

# Document Clustering for Mediated Information Access

David J. Harper, Mourad Mechkour  
Gheorghe Muresan

School of Computer and Mathematical Sciences, The Robert Gordon University  
Aberdeen AB25 1HG, Scotland, UK  
{djh,mrm,gm}@scms.rgu.ac.uk

## Abstract

This paper addresses the problem of accessing very large heterogeneous document collections by proposing a new approach to using clustering for information retrieval: **mediated access through a clustered collection**. In what is actually an information access environment, the user can explore a relatively small, well structured, pre-clustered collection covering a particular subject domain, in order to understand the concepts encompassed and to clarify and refine his/her information need. The user can ostensibly indicate clusters and documents of interest and be assisted in formulating a query, based on which a search can be done on a large, non-structured collection. Finally, the original cluster structure is the basis for visualisation tools that allow the user to explore search results. WebCluster, the system implementing these ideas, is presented, together with results of an initial formative experiment and plans for future experiments.

**Keywords:** Information Retrieval, Document Clustering, Mediated Access, World Wide Web.

**Acknowledgement.** The WebCluster Project is sponsored by Ubilab, Union Bank of Switzerland, Zurich.

## 1 Introduction

The initial aim of introducing *document clustering* in information retrieval was to increase the speed of retrieval. After an initial overhead, which groups the documents in clusters based on their reciprocal similarity, the search identifies clusters of documents that best match a query rather than individual documents[26].

It was Jardine and van Rijsbergen[13] who first suggested that the associations between documents convey information about the relevance of documents to requests and formulated the **cluster hypothesis**: “closely associated documents tend both to belong to the same clusters and to be relevant to the same request”. They experimentally showed that *cluster-based retrieval* could yield more effective results than best-match retrieval. A lot of research went into developing methods, algorithms, similarity measures and weighting schemes with view to increasing the *efficiency* and *effectiveness* of cluster-based retrieval. However, cluster-based retrieval has not proved consistently superior to best-match retrieval in terms of effectiveness[3, 22], and it is much more expensive than the inverted index search in terms of disk space and processing time[23]. However, we believe that abandoning cluster-based retrieval on the basis of this earlier work is premature. Much of this experimental work explored retrieval in a batch mode fashion, and we believe that cluster-based approaches have the capacity to exploit the inherent structure of document collections through interactive searching.

Recently, the system-centred approach has been giving way to the user-centred, task-oriented approach in IRS design. Users are seen as searchers of information and problem solvers in a certain cognitive space. Their interaction with the IRS is not limited to a simple interface for sending queries and receiving results, but is based on an *information-seeking environment*[17, 10, 11] or an *information workspace*[19]. This environment assists the user in defining and planning his/her tasks, developing a search strategy, monitoring the progress, analysing results, re-formulating queries

and even tasks if necessary. This standpoint creates new opportunities for clustering, in that users can be assisted in understanding the topical structure of a collection (and thus the subject domain), in refining their information need and formulating queries, and in exploring semantic relationships between documents. Systems designed according to this view typically combine searching and browsing, as searching can identify starting points for browsing, while browsing may reveal unexpected aspects of the domain and may suggest keywords for searching[5, 15]. Consequently, much work has gone into developing interfaces that support visualisation and browsing[19, 7]. Some authors eliminate searching completely and have proposed new retrieval paradigms, which focus either on the organised display of *all* documents[14] or on progressive refinement of the information need while exploring the information space[4].

However, clustering and visualizing the entire collection meet with scalability problems. As the size of the collection on which retrieval is done has increased dramatically, clustering the collection or manually categorising it may be difficult and expensive or even impossible (as is the case with the World Wide Web). The most common approach seems to be the *query-specific clustering* approach, proposed by Willett[24]. The results of a 'traditional' Boolean or best-match search, containing only the documents considered relevant to the query, are clustered and presented to the user[6, 9, 28, 27]. A consequence of this is that the clustering needs to be done on-the-fly, so fast clustering algorithms are needed, and this results in reduced accuracy and, implicitly, effectiveness. There are other problems with this approach, namely the user has no idea of the structure of the collection and is not assisted in clarifying or formulating his/her information need. Moreover, this approach does not exploit the full potential of cluster-based searching to retrieve closely related relevant documents, some of which may not match the query.

In this paper we propose a new approach to using clustering, namely *mediated retrieval through a clustered collection*. We address some of the shortcomings highlighted above, of previous work on cluster-based retrieval.

