

Topic Modeling for Mediated Access to Very Large Document Collections

Gheorghe Muresan

Department of Library and Information Science, Rutgers University, New Brunswick, NJ 08901. E-mail: muresan@scils.rutgers.edu

David J. Harper

School of Computing, The Robert Gordon University, Aberdeen AB25 1HG, Scotland, United Kingdom. E-mail: djh@comp.rgu.ac.uk

Clear and precise queries are a necessity when searching very large document collections, especially when query-based retrieval is the only means of exploration. We propose system-mediated information access as a solution for users' well-documented inability to formulate good queries. Our approach is based on two main assumptions: first, on the ability of document clustering to reveal the topical, semantic structure of a problem domain represented by a specialized "source collection," and, second, on the capacity of statistical language models to convey content. Taking the role of the human mediator or intermediary searcher, a mediation system interacts with the user and supports her exploration of a relatively small source collection, chosen to be representative for the problem domain. Based on the user's selection of relevant "exemplary" documents and clusters from this source collection, the system builds a language model of her information need. This model is subsequently used to derive "mediated queries," which are expected to convey precisely and comprehensively the user's information need, and can be submitted by the user to search any large and heterogeneous "target collections." We present results of experiments that simulated various mediation strategies and compared the effect on mediation effectiveness of a variety of parameters, such as the similarity measure, the weighting scheme, and the clustering method. They provide both upperbounds of performance that can potentially be reached by real end users and a comparison between the effectiveness of these strategies. The experimental evidence suggests that information retrieval mediated through a clustered specialized collection has potential to improve effectiveness significantly.

Introduction

Various studies of search engine logs (Jansen, Spink, & Saracevic, 2000) and comparisons of mediated and unmediated searches (Nordlie, 1996, 1999; Spink, Goodrum, & Robins, 1998) indicate that assistance given to users in their query formulation is an essential factor in retrieval success. Helping the users formulate their information need clearly and precisely is especially required for searching very large document collections, where retrieval precision is crucial.

Unassisted users have problems in formulating high quality queries due to a lack of sufficient knowledge of the appropriate vocabulary, inability to use advanced query language syntax, a false impression that "the computer knows what I want," or a lack of clear understanding of the system's conceptual model. Rather than providing precise queries, containing terms with a high power of discrimination, users typically submit very short queries, often made up of ambiguous or common words.

Moreover, in the case of exploratory searches, the users often 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* (Belkin, Oddy, & Brooks, 1982). In such cases, expressing one's lack of knowledge in a coherent and precise manner is impossible. Clarification and refinement of the user's need are part of the information-seeking process.

In the case of mediated searches, the librarian or intermediary searcher initiates a dialog with the user and elicits information in order to establish the user's context and problem domain. During this interaction, the user's information need usually evolves from the initial visceral, or conscious form to a clearer formalized form, representing a qualified and rational statement of the user's question (Taylor, 1968). Moreover, the mediator can derive the compromised need, dependent on the available retrieval system's

Accepted November 7, 2003

© 2004 Wiley Periodicals, Inc. • Published online 19 March 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20034

coverage, specialization, indexing model, and format, in anticipation of what the system can deliver.

The much higher rate of success observed for mediated searches (Nordlie, 1996) prompted work into means of assisting the user in exploring the domain of her problem, in clarifying and refining her information needs, and in formulating queries that reflect her information need accurately. This research effort has mainly taken two directions¹:

1. User interfaces and visualization tools that support the user's exploration. They typically provide tools for filtering and structuring the information space on various metadata or topical axes and for highlighting relationships between documents or between documents and formulations of the information need at various stages of the interaction.
2. Theoretical work geared at proposing models for representing documents, topics, collections, and user queries, for capturing the user's relevance feedback, and for building the user's context or profile.

We combined results from these two research streams by proposing an interaction model based on *system-based mediation* through structured specialized collections.

In previously reported work (Harper, Mechkour, & Muresan, 1999; Muresan & Harper, 2001), we described the design and architecture of WebCluster, a system that reifies our proposed interaction model,² discussed potential applications, and reported results from informal user experiments. The results of those experiments were encouraging: The users found the system usable and also useful. Especially for unfamiliar topics, 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. In this report, we discuss in more detail mediation as an interactive retrieval model and our approach to modeling relevant topics, propose an evaluation framework, and report more formal and more extensive experimental results.

The remainder of this report is structured as follows. In The Interaction Model, we briefly review and then discuss in some detail the proposed interactive model for system-mediated retrieval and compare it to related approaches. The theoretical model used for representing documents, collections and, especially, the user's topic of interest is described in Topic Models. We then present the rationale for our experiments and describe the experimental setting in Experimental Setting. The results are presented and discussed in experiments. Finally, we draw conclusions and propose future work.

