Resource Selection and Data Fusion in Multimedia Distributed Digital Libraries

Jamie Callan[1], Fabio Crestani[2], Henrik Nottelmann[3], Pietro Pala[4], Xiao Mang Shou[5]
[1]School of Computer Studies, Carnegie Mellon University, Pittsburgh, PA, USA
[2]Dept. Computer and Information Sciences, University of Strathclyde, Glasgow, UK
[3]Institute of Informatics and Interactive Systems, University of Duisburg-Essen, Duisburg, Germany
[4] Dip. Sistemi e Informatica, Università degli Studi di Firenze, Firenze, Italy
[5]Dept. of Information Studies, University of Sheffield, Sheffield, UK

callan@cs.cmu.edu, f.crestani@cis.strath.ac.uk, nottelmann@uni-duisburg.de, pala@dsi.unifi.it, x.m.shou@shef.ac.uk

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software

General Terms

Algorithms

Keywords

Networked Retrieval, Data Fusion, Resource Selection

1. THE MIND PROJECT

MIND is a EU funded project that addresses some of the issues that arise when people have routine access to a large number (possibly thousands) of heterogeneous and distributed multimedia Digital Libraries (DLs) over the Internet and the Web. When so many DLs are available, the first information access task is resource selection. This is predominantly an ineffective manual task as users are unaware of the contents of each individual library in terms of quantity, quality, information type, provenance and likely relevance. People need accurate automatic resource selection tools to assist them in this task, but resource selection requires accurate resource descriptions. The acquisition of resource descriptions is already a difficult task when dealing with cooperative DLs. It is much more difficult when dealing with non-cooperative DLs; in this case document sampling techniques need to be used. Once the Digital Libraries have been selected and the query forwarded to them, the results returned have to be merged by a process called data fusion, such that the overall retrieval quality is maximised. Todays Digital Libraries are rather heterogeneous with regards to schemas, retrieval capabilities and indexing methods, which makes both the data fusion and the following presentation of results particularly difficult. Research carried out in MIND attempts to addresses all these issues.

A schematic view of the architecture of the first MIND prototype is presented in figure 1. In syntheses, user queries issued via one of a number of available user interfaces (UIs) are passed to the dispatcher that selects the most appropriate DLs to which the query should be forwarded. The resource selection is based on information contained in proxies (one proxy for each DL). Each proxy translates the user query in the DL proprietary query language and schema. The results obtained from the different DLs are passed to a data fusing module which produces a single ranking of the documents to be presented to the users. In the rest of this paper we will briefly show how the MIND system works.

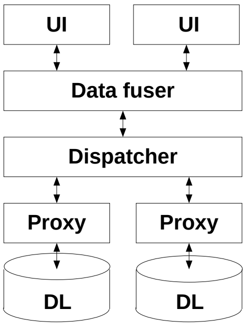

Figure 1: MIND architecture.

This work is supported by the IST project MIND (IST-2000-26061). More information about MIND can be found at `http://www.mind-project.org`.

Resource Description

Resource description are acquired at the proxy level using query-based sampling, a technique that does not require the cooperation of resource providers nor does it require that resource providers use a particular search engine or representation technique. The full details of this technique are presented in [3]. Each proxy has a specific policy of updating the resource description of its DL, so that resource description are kept up-to-date. Proxies are completely independent in this task.

Resource Selection

As it is too expensive to query all accessible libraries, resource selection is the task to find the DLs which are most-relevant to a user query. Current approaches rank libraries

SIGIR'03, July 28–August 1, 2003, Toronto, Canada.
ACM 1-58113-646-3/03/0007.

w.r.t. their similarity to the query and retrieve the same number of documents from a fixed number of top-ranked DLs. We follow a different approach on a better theoretic foundation: MIND's dispatcher computes a selection (number s_i of documents to be retrieved by DL_i) which minimises the overall costs (including retrieval quality, time, money) of the distributed retrieval [1]. We developed different methods for estimating retrieval quality (number of relevant documents $r_i(q, s_i)$ when retrieving s_i documents). In all methods, a probabilistic retrieval model is used for computing document scores, and these scores are then mapped onto the probability of relevance with a logistic function (which is a better approximation than the linear one used in former publications). The best performing method for estimating retrieval quality assumes that the indexing weights follow a normal distribution, leading to a normal distribution for the document scores. For non-text data, e.g. facts and images, this does not work. In contrast, a recall-precision function can be for mapping the average score onto the expected the number $r_i(q, s_i)$ of relevant documents in the result set. This technique can also be used for text. Our experiments showed that the MIND resource selection approach improves effectiveness compared with the state-of-the-art resource ranking method CORI.