## 2 Mediated access through a clustered collection

The idea of *mediated access*, in the sense described in this paper, is based on the existence of specialised collections, administered and maintained by specialised organisations and companies, that cover certain domains of interest. These collections (which we are going to call *source collections*) are assumed to be up-to-date, comprehensive and homogeneous in content. By clustering such a collection, the semantic and topical structure of the domain is revealed to the user. This is essential for the user in exploring and understanding the domain. Figure 1 depicts the idea of mediated access as a two-stage process. First, the user searches and browses a relatively small, clustered source collection, and thereby refines his/her information need relative to the domain. The user chooses documents and/or clusters of interest, thus ostensibly indicating his/her information need. Second, the system proposes a query based on the selection made by the user, which is submitted to the target collection, for example a subcollection of the Web, indexed by a search engine.

It is expected that clustering produces a meaningful structure of the collection, revealing themes and concepts covered, especially when the collection contains abstracts or short documents[8]. For collections of long documents, each covering several topics or subtopics, manual categorisation may be more appropriate for revealing the structure of the domain[6].

## 3 Users and scenarios of use

In designing the WebCluster system and implementing the *mediated access* paradigm, we have addressed some well-known problems:

- the difficulty many users have in clarifying and refining their information need, and especially in an unfamiliar domain. This can happen if the information need is vague or ill-defined, if the user is learning about a domain, or if the information request is assigned to a search intermediary.

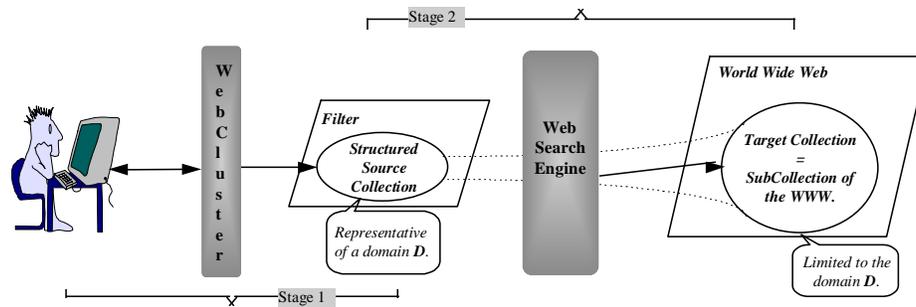


Figure 1: Mediating Access to a Document Collection.

- the difficulty the users have in formulating the right query, using the appropriate vocabulary and syntax[12] and conveying the right characteristics of the information need: precision, coverage, granularity, comprehensiveness.
- the difficulty the users have in exploring and understanding search results.
- the difficulty or impossibility of directly clustering or categorising very large collections in order to reveal their structure.

The solution that we propose is rather general. In order to assess its validity, it will have to be evaluated in various contexts, with different classes of users. However, before the proposed paradigm matures and the reaction to it can be better understood, formative evaluation is necessary in order to test our assumptions on some aspects of its use. This evaluation will help us shape our understanding of the users' behaviour in the face of the new paradigm, and it will dictate some design decisions concerning the implementation of the paradigm in retrieval systems and tools.

Three dimensions are proposed in order to discuss several possible ways in which mediated access can be presented to, and accessed by, end users:

1. Explicit vs. implicit search of the target collection.

In the explicit case, illustrated in Figure 2, the user is aware of the two stages of retrieval: the exploration of the source collection and the ostensive formulation of the information need, based on the selection of relevant clusters, followed by the search of the target collection, based on the query proposed by the system and possibly edited by the user. The user is in control - he/she initiates the search when he/she considers that his/her information need has been reasonably well expressed in terms of precision and comprehensibility.

In the implicit case, illustrated in Figure 3, the searching of the target collection is done by the system in the background, outwith the user's control, and the target documents are presented to the user when needed. Typically, the structure created by clustering the source collection is pre-populated with some target documents; if it is clear that the user wants to see more documents, a more extensive search is generated. The user does not see the query generated and cannot correct it, and thus the requirement for a very good query generation algorithm is paramount.

Experiments are necessary in order to assess the quality of the queries generated by the system and also the improvement that the user can bring by editing the query.