¹A review of the relevant literature is available in Muresan (2002).

²Several versions of the system were demonstrated at SIGIR 1999 (Muresan, Harper, & Mechkour, 1999), SIGIR 2000 (Muresan, Harper, Goker, & Lowit, 2000), and ECDL 2001 (Muresan, Harper, & Goker, 2001).

The Interaction Model

Overview of the Model

System-based mediated information access refers to the system assisting the user in investigating a domain of interest, in exploring and refining an information need, and in generating a query that conveys the information need accurately.

Our main target is the user framed by Belkin's ASK model (Belkin et al., 1982), who has a problem to solve but does not know what information is needed or how that information could be obtained. We supply a highly interactive interface and visualization tools, which allow the user to explore the domain of her problem, to become familiar with its terminology and topical structure, and to start developing a feel for the kind of documents or solutions that she is looking for. We rely on the existence and availability of specialized collections of documents or abstracts maintained by various companies and organizations. These collections are representative for their domain and their structure conveys the topical structure of the domain. Some of these collections were manually classified by their creators.³ Automatic classification or document clustering can be applied to the others in order to reveal their topical structure.

The mediated retrieval process is depicted in Figure 1. The exploration of the specialized source collection supports concept learning for the user unfamiliar with the problem domain. Moreover, based on the user's exploration of the collection, and on her selection of relevant "exemplary documents," the system builds a model of the topic investigated (a statistical language model, to be more exact, as we will see in the next section). The model can then act as a mediator 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, such as the World Wide Web. We conjecture that mediation through the right source collection has the potential to generate a very precise query and to increase significantly the quality of the retrieval effectiveness and the perceived completeness of the user's task.

Our model is somewhat similar to *relevance feedback* (RF) in that it allows the user to mark relevant documents, based on which the system builds or improves a representation of the user's information need (Harman, 1992). However, RF is a technique for *query reformulation*, and it typically offers insufficient support for completely novice users, who are unable to formulate an initial query. In con-

³Although widely used in the context of indexing or classification, the term "manual" refers, in fact, to an intellectual activity. Throughout this report, "manual" refers to an activity performed by a human, rather than a machine.

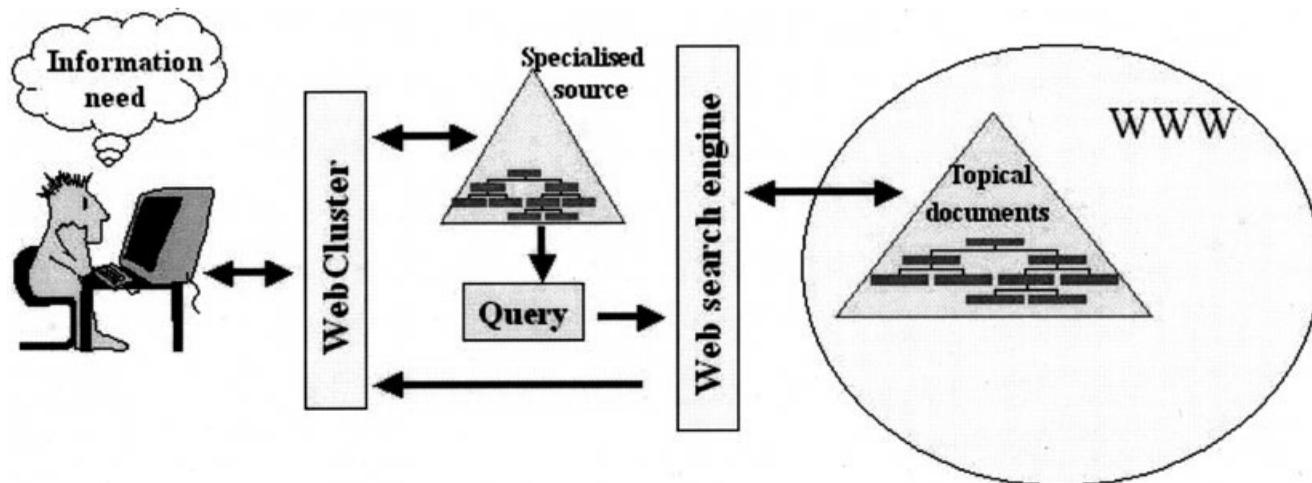


FIG. 1. Mediating access to the World Wide Web.

trast, mediated access proposes a new interaction paradigm, based on the user's interaction with a structured specialized collection, representative of the user's problem domain and rich in documents that can support the user's learning of the terminology, concepts, and topics of the domain.