Data Fusion

In MIND, data fusion combines the results from multiple search engines into a unified list. Existing merge methods rely on collection overlap or score information being returned by search engines. However in MIND, we do not rely on such information being present. Two fusion methods were tested: local headline search, and cross rank similarity comparison approximating document overlap by measuring the similarity of documents across the source rankings to be merged. The approach places documents higher in the fused ranking if they are similar to each other. The experiment results on the short topic category-A adhoc part of TREC-5 showed respectively a 2% and 41% improvement using cross rank similarity comparison and local headline search over round robin method [2]. Other methods such as local full text search, mixture of cross rank and local headline search were also tried. In general, the more text used, the better fusion became. Among all the methods, local search on retrieved headlines was found to be the most cost effective and efficient method and was employed for fusing text results in MIND. Conversely, fusion of results returned by image libraries relies on linear regression to learn and approximate normalisation coefficients used to normalise document scores.

Presentation of Results

Access to digital libraries in a heterogeneous, distributed, multimedia scenario, exacerbates the need for coping with different users, different tasks and different document representations. This might require different ways of presenting retrieval results, ways that support effective exploration of retrieved documents by enabling simultaneous display of multiple document properties as well as relationships between documents. In fact, in this scenario there are several potentially relevant characteristics of documents the user could be interested in, including document topic, source digital library, document medium, cost and presence of duplicates.

When considering methods for data fusion of heterogeneous collections, it is important to consider how data retrieved from multimedia collections should be presented in the UI. A study of existing multimedia retrieval systems was conducted. Users were interviewed about their attitudes to existing systems. Information gathered informed the design of the MIND interface. It was prototyped using low fidelity methods where user reactions to the interface were recorded. A conclusion of the study was that users do not wish to have search results across media fused into a single list. Therefore, the interface of MIND presents text, image and audio documents in separate columns rather than mixed into a single list. For textual documents, local headline search fusion was applied to the results, but other approach have also been explored.

To complement presentation of retrieved documents through multiple, single-media ranked lists, retrieved documents can also be presented in a graphic 2D Information Display Space (IDS). In this IDS, particularly powerful for images, each retrieved document is represented through a visual object (i.e. an icon) whose visual features are used to encode relevant information about the document. In so doing, visual features such as the position, size, shape and colour of a document icon are used to encode document properties. Values of visual features for each document icon should be chosen so as to associate the same feature to documents sharing some property (related documents) and different features to documents not sharing any property (unrelated documents). Thus, computation of the optimum combination of icon visual features is equivalent to the minimisation of a compound cost function accounting both for the visual similarity of related documents and for the visual dissimilarity of unrelated documents.

In addition, textual documents can be presented in a topical hierarchical structure and browsed in a "scatter-gather" way in order to identify the different interpretations of the information need expressed in the query that different DLs produced. This way of presenting documents is also very useful to identify document duplicates, which could not be picked up by the data fusion. Despite the highly heterogenous document representations, the clyster hypothesis at the basis of the hierarchical clustering appears to be quite robust.

Evaluation

We are currently evaluating the first MIND prototype system using a user-oriented and task-based evaluation methodology. Three evaluation scenarios involving different user groups, tasks and DLs are considered in the domain of news, arts and computer science.

2. REFERENCES

[1] N. Fuhr. A decision-theoretic approach to database selection in networked IR. *ACM TOIS*, 17(3):229–249, 1999.

[2] X.M. Shou and M.Sanderson. Experiments on Data Fusion Using Headline Information. *Proceedings ACM SIGIR'02*, p. 413-414, Tampere, Finland, August 2002;

[3] J.P. Callan and M.E. Connell. Query-based sampling of text databases. *ACM TOIS*, 19(2):97–130, 2001.