2. Seamless source and target collections vs. distinguishable source and target collections.

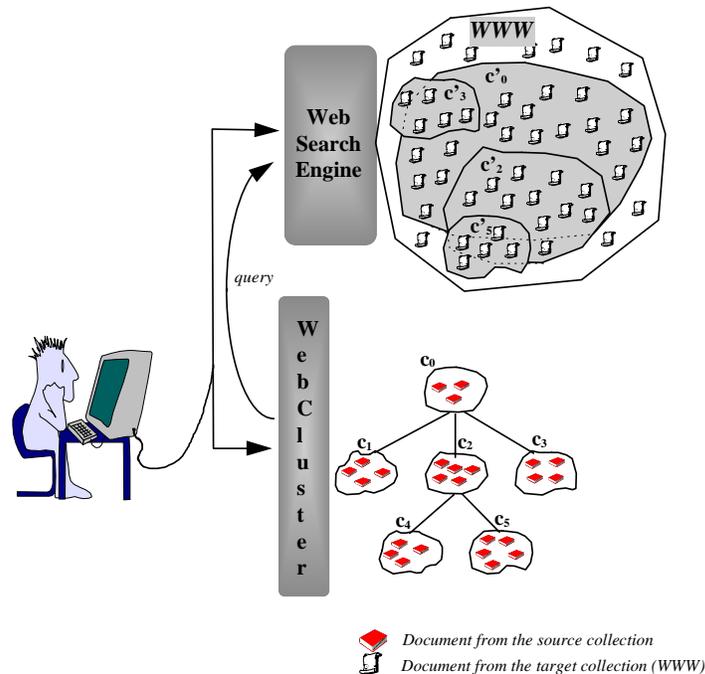


Figure 2: Explicit Mediated Access to the World Wide Web.

In the former case, the user is interested in getting relevant documents, no matter their source. In the latter case, the user is aware of the existence of the two collections and of their specific roles: the source mediates the retrieval, the target provides the bulk of the documents. There is a certain dependence between this dimension and the previous one: the former case goes naturally with the implicit scenario, while the latter one with the explicit scenario.

Interviews and observations of users doing different tasks should shed light on the appropriateness of each case.

### 3. Visibility vs. non-visibility of the source documents.

One point of view is that, once the source collection has been clustered and the structure of the domain of interest revealed, the actual source documents are not needed for mediating the retrieval. The structure of the domain should be clear from the structure of the cluster hierarchy built and the themes and concepts of the domain should be indicated by the topical keywords that form the cluster representatives. An objection is that typical documents may be, in the users' perception, more indicative of the contents of a cluster than the topical keywords.

Experiments should be conducted in order to assess the representativeness of keywords obtained with different indexing methods and weighting schemes, of document titles and of document bodies.

## 4 The WebCluster architecture

We have designed a general architecture, for realising systems that deliver mediated access through a clustered collection. The resultant client-server architecture is depicted in Figure 4.

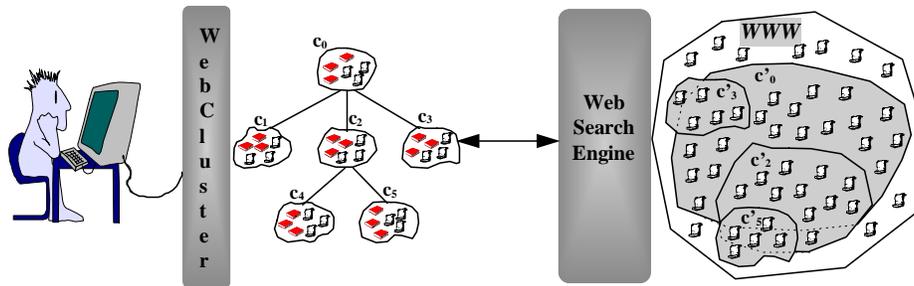


Figure 3: Implicit Mediated Access to the World Wide Web.

The client, implemented in Java, represents the user interface providing the user with functionality for choosing a source and a target collection, for visualising the structured source collection, for searching and browsing it, for exploring the results retrieved, for bookmarking items, for saving and printing results. A future extension of the client will be coupled with a Web browser, so that the Web can be browsed starting from retrieved pages.

The server, implemented in C++, runs on a fast Unix machine and is able to serve multiple clients simultaneously. It provides access to different document collections, including the Web, and incorporates a *clustering framework* that provides clustering functionality. Several clustering methods, similarity measures and cluster representation methods are implemented. The server also executes the best-match and cluster-based searches requested by the client.