The idea of *exemplary documents*, proposed by Blair and Kimbrough (2002), is also related to our approach. The source collection used by the mediation system to guide the user's exploration is a structured collection of exemplary documents representative of the domain of interest. During the exploration, the user selects relevant documents and clusters, providing the system with exemplary documents of her search topic.

Supporting that idea is Wittgenstein's theory of language acquisition, which states that language is acquired not by definitions and explanations alone but by having the terminology and expressions in question demonstrated in ordinary or typical use. This happens to fit very well the language model approach to representing documents and topics: A statistical language model of a topic can be built based on the exemplary documents selected as representative for it.

Let us close this overview of the mediated retrieval approach with a brief look at WebCluster's user interface, depicted in Figure 2. For exploring the source collection, our system allows the user to browse the structure and to select documents and clusters that are relevant, but it also affords querying, in case the user is able to formulate a query. The user can combine search strategies: Browsing can suggest search terms and has potential for serendipitously discovering relevant documents, while searching can reveal starting points for browsing. The overview panel (4) supports the browsing of the hierarchic, topical structure of the source collection. The currently selected document or cluster is shown in detail in the local view panel (5). The query panel (2) shows either the query specified by the user or the mediated query, automatically generated by the system based on the documents and clusters selected by the

user. The ranked view panel (6) displays the hits following a query-based search. If a hit is selected, the corresponding document is displayed in the local view and also highlighted in the overview, indicating a potential "pocket" of relevant documents. To complete the description of the user interface, the source collection panel (1) and the target collection panel (3) are used for selecting the source and the target collections.

The Source Collection

The source collection plays a crucial role in mediated access. An analysis of the part it plays in mediation reveals the characteristics of the ideal source collection. It should be large enough to be comprehensive relative to the domain of interest, but small enough to afford operations such as filtering, clustering, classifying, or sorting in reasonable time. It should have a clear topical structure to support exploration via a combination of browsing and searching, and it should be representative of the user's domain of interest, so that the user can learn the domain's terminology, its concepts and topics, and understand her problem and its context better. In practice, the ideal source collection may not always be available, so we have to examine possible choices of source collections for various domains of user interest.

One type of source collection is the *manually classified specialized collection* that covers the user's domain of interest. Specialized libraries and information organizations collect, classify, and maintain collections that cover various fields. For example, MEDLINE is a collection covering medicine, which is associated with a terminology model (Unified Medical Language System, UMLS) and a hierarchical classification of medical concepts (Medical Subject Headings, MeSH)⁴; Communications of the ACM (CACM)

