

Mediated access to very large document collections

David J. Harper
School of Computing
The Robert Gordon University
Aberdeen, Scotland, UK
djh@scms.rgu.ac.uk

Gheorghe Muresan
Department of Library and Information Sciences
Rutgers, The State University of New Jersey
New Brunswick, NJ 08901 USA
muresan@scils.rutgers.edu

ABSTRACT

Mediation based on structured specialised collections is proposed as an approach to supporting the exploration of very large document collections such as the Web. The main techniques employed are statistical language modelling and document clustering. The paper discusses the new interaction model and presents experimental results obtained with a prototypical system. References to papers which describe our work in more detail, discuss potential applications and propose future work are also included.

Keywords

Interactive information retrieval, Topic Modelling, Document clustering

1. INTRODUCTION

Various studies comparing mediated and un-mediated searches indicate that assistance given to users in their query formulation is a crucial factor in retrieval success. Unassisted users often have problems in formulating good queries due to a lack of sufficient knowledge of the appropriate vocabulary, an inability to use advanced query language syntax, or a lack of clear understanding of the system's conceptual model. Moreover, for exploratory searches, the users have the additional problem that they do not quite know what they are looking for and how they should go about resolving their *anomalous state of knowledge*.

A common solution to this problem is providing the users with interactive user interfaces based on visualization tools, through which the target collection can be explored. However, this approach is not practical for document collections that are too large to be structured, or for those that lack a semantic, topical structure. In such cases query-based searching is the sole means of access, so providing good queries is essential.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

JCDL '02 Portland, Oregon, USA

Copyright 2002 ACM X-XXXXX-XX-X/XX/XX ...\$5.00.

We propose a system that emulates the role of the human mediator by assisting the user in exploring a problem domain, in refining an information need, and in generating high quality queries for searching the target collection.

2. THE INTERACTION MODEL

We rely on the existence and availability of specialised collections of documents or abstracts maintained by various companies and organisations. These collections are representative for their domain and their structure conveys the topical structure of the domain. Some of these collections are already classified, either manually or automatically; for the others, we propose the use of document clustering in order to reveal the semantic structure of the domain represented.

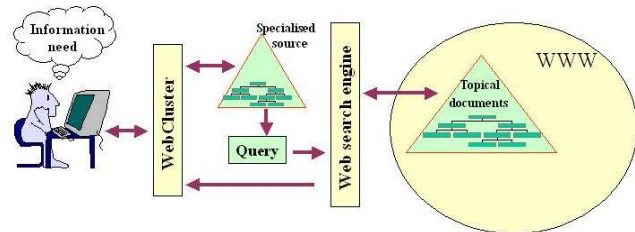


Figure 1: Mediating access to the World Wide Web.

Mediated retrieval, depicted in **Figure 1**, is a two-stage process. In the first stage, the searcher is offered visualization tools and the use of a combination of retrieval strategies for exploring a specialised 'source collection'. By browsing the structure of the collection, the user can learn the topics and the terminology of the problem domain. If more or less familiar with the domain, the user can also employ query-based searching in order to identify 'pockets' of relevant documents, and starting points for browsing[3].

Based on the user's exploration of the source collection, and on her selection of relevant 'exemplary documents', the system builds a (*statistical language*) model of the topic investigated. It can then act as a *mediator* by possibly eliciting more information for context disambiguation, and by generating a query that comprehensively, clearly and precisely reflects the contents of the documents selected by the user. This *mediated query* can then be used to extend the search to any 'target collections' that are heterogeneous, unstructured and too large to readily afford exploration strategies

other than query-based searching.

Mediation through the right ‘source collection’ has the potential to generate a very precise query and to significantly increase the quality of the retrieval effectiveness and the perceived completeness of the user’s task. If used over multiple search sessions, the mediation system has the potential to build user models, to learn the users’ interests, and to disambiguate and expand future queries.