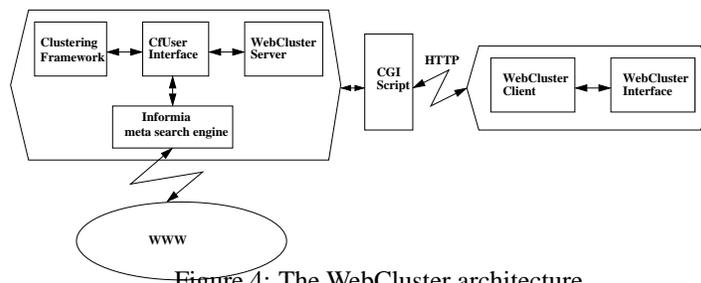


Figure 4: The WebCluster architecture.

**WebClusterClient** and **WebClusterServer** implement the connectivity between the client and the server through a CGI script, using the HTTP protocol. On the client side, **WebClusterInterface** implements the user interface. On the server side, **CfUserInterface** is the *facade* that ensures access to different modules of the server. **ClusteringFramework** is, as the name suggests, an integrated tool that can be used to cluster different document collections and to search the cluster structure using different search strategies. **Informia** is a fusion meta-search engine developed by our collaborators at Ubilab. It sends a query to a range of information sources, including Web search engines, which can be fixed for an application or set by the user, and it fuses the results returned by these sources[1].

A detailed description of the system design and architecture will be available in a different paper[18]. For the purpose of introducing the concept of *mediated access* the presentation of the user interface is sufficient.

## 5 The user interface

The principles of user-centred design command that the requirements for different classes of *users* in different *contexts* and for different *tasks* be analysed and well understood. They drive the design of the interface and, subsequently, the design of the system. As we are in the stage of formative experiments, various scenarios need to be tested and analysed with various classes of users before final decisions can be made regarding the final, operational system. Several interface prototypes, modelling different scenarios and employing different metaphors, need to be implemented and tested. From a software engineering point of view, the *model-view-controller* paradigm has been adopted, so that changes in the view can be made easily.

The first prototype, depicted in Figure 5, implements the explicit scenario, in which the user is in complete control of all actions. He/she scans the structure of the source collection and asks the system to generate an appropriate query based on the cluster which best describes his/her information need. He/she can inspect this query and may choose to edit it before submitting it to the target collection.

The screenshot displays the WebCluster End User Interface. At the top, there are three main sections: 'Source collection' (containing 'Restor000\_NoStemComplete'), 'Query' (containing 'underwriting bond issued'), and 'Target collection' (containing 'WWWBellevad'). Below these are buttons for 'Select', 'Info', and 'Search'. The main interface is divided into three columns: 'Overview' (a hierarchical tree of clusters), 'Local view' (showing 13 documents with a preview of the top document), and 'Ranked view' (showing 22 results with scores and URLs). The 'Overview' column shows a tree structure with clusters like '(13) doc underwriting selling payment management issuing' and '(11) doc underwriting selling payment management issuing'. The 'Local view' shows a list of documents with a preview of the top document: 'underwriting selling payment management issuing issues eurobond date manager listed lead concession convertible priced paid deamissioes: bond available'. The 'Ranked view' shows a list of results with scores and URLs, such as 'Score: 0.6 [Ivuc0.44,Infoec0.5d] Ref: http://www.law.indiana.edu/codes/ind'.

Figure 5: The WebCluster End User Interface: the Explicit Scenario.

Perhaps the best way to describe the interface and its functionality is by a *walkthrough*. Suppose a user is interested in finding on the WWW information on underwriting (in the banking context). The search would proceed as follows:

1. Use the **Source Collection Panel** to select a source collection appropriate for the domain, e.g. the Reuters collection of news stories might be appropriate for a financial search.
2. Browse the clustered source collection, and find a cluster of relevant source documents corresponding to the information need. The clustered source collection is visualized in the **Overview Panel**, and the hierarchy can be explored by the user by expanding/collapsing clusters. The **Local View Panel** is used to display the content of a cluster, and to display the content of a document. In our case, the cluster representative containing the keywords “underwriting selling payment management . . .” attracted the user’s attention and the exploration of the documents contained in the cluster confirmed the identification of a relevant cluster.
3. Ask the system to generate a query based on the keywords found in the cluster representative, derived from the source documents in the cluster. This query may be edited by the user, according to how precise or how comprehensive the search needs to be. In the **Query Panel** the query has been edited and reduced to three terms, considered relevant and sufficient to express the information need.
4. Use the **Target Collection Panel** to select a target collection and query it using the generated query. Note, there may be a number of target WWW collections, corresponding to different WWW search engines (or combinations of engines)<sup>1</sup>.
5. A ranked list of retrieved target documents are displayed in the **Ranked List Panel**, and individual target documents (actually text snippets currently) can be selected for display in the **Local View Panel**.

The above procedure was used in formative experiments. At the time, the exploration of the source collection was only possible by browsing. In the future, cluster-based searching will also be available.

In the case the user knows the domain and has no problems in formulating a query, it is also possible to search the target collection(s) directly, by entering a query in the **Query Panel** and initiating a search of the target collection.

During a retrieval session, a user can bookmark retrieved documents and/or clusters, and thus save the results of a search session. Printing the results will also be possible.

## 6 Experiments

The evaluation of an IR system looks at how well the system meets its purpose of satisfying an information need. The overwhelming majority of IR evaluations have used the *system-centred approach*, which considers the information need given and unchangeable during the experiment, and the presentation of the results of minor importance. The quality of a system, according to this approach, is given by *efficiency*, based on computer resources used (storage and CPU time) and especially *effectiveness*, based on **recall** (the percentage of relevant documents that are retrieved) and **precision** (the percentage of retrieved documents that are relevant)[20]. More recently, the *user-centred approach* has been gaining ground - it considers the user and his/her satisfaction during an *interactive* retrieval session paramount. Saracevic[21] is adamant that both aspects are needed, and that together they should provide a comprehensive picture of IR performance.

In our case, the main hypothesis that we propose is: **In an interactive setting cluster-based mediated retrieval can be more effective than traditional retrieval.** The evaluation of this hypothesis will have to take into consideration all the aspects of the retrieval process: query formulation, examination of the retrieved documents, selection of the relevant ones, and to include the possibility of an information need shift or refinement during the search process or even a berry-picking strategy[2] of searching and collecting information. This will be the major experiment of the WebCluster project, for which the appropriate context and scenario need to be designed, coupled with a representative sample of users and set of tasks to be solved. The analysis of the experiment (based on the direct observation of

---

<sup>1</sup>In the current system, we access the WWW through Informia.

the users, on the think-aloud protocol, on logs of user actions and of results and on post-task questionnaires) should look for possible differences in retrieval performance and in user satisfaction in terms of effort, time, pleasure, and perception of completion of the task.

However, before such a complex evaluation can be done, a set of smaller scale tests are necessary in order to assess the usability of the interface, the acceptance by the users of the mediated access paradigm, in the explicit and in the implicit form, and also to select parameters such as the clustering method, the weighting scheme, and the method for generating cluster representatives.

So far, we have conducted an informal, proof-of-concept experiment in order to test the reaction of the users regarding the new concept we are proposing. In the next subsection we present this formative experiment, together with a discussion of the results. Based on these, a set of more formal experiments to be conducted in the future are proposed in the second subsection.

### 6.1 Informal proof-of-concept experiment

#### Objectives

The objectives of the first, informal experiment were to get an initial reaction to WebCluster, to gauge the usability of the system, and to see if the users were more satisfied with the results obtained by mediated search compared with a direct search. By observing the users using the system and by recording their questions and comments, we also hoped to uncover issues that needed to be addressed in order to make the system useful and usable and to satisfy the users.

#### Procedure

1. The subjects were introduced to the idea of mediated retrieval and to the use of the system.
2. The work task situation was described. The subject had to imagine that he/she was a journalism student on a work placement at a large national daily newspaper. His/her job was to support the journalists in writing articles by finding relevant information on assigned topics. The sample tasks were:
  - The journalist is writing an article on the coffee industry. He/she wants to know (if and) how quotas for growing coffee are set and controlled on a world-wide basis.
  - The journalist writing an article on strategic stocks of raw materials in the US wants to know details of the US oil reserves.
  - The journalist want details on the history of the Brazilian debt crisis.
  - George Shultz visited the Soviet Union for talks with Gorbachev on missile reduction. Details of the visit and of subsequent related visits are needed.

The topics were manually selected by the experimenters from a collection of Reuters articles.

The subject (journalism student) was expected to retrieve as much relevant information as possible, so that a journalist could find enough useful material to write an article on the subject.

3. The subject was asked to pick a topic (the recommendation was that the subject knew as little as possible on the topic) and to write down the query that he/she would be likely to submit to a search engine.
4. The subject was asked to select the Reuters collection from a pool of clustered source collections and to browse it<sup>2</sup> in order to find the best cluster that matched the description of the topic (he/she could explore the cluster representatives and also the documents that made up the cluster). The user could ask the system to generate a query based on the chosen cluster, and could edit the query before submitting it for a search on the target collection (the Web, through Informia).