⁴<http://www.nlm.nih.gov/pubs/factsheets/pubmed.html>

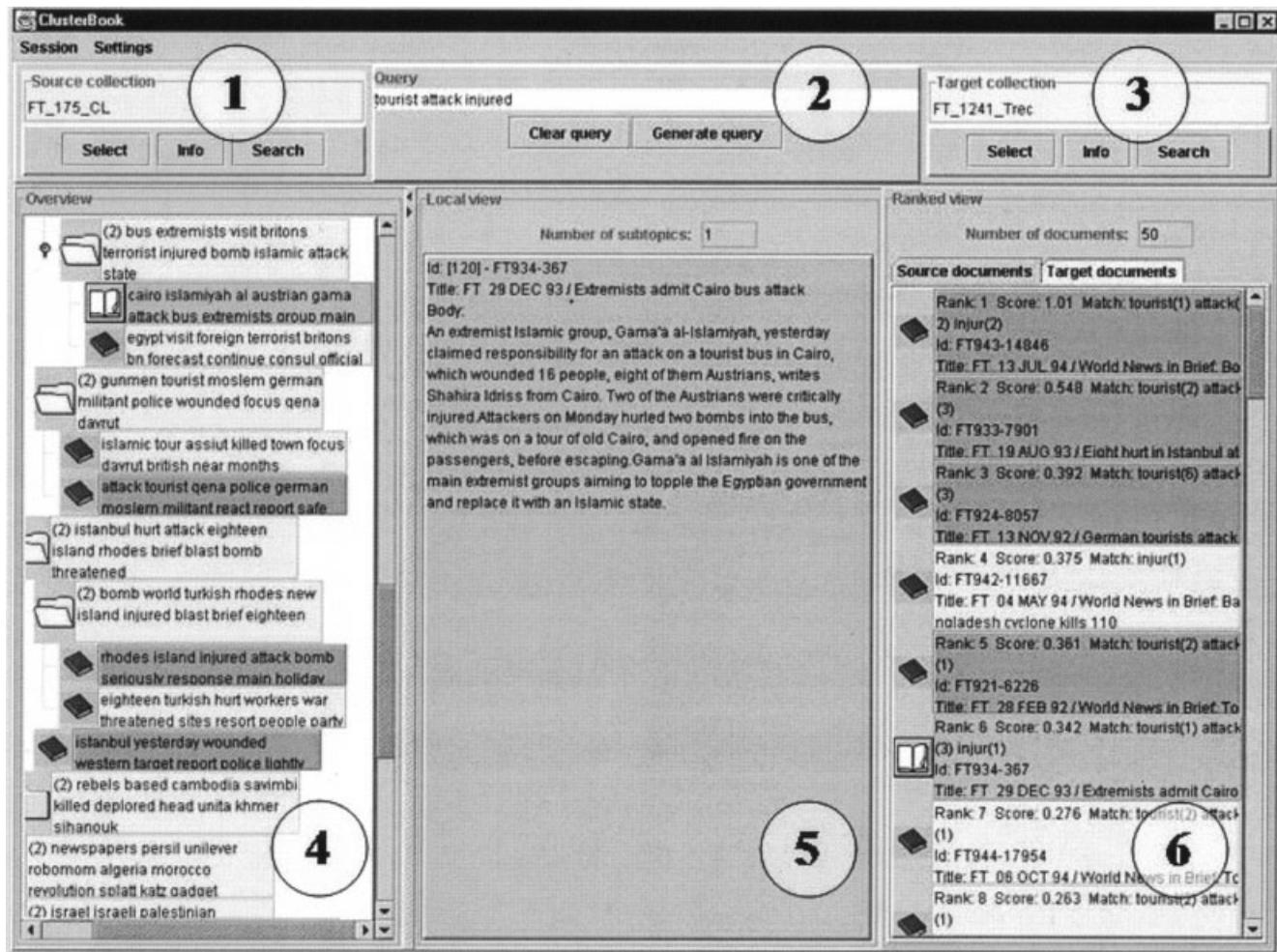


FIG. 2. One version of the mediated access user interface.

is a collection of articles on computing, with the associated ACM Computing Classification System.⁵

Specialized collections and taxonomies can be seen as loosely coupled: Although every specialized domain has a terminology and an underlying structure, an explicit ontology (or classification of the domain's concepts) may not exist. Some domains are well represented by specialized collections, but no ontologies are available to support the indexing and browsing of these collections. On the other hand, ontologies can be built for a domain by human experts based on their knowledge and experience without a certain document collection being available for support or exemplification. The consequence is that a classification system developed for a certain domain can be used to classify any document collection covering that domain. For example, MeSH can be used to classify any medical document collection (manually or automatically), not only MEDLINE. Alternatively, browsing an ontology such as MeSH can be used for exploring the medical domain and for mediation if no appropriate source collection can be found.

⁵<http://www.acm.org/class>

If no ontology and no manually classified representative collection are available for a certain domain, an alternative approach to structuring a collection in order to build a source collection for mediation is *document clustering*. For example, Cranfield is a collection of abstracts on aerodynamics, RAPRA covers polymers, and CABI covers agriculture issues. These collections can be clustered in order to reveal the topical structure of their domains and to support exploration of the documents they contain. In the process of clustering, support tools like an index or a thesaurus can be built for supporting exploration of the vocabulary and of the topical structure of the domain. Some critical issues are the choice of clustering method and parameters, the similarity measure between documents, as well as the method for creating cluster labels that are representative and accurate and have discriminatory power.

A mixed approach to building the source collection based on a document collection can also be imagined. In the first stage, a clustering algorithm is used in order to automatically group documents based on reciprocal similarities, as estimated by the system. In a second step, human experts can adjust the obtained structure by exploring the structure

and moving to the right cluster the documents whose semantic content does not fit their place, or making copies for documents that should belong to more than one cluster.

If no explicit source collection is available, one can be built based on the target collection. A possible approach is to produce a sample of the target collection that is representative enough and covers all the subdomains, topics, and concepts of the domain, but at the same time is small enough to afford manual or automatic classification and exploration through a combination of searching and browsing. Another approach is to apply an initial user query as a filter and to classify the obtained source collection on the fly. Consideration must be given to the fact that the user may not have a clear information need or a good grasp of the domain vocabulary. Therefore, the filtering should be rather generous, including in the retrieved set even documents with a low estimated relevance for the initial query. The user should also be encouraged to supply as many words as possible, or a thesaurus should be used for query expansion.

Structuring the Source Collection and the Aspectual Cluster Hypothesis

There is quite a rich literature on interactive information retrieval systems based on manually classified collections (Hersh, Leone, & Hickam, 1994; Pollitt, 1997; Pollitt, 1998; