criteria. In a typical information retrieval setting using a (rather small) set of keywords to retrieve documents on a specific topic, a huge number of documents may satisfy the specified query criteria. Yet, the retrieved documents may cover a rather broad range of topics because the small number of keywords provided usually does not define the entire content. We may thus need to analyze the retrieved documents in order to characterize the various topics retrieved from a repository to allow the user to select the documents most relevant to her or him. Compared to a conventional IR-setting, i.e. a conventional library, we find exactly this situation: asked for books about a specific topic, such as 'computer graphics', librarians usually will not leave you with a pile of 50 books, stating that the books on the top are the best, and the ones on the bottom are the least important with respect to computer graphics. Rather, they will provide you with several piles, stating that some books deal with virtual reality and animation or others cover the concepts of information visualization and so on.

We thus find topical organization one of the key concepts in an IR setting, which is being used in a number of projects such as the Information Visualization Project at Xerox PARC [9], where information is depicted in a 3-dimensional space with the focus being on the amount of information being visible at one time and an easily understandable way of moving through large information spaces. A different approach for visualizing the contents of texts is presented in [10], where the main concepts of a text are used to span a multidimensional shape which is rendered to form 3-dimensional shapes, allowing the detection of documents on similar topics as documents exhibiting a similar shape. A second problem users encounter when interacting with IR systems, is understanding the returned documents. A lot of information is usually presented to the user, such as the size of the document, its date of creation, its location in terms of a URL or repository section. These concepts are provided in a textual and, even worse, in a rather technical form, listing document sizes in kilobytes, locations in form of incomprehensible domain listings, forcing the user to read and abstract from the provided textual information. Comparing this again to a more conventional setting, we find most of this information provided in a rather intuitively interpretable way. Books have entirely different sizes, allowing us to estimate the amount of information provided by them without having to think about the number of pages they are made up of. The age of the book can be easily told by taking a look at it's binding, and information like publishers' logos are interpreted without specifically thinking about it. The actual representation of a document thus allows us to tell a lot about its contents without having to read any description.

This problem of graphical document representation has been analyzed in several projects with the focus ranging from real-world like representations, such as in the antiquarian Sar-

0-7695-0680-1/00 $10.00 © 2000 IEEE

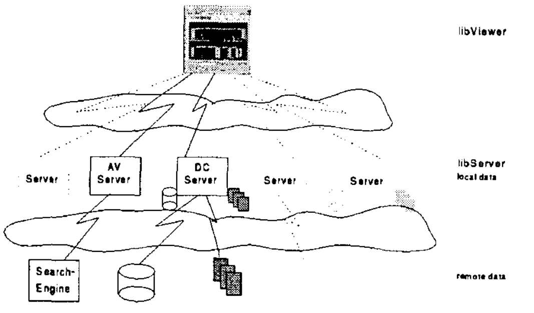


Figure 1. *libViewer* Architecture: Applet connecting to a number of servers

tiaux Collection [1], to the use of abstract metaphors to allow intuitive interaction with search mechanism and electronic documents. A map-based representation of documents in the spirit of multi-dimensional information visualization is provided by the Nemo project ([2]), showing the main attributes of a set of documents as icons of different color, patterns and text. An integrative approach to library representation covering both document visualization as well as interaction, is covered by the Bookhouse project [5], drawing the library as a storehouse and using various metaphors for interaction.

In this paper we present an approach to combining these two characteristics of "conventional" information retrieval for a rather modern setting. With the SOMLib digital library system a neural network, the self-organizing map (SOM), is used to organize documents into topical clusters. This approach of using the SOM to organize large document collections has shown to be succesful in a number of projects, c.f. [3, 7, 8]. The libViewer, a metaphor-graphics based representation of documents, is used to visualize meta-information on the documents in an intuitively interpretable way. We use the AltaVista search engine as a sample document repository to demonstrate the benefits of this approach.

The remainder of this paper is organized as follows: Section 2 provides an overview of and introduction to the various modules of the system, presenting document retrieval, content analysis and graphical representation. Section 3 then presents an example query to the AltaVista search engine and the graphical representation provided by our system, followed by an analysis and further research directions in Section 4. Some conclusions are given in Section 5.

2 Architecture

2.1 Overview

The presented system relies on a client-server based architecture for query processing, result analysis and representation. A conceptual view of this architecture is depicted in Figure 1. The libViewer [6] is a Java-Applet interfacing with a number of servers providing the data to be visualized.[1] Its main task is to provide a library representation that is intuitively understandable by the untrained user by relying on concepts and metaphors taken from conventional, real world libraries. It relies on a server to provide the appropriate mapping from the metadata attributes available in a specific document repository onto (a subset of) the supported metaphors.

A query is presented to the system via the libViewer interface, which forwards it to the selected server. This server either has the appropriate data available locally, or, as in the case of the AVServer, contacts a second server, such as e.g. the AltaVista search engine, providing the requested information. The retrieved documents are processed by the server to allow graphical representation. In the case of search results representation, a neural network may be trained to provide a content-based clustering of the retrieved documents, sorting the documents into different shelves. The resulting data description is then forwarded to the libViewer applet for graphical representation.

While the libViewer applet was designed as a generic interface to digital document repositories by relying on various servers to supply appropriate mappings, we will restrict our discussion in this paper to its interaction with the AVServer, presenting it as a front-end tool to the AltaVista search engine. For further information on its application in more general domains, please refer to [6].

2.2 Query processing and graphical result description

To use the libViewer as a front-end to AltaVista, the appropriate server, i.e. the AVServer, is selected from the list of available servers in the libViewer applet menue. A query is entered into the query field and sent to the server. The AVServer forwards the query to AltaVista, retrieving the list of the first 100 or 200 documents. Their descriptions are then mapped onto the according graphical metaphors supported by the libViewer. AltaVista currently returns the following meta-information:

- **Title:** Title of a html-page as specified by the <TITLE> </TITLE> mark-ups.
- **Size:** Size of the page in Kilobytes.
- **URL:** The location of the page on the World Wide Web.
- **Date of last modification:** Date when the page was last modified.
- **Description:** A short description of the page consisting of its first few words.
- **Language:** The language that the page is written in.
- **Ranking:** An implicit relevance information of the document with respect to the query by listing the highest ranking documents first.

The metaphors supported by the libViewer basically comprise different types of documents, such as hardcover and paperback books, binders, papers etc. These graphical representations of document types are modified according to the document descriptions, i.e. the size of the graphical representation is used to represent the size of the document, converting, e.g., the size of the document in terms of kilobytes into the width

[1] A prototype of the libviewer is available at www.ifs.tuwien.ac.at/~andi/somlib/libviewer.html for interactive exploration

title	text on spine
description	textual description
date	highlighting glare
language	color
topic section	shelf location
size	spine width
ranking	position within shelf
URL	link + domain logo

Table 1. Mapping of AltaVista metadata onto libViewer metaphors

of the spine. The language of a document as provided by AltaVista is represented by the color of the according document, painting, for example, all german-language documents with a yellow spine. Further information, such as the title of the document, is set to be printed on the spine of the book. While a logo on the back of the spine is conventionally used for representing publisher information, in the case of the AVServer we use it to convay domain information, drawing an appropriate flag on the spine if a national domain is encountered, or an appropriate logo for other domains such as the educational or commercial *.edu* and *.com* domains, respectively. The relevance, i.e. the position of the document within the returned result list, is represented by its position within the shelf, with less relevant documents being moved more to the back in the resulting graphical representation. The distribution of documents into the various shelves is provided by the content analysis performed by the neural network as is described in Section 2.3.

An overview of the mappings of metadata onto metaphors is provided in Table 2.2. Please note, that, due to the rather small amount of meta-information provided by AltaVista, only a very limited subset of the graphical metaphors provided by the libViewer is actually used. For a more detailed presentation of the set of available metaphors that may be used by servers, please refer to [6].

2.3 Content classification

While the graphical representation allows users to intuitively interpret the available data on a document, we need an automatic way to overcome the restricting linear ranked listing of retrieved documents. This is provided by a popular unsupervised neural network, namely the self-organizing map (SOM) [4], which provides a topology-preserving mapping from a high dimensional feature space onto a usually two-dimensional output space.Documents can be thought of as forming topical clusters in the high-dimensional feature space spanned by the words that the documents are made up of. During the training process the SOM organizes the documents on the map space in such a way, that documents that are close to each other in the feature space, and thus cover similar topics, are also mapped onto neighboring regions on the map. After the AVServer received the list of result hits from AltaVista, a download process is started, trying to retrieve all documents in parallel. In order to keep the response time of the system small, no retries are performed, i.e. documents that could not be retrieved at first contact are ignored.

We then use full-text indexing to represent the various documents according to the vector space model of information retrieval. The terms are roughly stemmed and weighted according to a $tf \times idf$, i.e. term frequency × inverse document frequency, weighting scheme, which assigns high values to terms that are considered important in describing the contents of a document. These feature vectors are then used to train a self-organizing map. To be able to provide acceptable response times, we developed the parSOM [11], a parallel implementation of the self-organizing map, allowing fast classification of large document collections. This way classification can be achieved in less than a minute, making the download times rather than the classification times a bottleneck in the system.

As a result of the SOM training process each document is assigned a location in a grid according to its content. This information is used to assign each document an appropriate shelf by setting the corresponding location metaphor as listed in Table 2.2.We furthermore use the LabelSOM technique [8] to extract keywords that turn out to be most descriptive for a cluster, which are then depicted as shelf labels.

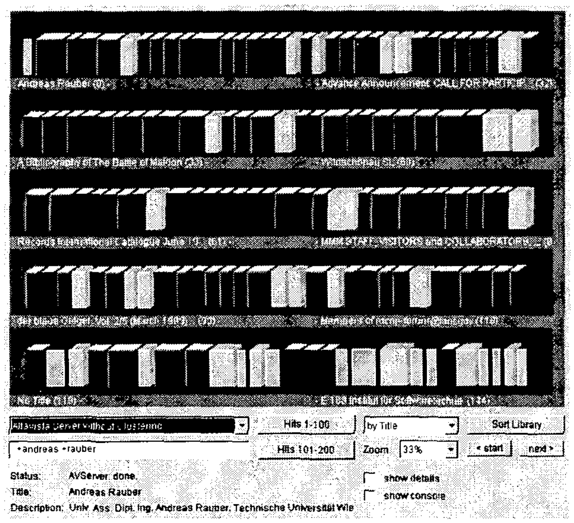
2.4 Result representation

The resulting metaphor graphics description is then transferred to the libViewer applet, which draws the documents accordingly. The resulting library-like representation of documents can then be viewed from both a distant as well as a close-up perspective, with the distant representation providing an overview of the various documents, and the close-up representation revealing more graphical details. The documents in the shelves can be sorted according to various criteria such as size, domain etc. While the graphical representation itself provides the most important metadata, additional information such as the full title of a document and the textual description provided by AltaVista is listed in the libViewer window as the mouse pointer is moved across the spine of the books. A details window lists the complete set of metadata information. Clicking on the spine opens a separate browser window with the according URL.

3 Experiments

In this section we present an example session of using the libViewer as a query interface to the AltaVista search engine. The libViewer applet is loaded into a standard graphical web browser and the AVServer with Clustering is selected as the server to be contacted. Next, a query, such as *'+andreas +rauber'* is entered in the query interface and sent to the AVServer, which forwards the query to AltaVista and retrieves the result pages.

If the AVServer without clustering were selected, the returned hits would immediately be converted into their graphical description and forwarded to the libViewer. Since no topical classification was yet performed, the libViewer simply lists the documents as a sequential list as provided by the



AltaVista Server without Clustering

+andreas +rauber

by Title

Zoom 33%

Status: AVServer: done.

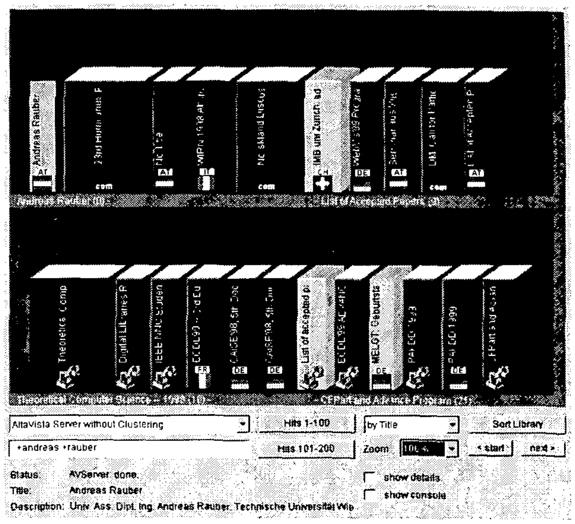
Title: Andreas Rauber

Description: Univ. Ass. Dipl. Ing. Andreas Rauber, Technische Universität Wie

□ show details

□ show console

Figure 2. Standard libViewer representation of AltaVista query results



AltaVista Server without Clustering

+andreas +rauber

by Title

Zoom 100%

Status: AVServer: done.

Title: Andreas Rauber

Description: Univ. Ass. Dipl. Ing. Andreas Rauber, Technische Universität Wie

□ show details

□ show console

Figure 3. Close-up libViewer representation of AltaVista query result without topic classification

search engine. Figure 2 depicts the resulting graphical representation of the search results from the distance view. Please note, that in the graphical representation a set of 144 documents is shown at one glance, with the remaining documents being revealed by scrolling through the library.

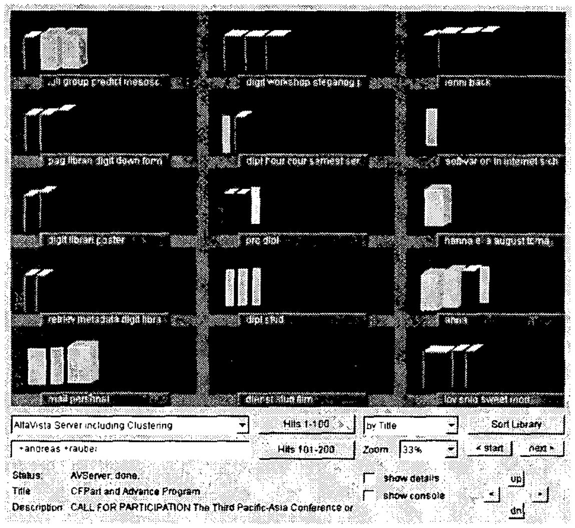
Even at this small-scale representation, a lot of information can be discerned, such as the existence of a number of (yellow) german language books amidst many (blue) english documents and a few (grey) documents for which no language tag was returned by AltaVista. Furthermore, we can easily discern large documents from small ones by their significantly broader spine. We also can identify new documents (i.e. less than a year old) by the high-lighting glare framing the spine, as for example on the 6th document in the first shelf. The titles of the documents and their descriptions are displayed in the libViewer window as the mouse moves across the spines of the books. Simultaneously, all information about a document is listed in a 'details window'. This feature facilitates very efficient browsing and analysis of the search results, as both a lot of information about a large number of documents is provided in a very condensed, but intuitively interpretable way, while at the same time providing all detailed information by pointing at the documents of interest.

Zooming into the library, more details are revealed as depicted in Figure 3. For every document we now find the title of the document printed on its spine, together with a logo indicating the domain that the document is located in. We find, for example, the first document entitled *Andreas Rauber* to be located in Austria, next to a rather large document on the *23 Hurricanes conference* from the *.com* domain. Other domains found are the national domains of Italy, Switzerland, Germany and France, as well as the Educational domain *.edu*. Overall, by providing a graphical representation of the search result, a much more intuitive representation and means of interaction can be provided by reliefing the user from having to read and interpret rather incomprehensable concepts such as size in terms of kilobyte, domain information, etc.

However, when looking at the documents returned by the search engine we find, for example, a large document about a hurricanes program, several conference sites on data mining, the homepage of IBM Zurich etc. A topically sorted list would be highly preferable. This requires the AltaVista Server with clustering to download all documents. The documents are parsed to form a vector representation and fed into a 7 × 7 SOM. As a result, each document is assigned a position within a grid, which can be directly translated to a bookshelf of 7 columns and rows. Returning this additional information to the libViewer allows it to represent the documents organized by topic as depicted in Figure 4.

We find that several identical documents are located on different servers and thus treated as different documents by the search engine. Using the content analysis those identical documents are obviously located in the same shelf. Apart from this obvious identification of identical pages, pages on related topics are grouped nicely together into the same shelves. For example, in the second shelf on the second row we find two documents from our department, whereas a number of documents on the ACM Digital Libraries conference are mapped onto the first shelves in rows 2 and 3.

While being rather crude identifiers and somewhat distorted by the stemming process, the labels on the shelves help identify certain subject areas. The shelves described above, for example, are labeled *digit(al), librari(es), poster* and *pag(e), librari(es), digit, down* for the two shelves on digital libraries, respectively, and *dipl, hour, cour(se), semest(er),*



+andreas +rauber

Status: AVServer: done.

Title CFPart and Advance Program

Description CALL FOR PARTICIPATION The Third Pacific-Asia Conference or

Figure 4. libViewer representation of AltaVista query result with topic classification

seminar, mittwoch, tjoa, miksch for shelf 2 in row 2, listing some keywords and members of our department as descriptions of the two documents in this shelf. Further topical clusters are e.g. the hurricanes conference pages on the shelf in column 4 and row 3, which are part of a larger cluster of conference pages in the upper right corner of the library including DEXA and IDEAL conference pages. Furthermore, we obviously find the yellow german language documents to be nicely separated in the bottom left area of the library from the blue english documents, with a few exceptions, which are either be bi-langual documents or documents that have an incorrect language tag assigned by the search engine.