---

<sup>2</sup>Searching had not been implemented at the time of the experiment.

5. The subject used the initial, self-generated query, for a search on the target collection and compared the results to the ones obtained based on mediated search. Although in the time used for mediation, the user could have alternatively re-formulated the query based on the initial results and improved his/her results, the comparison is not unrealistic; the analysis of typical searchers on the WWW revealed that the majority do not follow the initial search with successive queries[12]

The *think-aloud* protocol was used. The examiners took notes on the users' reasoning and actions. No logging of the actions and no post-task questionnaires were used.

### Users

The users were all IR researchers, participating in a European workshop on evaluation of interactive, multimedia IR systems (MIRA). They had a good understanding of search strategies and indexing, and were experienced in using IR systems and in formulating queries. They acted as expert reviewers and their comments on the usefulness and usability of the system during the think-aloud retrieval sessions proved extremely valuable. The fact that they chose topics that they knew little about made the experiment realistic, namely by simulating searching by an intermediary.

### Results

1. Most users felt comfortable with the idea of mediated access. They found it particularly useful when they were not familiar with the domain and therefore had problems in producing keywords for the query. This confirms our expectation that mediation can be valuable in assisting a user, especially a naive one, in formulating his/her information need.
2. When the topic was more familiar, the users felt they could generate good queries without assistance and showed a certain resistance to spending time with the mediation stage. However, constrained to using the experimental procedure, they did use mediation and were pleasantly surprised to see the improvement in retrieved results due to this stage.

The resistance noticed corroborates with other research results showing most users to be result-driven rather than process-driven, trying to get just the minimum required of a task, with a minimum of effort<sup>3</sup>. This would suggest that the interface implementing the implicit scenario, with only one retrieval stage, may be better received by the users than the one implementing the explicit scenario, which involves a two-stage process. However, the user interface implementing the implicit scenario was not ready for testing, so the results of the informal experiment cannot be conclusive. More experiments are needed, with a wider range of user expertise. Moreover, the information need should not be explicitly assigned to the users; rather, the users should be immersed in a context and given a task that forces them to formulate an information need and do a search. Also, the information needs should range from rather general to precise ones, and the conditions in which the task is considered solved should be more explicit. It may also be worth looking for a possible change in the users' reaction as the familiarity with the system grows. If the users perceive an improvement in the search results due to mediation, they may grow to want it.

3. Some users questioned the necessity of searching the target collection (the second step of the explicit scenario) when some documents retrieved from the source collection were relevant for the task topic. This may be due to the lack of an explicit stopping condition for the search, respectively to a somehow imprecise task. It may also be due to a lack of rigour from subjects in assessing if the task has been achieved: after starting the search, they never re-read the task or compared it to the documents retrieved, neither did they consider if the information gathered would be enough for writing an article. No user made any attempt to find information on related topics. They wanted to stop as soon as a minimal set of results seemed to be sufficient for satisfying the task. A more detailed experiment, with the precision and recall

---

<sup>3</sup>This was verbally reported by Victoria Mangano, City University, London.

requirements more clearly stated would be more conclusive (we expect that if a comprehensive search, is required, the users are more likely to extend the search to the target collection). So far, however, what we have is another indication of preference for the one-stage search (or rather for a quick search).

