Adding SOMLib capabilities to the Greenstone Digital Library System

Rudolf Mayer, Andreas Rauber

Institute of Software Technology and Interactive Systems
Vienna University of Technology, Vienna, Austria
mayer@ifs.tuwien.ac.at, rauber@ifs.tuwien.ac.at,
http://www.ifs.tuwien.ac.at/~[mayer,andi]

Abstract. Many conventional digital library systems offer access to their collections only via full text or meta-data search, or by browsing-access via a hierarchy of categories. With the increasing amount of digital content available, alternative methods to access the content seem necessary. The SOMLib system, which is based on using Self-Organizing Maps (SOMs), has been used to automatically organize documents of a digital library by their content. In this paper, we present an integration of this system into the popular open-source digital library system Greenstone, combining searching and explorative browsing through the thematically organized content using the map. We present the system on a demo collection consisting of the abstracts of papers and posters from the last 5 years from the JCDL, ECDL and ICADL conferences.

1 Introduction

Digital library systems provide a uniform way to organize, maintain and access collections of digital objects, may they be text, images, audio, or others. However, many conventional digital library systems have shortcomings on providing the user with ways to access and retrieve their content. This is due to many of the systems offering access to their collections only via full text or meta-data search or browsing-access via simple ordered lists or a hierarchy of categories.

The SOMLib system [1], which is based on using Self-Organizing Maps (SOMs) [2], has been successfully used to automatically organize documents of a digital library by their content. The user can explore the thus generated map, as she is used to exploring a conventional geographical map. However, this approach can only be understood as an addition to traditional ways of retrieving the information. Therefore, integrating both traditional ways like full text or meta-data search and explorative search via the SOMLib map is necessary.

In this paper, we present a system that integrates SOMLib features into the popular digital library system *Greenstone*. The system thus created combines full text or meta-data search with explorative browsing through the thematically organized content. This is achieved by visualizing the search results on the map, and by including map-selections into the list of results. This way, users can easily find documents which are similar in content to their search result and may therefore be relevant to them, but were not retrieved by conventional search methods. With an advanced SOMViewer interface, additional means of interaction become possible.

The remainder of this paper is organized as follows. In Section 2 we will give an introduction into retrieval capabilities offered by conventional digital library systems, based on the example of Greenstone. Section 3 shortly describes the Self-Organizing Map and the SOMLib digital library system, and presents the integration of SOMLib and Greenstone. We give conclusions and an outlook on future work in Section 4.

2 The Greenstone Digital Library System

Greenstone is a popular open-source digital library system for constructing comprehensive document collections [3]. As other systems, Greenstone supports the generation of indices of various kinds of media (e.g. text, images, audio and video). Documents of any kind can further be described by a set of additional meta-data information. Greenstone offers two widely used ways to locate documents within the collection: full text and meta-data based *search*, and *browsing*.

Search: The full text and meta-data search allows for proximity search and exact phrase search. Additionally, a fielded search is provided, which allows the user to combine searching on different indices at the same time. Although a very powerful tool, especially compared to conventional libraries, searching inherits a few problems: First, broadly formulated queries may lead to huge result lists. Although the documents the user is looking for will probably be part of the result list, it is quite unlikely that she will look through the whole list to find them. Another problem is known as the vocabulary problem: the same object or action may be named differently by different people, resulting in a low recall when query terms differ from the terms used in the documents.

Browsing: Browsing offers an (ordered) list of the documents in the collection, built on meta-data. Greenstone offers browsing by providing a simple scrollable list of the documents, ordered according to a given meta-data field. Moreover, Greenstone offers a hierarchical classication, which allows to define an arbitrary number of levels of hierarchies. However, the meta-data hierarchy has to be defined separately by the user. The quality of the meta-data is crucial for the meta-data based searching and browsing. If some meta-data is not available for a document, the document can not be easily found through that search or browsing list. Another problem is inconsistency, for example two documents having the same author might have different meta-data due to different spellings of the name, or different ordering of name and surname.

Some of the mentioned disadvantages can be solved by the proposed integration of the SOM into a digital library system.

3 A Self-Organizing Map for Greenstone

Self-Organizing Maps and SOMLib

The Self-Organizing Map [2] is a well known and widely used neural network model based on unsupervised learning. The SOM provides a mapping from a high-dimensional input space to a lower dimensional output space. In this mapping, the SOM preserves the topology of the input space, i.e. input patterns that are located close to each other in the input space will also be located closely in the output space, while dissimilar patterns will be mapped on to opposite map regions. In our application, the input space is formed by a vector-space representation of the documents of the digital library collection. The features of the vector are selected according to their document frequency, and the weights are computed using a standard $tf \times idf$ weighting scheme.

