Knowledge Navigation in Networked Digital Libraries

Mike P. Papazoglou and Jeroen Hoppenbrouwers
Tilburg University/Infolab
PO Box 90133, NL-5000 LE Tilburg
The Netherlands
{mikep,hoppie}@kub.nl

Abstract. Formulating precise and effective queries in document retrieval systems requires the users to predict which terms appear in documents relevant to their information needs. It is important that users do not retrieve a plethora of irrelevant documents due to underspecified queries or queries containing ambiguous search terms. Due to these reasons, networked digital libraries with rapid growth in their volume of documents, document diversity, and terminological variations are becoming increasingly difficult to manage. In this paper we consider the concept of knowledge navigation for federated digital libraries and explain how it can provide the kind of intermediary expert prompting required to enable purposeful searching and effective discovery of documents.

Keywords: digital library, meta-data, ontology, clustering, browsing, navigation, semantic indexing, concept searching.

1 Introduction

Digital libraries bring large volumes of information to the user, whether researcher, analyst, student or casual browser. The classical approach by Information Retrieval (IR) is to define scalable techniques such as the vector-space model for matching queries against many thousands of documents efficiently [21]. This technique attempts to maximize the relevance of a document to a query. All approaches have also been similarly intensiooned although focusing on applying domain knowledge and analogical reasoning rather than numeric matching techniques. For example, an analogical reasoning system can be used to construct the possible interpretations of query terms corresponding to alternative paths in the inference network and to negotiate them with the user. In this way the user is able to select his/her intended interpretation of an unstructured query.

The relevance of query terms to documents is only one part of a complex problem. Currently, there is a massive investment world-wide in making digital document repositories accessible over networks. The result of this is that users

---

* This research has been partially funded by the European Union under the Telematics project Decomate LIB-5672/B.

© Springer-Verlag Berlin Heidelberg 1999
of Digital Libraries (DLs) are overwhelmed by the amount of documents that are required to assimilate but also of the constant influx of new information. At the same time there is also a major investment in providing indexing, categorization, and other forms of meta-data for DL documents and a large number of IR techniques have been developed for automatic categorization of repositories for which human indexing is unavailable. These activities result in quite diverse meta-data vocabularies, e.g., index and thesaurus terms, that characterize documents. Therefore, the number of meta-data vocabularies that are accessible but unfamiliar for any individual searcher is increasing steeply.

1.1 Limitations of Index Terms

Despite user knowledge that several terms within a particular domain may have the same meaning, known IR technology can only match terms provided by the searcher to terms literally occurring in documents or indexing records in the collection. Unfortunately, keyword expansion techniques have shown no significant improvements over other standard IR techniques as it is usually very difficult to choose which keywords to expand [3]. This implies that there are too many potentially matching documents which may not be retrieved due to the variation of the index terms used, and the fluidity of concepts and vocabularies in different domains.

The situation described above is particularly acute in digital libraries with spatial distribution which aim to make widely distributed collections of heterogeneous documents appear to be a single (virtually) integrated collection. Such federated digital libraries (FDLs) typically specialize in a fairly narrow and specific domain area, e.g., Biomedicine, Computer Science, or Economics. Although the amount of searching in FDLs is expected to rise, diminishing search effectiveness and less reliable answers is the predictable result as a consequence of the explosive increase in meta-data heterogeneity due to terminology fluctuations. The challenge is to provide automatically the kind of expert assistance that a human search intermediary, familiar with the source being searched, would provide. In [3] has been argued that the most effective solution to improving effectiveness in the search of digital repositories would be technology to assist the information searcher in coping with unfamiliar meta-data vocabularies.

1.2 From Terms to Knowledge

A particularly promising methodology for addressing these objectives is knowledge navigation. This methodology relies on the use of computer assisted support for acquiring and relating digital information originating from diverse heterogeneous document repositories. Knowledge navigation combines techniques from knowledge representation and natural language processing with classical techniques for indexing words and phrases in text to enable a retrieval system to make connections between the terminology of a user request and related terminology in the information provided in an FDL.