4. Evaluation and further research directions

Current evaluation has shown, that the proposed system may provide a more intuitive approach to represent query results. While it is not intended as a standard interface to search engines, it might prove to be a helpful alternative for smaller or special purpose document repositories. Obviously, the download times required to obtain a local copy of each document for topic classification limit its applicability. However, basically the search engine providing the document description can also provide the topic classification, saving the user from having to download the documents as such.

Discussions revealed, that the 2-dimensional SOM, which organizes documents by topic on a two-dimensional grid, does not reflect the organization of documents in the real world, which are rather sorted on a 1-dimensional shelf, which is only organized as several consecutive rows of shelves due to physical limitations. This situation could be addressed by using a 1-dimensional SOM rather than a 2-dimensional one for document organization. However, we have the impression, that the limited view represented by the screen actually makes the two-dimensional organization preferable by having shelves beneath each other represent similar topics rather than having to scan consecutive rows of shelves horizontally. However, these considerations definitely merit further analysis.

We are currently investigating a structural analysis of documents in addition to the topical analysis in order to be able to depict, say, pages containing lots of links or graphics using a different metaphor than for, say, pages with lots of text, utilizing, for example, the libViewer binder or paper metaphors. This should provide the user with some more feedback on the type of document to be expected before actually opening it.

5 Conclusion

We presented a system combining content based document analysis with metaphor graphical representation to provide a more intuitive interface to document repositories. Documents are grouped by contents into several shelves using a neural network. Rather than providing textual descriptions of the documents, graphical real-world like metaphors are used to convey this information. This makes the meta-information available on the documents intuitively graspable and interpretable. Concepts such as the size of a document in kilobytes or the location of documents in terms of domain information can thus be conveyed to a non-expert audience.

We demonstrated the benefits of this system using it as a front-end to the AltaVista search engine.

With no significant additional processing time required, the graphical depiction of the result list provides the user with a much more intuitive representation of the documents in a rather condensed space.

References

[1] P. Cubaud, C. Thiria, and A. Topol. Experimenting a 3d interface for the access to a digital library. In *Proc. ACM Conf. on Digital Libraries (DL98)*, Pittsburgh, PA, 1998.

[2] M. Hascoët and X. Soinard. Using maps as a user interface to a digital library. In *Proc. Int'l ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 1998.

[3] S. Kaski, T. Honkela, K. Lagus, and T. Kohonen. WEBSOM—self-organizing maps of document collections. In *Elsevir Publ.* Elsevier Publications, 1997.

[4] T. Kohonen. *Self-Organizing Maps*. Springer Verlag, Berlin, Germany, 1995.

[5] A. Pejtersen. A library system for information retrieval based on cognitive task analysis and supported by an icon-based interface. In *Proc. of the ACM SIGIR Conference on Research and Developement in Information Retrieval (SIGIR'89)*, 1989.

[6] A. Rauber and H. Bina. A metaphor graphics based representation of digital libraries on the World Wide Web: Using the libViewer to make metadata visible. In *Proc. DEXA-Workshop on Web-based Information Visualization (WebVis99)*, Florence, Italy, 1999.

[7] A. Rauber and D. Merkl. The SOMLib Digital Library System. In *Proc. Europ. Conf. on Research and Advanced Technology for Digital Libraries (ECDL99)*, Paris, France, 1999. LNCS, Springer Verlag.

[8] A. Rauber and D. Merkl. Using self-organizing maps to organize document collections and to characterize subject matters: How to make a map tell the news of the world. In *Proc. 10th Int'l Conf. on Database and Expert Systems Applications (DEXA99)*, Florence, Italy, 1999.

[9] G. Robertson, S. Card, and J. Mackinlay. Information visualization using 3d interactive animation. *Communications of the ACM*, 36:57 – 71, April 1993.

[10] R. Rohrer, D. Ebert, and J. Sibert. The shape of shakespeare: Visualizing text using implicit surfaces. In *IEEE Symposium on Information Visualization (INFOVIS'98)*, North Carolina, 1998.

[11] P. Tomsich, A. Rauber, and D. Merkl. parSOM: Using parallelism to overcome memory latency in self-organizing neural networks. In *Proc. of the European Conf. on High-Performance Computing (HPCN00)*, 2000.