4. When the user edited the query proposed by the system, by deleting the words that were not relevant in the context, a significant increase in precision and recall was noticed, compared to the search on the user's original query. However, when the query was not edited, the results were worse. This suggests that more work needs to be done on the query generation algorithm, and thus on the algorithm that calculates cluster representatives. The current algorithm considers keywords that are common among documents of each cluster. A probabilistic model, which considers statistical information of the collection will be tested in the future (see subsection 6.2). If the user's intervention remains necessary, then the implicit scenario may prove infeasible.
5. The users considered that there were too many clusters at the top level and it was difficult to select one for top-down browsing. While this problem will be partially solved by adding a search function, it is worth investigating alternative clustering methods, which produce different structures. The complete-link clustering method has been used because it produces a structure with small, tightly bound, well defined clusters, hopefully more easily identifiable by the users. It has also been shown to produce good search results[25]. The drawback is the wide shallow structure that it produces. The use of other methods may reveal other advantages: the single-link methods, because of its *chaining* effect[13], may reveal unexpected links between apparently unrelated documents, topics, or sub-domains, while a method that allows overlapping may generate a more realistic structure. For the purpose of browsing, a clustering method that produces a better balance between the *width* and the *depth* of the structure would be useful. Too many alternatives cannot be stored in the short-term memory, so selecting a sub-cluster to explore may be problematic for the user. On the other hand, too deep a path can create problems in assessing the context, so that the user can get 'lost'.
6. The users found that often the cluster representatives did not convey well the contents of the clusters and had to look at sample documents to judge the relevance of the cluster. Finding the best cluster (and document) representative for browsing, respectively for generating the best query is definitely one of the next challenges for us.
7. Many keywords recommended by the system were observed to improve retrieval, and this provides some evidence of the benefits of mediated search for assisting query formulation. However, users sometimes rejected proposed terms, even though these terms were known to be useful by the experimenters (who were familiar with the topics and with the collection). Research in *interactive query expansion* has shown inexperienced users' failure to recognize 'good' terms proposed by the system[16]. Our informal experiment seems to indicate the same failure, especially for users unfamiliar with the domain. Therefore, if the explicit scenario is to be successful, tools will have to be provided to assist the user in making good decisions e.g. by showing terms in context as in snippet searching, or by ranking the list of proposed terms, etc[19].
8. Of the pages retrieved from the Web, a high proportion were newspaper articles. This supports anecdotal evidence that different characteristics of documents such as the structure, the genre, the language, the formatting style, etc may act as features in clustering[8]. More experiments are needed to confirm this.
9. The users pointed out some lack of functionality in the system, such as a *colour model* for labeling different kinds of objects (e.g. previously viewed documents), an integrated *browser* for viewing the WWW documents, a *search* facility for the source collection, *bookmarks*, a *results collector*, a *history* function, and so on. Some of this functionality had been designed and it will be in place for future experiments.

## 6.2 Future experiments

The main experiment, discussed in the introduction to this section, needs to be preceded and supported by lower-scale experiments that would establish the version of the WebCluster that is accepted by the users and gives best performance.

One experiment should look at the user preference with regards to the two scenarios proposed: the two-step explicit scenario and the one-step implicit scenario. One of the two systems and one or more from a set of tasks will be randomly allocated to each user. Performance and user preference are to be assessed.

More evidence needs to be sought regarding the users ability to distinguish between 'good' and 'bad' terms for query formulation. The effectiveness of tools to support the users' decision by showing the context of a term, based on document or cluster content, or by ranking potential terms, should be assessed.

Other experiments are needed regarding an issue that is central to the idea of *mediated retrieval*: the generation of **cluster representatives**. In WebCluster, the cluster representative serves two purposes:

1. for searching - it has to accurately represent the contents of the cluster, and it has to produce, when used as a query, high retrieval effectiveness.
2. for browsing - it has to allow the user to make the distinction between sibling clusters.

A probabilistic model (described in [18]) has been developed in order to produce a 'searching representative' of a cluster and a 'browsing representative'. User-less experiments are necessary in order to determine the coefficients and cut-off threshold that generate best retrieval performance for the former, while user experiments are necessary in order to test the power of discrimination, in the user's perception, of the latter.

## 7 Conclusions

The paper examines the use of document clustering in IR and proposes a novel use of it for accessing very large heterogeneous document collections in an interactive environment. The new paradigm of *mediated retrieval based on a clustered collection* is presented and various versions of it are analysed. WebCluster, the system that implements the idea is presented, with an emphasis on the user interface, in which the paradigm has the major effect.

An informal proof-of-concept experiment is presented and analysed. The results and the reaction of the users were encouraging, showing the potential of mediated access as a tool for query formulation support, for searching and browsing a domain of interest. The experiment also revealed a number of issues to be analysed and indicated a set of more formal experiments whose results will improve and refine our solution. They will shape design choices regarding the scenario to be made operational, the algorithm for the cluster representative and the clustering method of choice.

Finally, a new 'clustering hypothesis' is proposed - in an interactive setting cluster-based mediated retrieval can be more effective than traditional retrieval - and some aspects of the experiments that would test it are discussed.

## 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] M. J. Bates. The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5):407-424, 1989.

