

WebCluster, a Tool for Mediated Information Access

Gheorghe Muresan, David J. Harper and Mourad Mechkour

School of Computer and Mathematical Sciences

The Robert Gordon University, Aberdeen

{gm,djh,mrm}@scms.rgu.ac.uk

1. THE CONCEPT

Mediated access through a clustered collection is a new approach to accessing very large heterogeneous document collections. The user explores a relatively small, homogeneous, pre-clustered, and thus well structured, document collection covering a particular subject domain (called the *source collection*), in order to understand the concepts encompassed and to clarify and refine his/her information need. The user ostensibly indicates clusters and documents of interest and is assisted in formulating a query, based on which a search is done on a large, heterogeneous, non-structured collection (called the *target collection*). The original cluster structure is the basis for visualisation tools that allow the user to explore search results.

2. THE IMPLEMENTATION

WebCluster, the system implementing this idea, has a client-server architecture. The server provides access to different target collections, including the Web, and incorporates a clustering framework that provides access to clustered source collections. The server also executes the best-match and cluster-based searches requested by the client. The client implements the user-interface. Two versions of the interface are available, implementing scenarios for different classes of users.

The *explicit scenario* is targeted at experienced users, who are expected to understand the idea of mediated access, to distinguish between the source and target collections, and to have initiative in generating and editing queries and in sending queries to the target collection. In a typical search session, the user selects a clustered source collection, appropriate for the domain of interest, browses or searches it, and thus identifies the cluster or documents that best meet his/her information need. Based on the user's selection, the system proposes a query, which may be edited, according to how precise or how comprehensive the search needs to be. The

user then selects a target collection and submits the query to it via a meta search engine incorporated in the server. The client displays the snippets returned and can retrieve the full document if the user requires so.

The *implicit scenario* is targeted at naive users. The query generation and the searching of the target collection are done in the background, based the user's actions, and relative to the ostensive indication of an information need. The documents retrieved from the target collection are integrated in the structure of the source collection, so that the user is not aware of the existence of two different collections.

3. ACKNOWLEDGEMENTS

The WebCluster Project is sponsored by Ubilab, Union Bank of Switzerland, Zurich. WebCluster uses Ubilab's fusion meta-search engine, Informia[1,2].

References

- [1] M. Barja, T. Bratvold, J. Myllymaki, and G. Sonnenberger. A mediator for integrated access to heterogeneous information sources. In *ACM Conference on Information and Knowledge Management (CIKM '98)*, November 1998.
- [2] 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 (to appear)

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.
SIGIR '99 8/99 Berkeley, CA, USA
© 1999 ACM 1-58113-096-1/99/0007...\$5.00