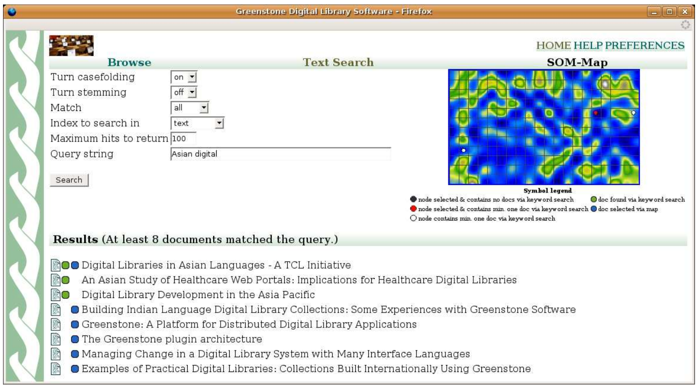
The SOMLib digital library system [1] creates maps of a document collection. A rich desktop client application allows interactive exploration of the data space by zooming into and selecting areas of documents. The system can organize any kind of objects that can be represented numerically by feature vectors, such as text, images and music [4].

Greenstone Integration

Greenstone offers in its current version 3 an open architecture that allows developers to provide additional services to a Greenstone collection. We make use of this plug-in architecture to provide our own service as an extension, based on the existing query services. That way, the user can still use all the basic functionality provided by Greenstone. Additionally, she will be able to use the wealth of additional information the SOM mapping provides about the documents matching the query results and the whole collection itself.

The map can be used in two different ways. First, results of the Greenstone search will be highlighted on the map: map nodes that contain at least one of the documents matched by the query are indicated by a white circle. With this visualization of the search results, the user can immediately see which documents have a topical similarity - these documents will all be located close to each other and form a cluster on the map. Additionally, outliers become visible as isolated spots on the map. Secondly, the user can explore the map - she can select nodes, upon which the documents lying on that nodes will be added to the result list of the Greenstone search. Documents that have been matched both by the map selection and the search result will be marked especially. The user can get additional information on the content of the collection by just moving over a node of the map with the mouse, upon which a pop-up will display terms that describe the documents on that node the best. This is achieved by utilising the LabelSOM algorithm [5].

Figure 3 depicts the standard search interface of Greenstone, extended by the SOM map on the top-right corner. The collection used in this example consists of the abstracts of the accepted papers and posters of the three major conferences on digital libraries during the last years: JCDL (2001-2005), ECDL (2001-2005), and ICADL (2002-2005). It contains 1051 documents. In the given example, the user has issued a query



Turn casefolding on
Turn stemming off
Match all
Index to search in text
Maximum hits to return 100
Query string Asian digital

Symbol legend
● node selected & contains no docs via keyword search
● node selected & contains min. one doc via keyword search
○ node contains min. one doc via keyword search
● doc found via keyword search
● doc selected via map

Results (At least 8 documents matched the query.)

- ●● Digital Libraries in Asian Languages - A TCL Initiative
- ● An Asian Study of Healthcare Web Portals: Implications for Healthcare Digital Libraries
- ● Digital Library Development in the Asia Pacific
- ● Building Indian Language Digital Library Collections: Some Experiences with Greenstone Software
- ● Greenstone: A Platform for Distributed Digital Library Applications
- ● The Greenstone plugin architecture
- ● Managing Change in a Digital Library System with Many Interface Languages
- ● Examples of Practical Digital Libraries: Collections Built Internationally Using Greenstone

Fig. 1. Enhancing the traditional Greenstone query search with a SOMLib map.

’Asian digital’, wanting to search for documents dealing with Asian digital libraries, and got 3 hits. However, manually inspecting the map on the search hits, the user however is possible to detect five more documents. Of those, one deals with an Indian digital library, and two others deal with international digital libraries and may be relevant for her.

4 Conclusion and future work

In this paper we presented an integration of a SOM map into a conventional digital library system such as Greenstone. This combination allows the user to use both traditional search and browsing with a wealth of new exciting possibilities to exploratively search in the digital library’s content. Future work will focus on increasing the level of interaction with the map, as it is already now possible in the desktop application (e.g. zooming functionality).

References

1. Rauber, A., Merkl, D.: The SOMLib digital library system. In: European Conference on Digital Libraries, Paris, France (1999) 323–342
2. Kohonen, T.: Self-Organizing Maps. Volume 30 of Springer Series in Information Sciences. Springer, Berlin, Heidelberg (1995)
3. Witten, I.H., Bainbridge, D., Boddie, S.J.: Greenstone: Open-source digital library software. D-Lib Magazine **7**(10) (2001)
4. Neumayer, R., Dittenbach, M., Rauber, A.: PlaySOM and PocketSOMPlayer: Alternative interfaces to large music collections. In: Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR 2005), London, UK (2005) 618–623
5. Rauber, A., Merkl, D.: Text mining in the SOMLib digital library system: The representation of topics and genres. Applied Intelligence **18**(3) (2003) 271–293