- [3] R. Burgin. The retrieval effectiveness of five clustering algorithms as a function of indexing exhaustivity. *Journal of The American Society for Information Science*, 46(8):562–572, 1995.
- [4] I. Campbell and C. J. van Rijsbergen. The ostensive model of developing information needs. In Peter Ingwersen, editor, *Proceedings of Conceptions of Library and Information Science (CoLIS-2)*, Copenhagen, October 1996.
- [5] J. Furner, D. J. Harper, and D. G. Hendry. Coordinated support for browsing, searching and monitoring: A user interface for networked information retrieval. IR/HCI workshop in Glasgow, September 1996.
- [6] M. Hearst. Using categories to provide context for full-text retrieval results. In *Proceedings of RIAO '94*, pages 115–130, Rockefeller, 1994.
- [7] M. Hearst. Cat-a-cone: An interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. In *Proceedings of SIGIR '97*, pages 246–255, Philadelphia, July 1997. ACM.
- [8] M. A. Hearst. *Natural Language Information Retrieval*, chapter The Use of Categories and Clusters for Organizing Retrieval Results, pages 333–373. Kluwer Academic Publishers, 1999.
- [9] M. A. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In H.-P. Frei, D. Harman, P. Schauble, and R. Wilkinson, editors, *Proceedings of SIGIR '96*, pages 76–84, Zurich, Switzerland, August 1996. ACM.
- [10] D. G. Hendry. *Extensible Information-Seeking Environments*. PhD thesis, School of Computer and Mathematical Sciences, The Robert Gordon University, Aberdeen, September 1996.
- [11] D. G. Hendry and D. J. Harper. An informal information-seeking environment. *Journal of The American Society for Information Science*, 48(11):1036–1048, November 1997.
- [12] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: A study of user queries on the web. *SIGIR Forum*, 32(1):5–17, 1998.
- [13] N. Jardine and C. J. van Rijsbergen. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, 7:217–240, 1971.
- [14] R. R. Korfhage. To see, or not to see - is that the query? In *Proceedings of SIGIR '91*, pages 134–141, Chicago, October 1991. ACM.
- [15] Z. Lacroix, A. Sahuguet, R. Chandrasekar, and B. Srinivas. A novel approach to querying the web: Integrating retrieval and browsing. ER97 Workshop on Conceptual Modeling for Multimedia Information Seeking, Los Angeles, November 1997.
- [16] M. Magennis and C. J. van Rijsbergen. The potential and actual effectiveness of interactive query expansion. In *Proceedings of SIGIR '97*, pages 324–332. ACM, July 1997.
- [17] G. Marchionini. *Information Seeking in Electronic Environments*. Cambridge University Press, 1995.
- [18] M. Mechkour, D. J. Harper, and G. Muresan. Mediated access: A novel technique for searching the world wide web. Submitted to SIGIR '99.
- [19] R. Rao, J. O. Pedersen, M. A. Hearst, J. D. Mackinlay, S. K. Card, L. Masinter, P.-K. Halvorsen, and G. G. Robertson. Rich interaction in the digital library. *Communications of The ACM*, 38(4):29–39, April 1995.
- [20] C. J. van Rijsbergen. *Information Retrieval*. Butterworth and Co, London, 2nd edition, 1979.
- [21] T. Saracevic. Evaluation of evaluation in information retrieval. In *Proceedings of SIGIR '95*, pages 138–146. ACM, July 1995.

- [22] W. M. Shaw Jr, R. Burgin, and P. Howell. Performance standards and evaluations in ir test collections: Cluster-based retrieval models. *Information Processing and Management*, 33(1):1–14, January 1997.
- [23] E. M. Voorhees. The efficiency of inverted index and cluster searches. In *Proceedings of SIGIR '86*, pages 164–174, Pisa, Italy, September 1986. ACM.
- [24] P. Willett. Query-specific automatic data classification. *International Forum on Information on Documentation*, 10(2):28–32, April 1985.
- [25] P. Willett. Recent trends in hierarchic document clustering: A critical review. *Information Processing and Management*, 24(5):577–597, 1988.
- [26] S. Worona. *Scientific Report No. ISR-16 - Information Storage and Retrieval - to the National Science Foundation*, chapter Query Clustering in a Large Document Space, pages XV–1–XV–22. Department of Computer Science, Cornell University, Ithaca, USA, September 1969.
- [27] O. Zamir and O. Etzioni. Web document clustering: A feasibility demonstration. In *Proceedings of SIGIR '98*, pages 46–54, Melbourne, August 1998. ACM.
- [28] O. Zamir, O. Etzioni, O. Madani, and R. M. Karp. Fast and intuitive clustering of web documents. In *The Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*. AAAI, 1997.