3. TOPIC MODELS

The conceptual model of mediation access relies on the user making relevance judgements with regards to documents and clusters of the source collection, and thus conveying her topic of interest. In effect, the user provides an *exemplary* representation of the topic of interest, which consists of documents and clusters of documents that are typical for the topic investigated. The system performs a statistical analysis of the exemplary documents and derives a *statistical* or *language model* representation of the topic. Based on the context (the source collection and the target collection), the system derives a *keyword* representation of the topic, which consists of the terms that discriminate the topic in the given context, ranked based on their contribution to the topic. Finally, the mediated query, used for searching the target collection, is derived from this keyword representation of the topic, typically by applying a threshold on query size or term weight.

For computing term weights and consequently generating the keyword representation of a topic we use the *Kullback-Liebler (KL)* formula, or *relative entropy*, which offers several advantages. Firstly, it allows documents and clusters to be treated similarly, as bags of terms, so just one unified model is sufficient. Secondly, it offers a balance between *accuracy* in representation and *power of discrimination*. Thirdly, it allows the integration of context in the formulae, by highlighting the terms that distinguish a cluster from its context (details in [2]).

4. EXPERIMENTAL RESULTS

Successful mediation is based on several assumptions that we tested experimentally [1, 4]:

1. **System-based mediation is a usable paradigm.** In informal experiments, users found the system easy to use and useful. Especially when unfamiliar with the domain of the assigned retrieval task, the users felt that the terms suggested by the system helped them improve the quality of their queries and consequently the effectiveness of Web searching.
2. **Topics are identifiable in specialised source collection.** Informal user experiments and simulations of user searches provide experimental evidence that browsing and searching cluster representatives are effective in identifying the clusters most relevant for a topic.
3. **The mediated queries can support effective search of the target collection.** In simulations of various mediation strategies, the mediated queries proved significantly better than the queries derived from topic descriptions.

The explanation for our choice of search simulations, rather than real user experiments, is that the interactive mediation model is quite flexible and can be ‘enacted’ by a variety of user search strategies. A user experiment would have contained too many independent variables to be practical. Our approach was to simulate and compare various search strategies and other parameters such as weighting schemes or clustering methods and to find the combinations of parameters that looked most promising in terms of retrieval effectiveness. These optimal combinations will be used in future user experiments.

5. CONCLUSIONS

System-based mediation attempts to simulate the human mediation: the system interacts with the user and learns the user’s topic of interest by observing the user’s actions and selection of exemplary documents. Subsequently, the system tries to satisfy the user’s information need. The system helps the user both to gain a better understanding of the problem domain and to formulate high-quality queries.

Our simulations have shown that mediated access through a structured source collection has potential to improve the user’s query and to increase retrieval effectiveness. Such a solution is badly needed today, with Web searching tools widely available to people not trained in how to search.

Due to the small scale of our experiments, the results cannot be safely generalised. However, these results, the conclusions drawn from them and the ideas generated, can be used as a starting point in larger scale experiments, with a larger number of source collections, of different sizes and levels of heterogeneity, and a much larger number of test topics.

More details, as well as proposed applications and future work are described in [4].

6. REFERENCES

- [1] D. J. Harper, M. Mechkour, and G. Muresan. Document clustering for mediated information access. In *Proceedings of the 21st Annual BCS-IRSG Colloquium*, Glasgow, April 1999.
- [2] G. Muresan and D. J. Harper. Document clustering and language models for system-mediated information access. In P. Constantopoulos and I. T. Solvberg, editors, *Proceedings of ECDL’01*, pages 438–449, Darmstadt, Germany, September 2001. Springer. ISBN 3-540-42537-3.
- [3] G. Muresan, D. J. H Harper, A. Goker, and P. Lowit. ClusterBook, a tool for dual information access. In N. J. Belkin, P. Ingwersen, and M.-K. Leong, editors, *Proceedings of SIGIR 2000*, page 391, Athens, July 2000. ACM.
- [4] Gheorghe Muresan. *Using Document Clustering and Language Modelling in Mediated Information Retrieval*. PhD thesis, School of Computing, Robert Gordon University, Aberdeen, Scotland, United Kingdom, January 2002.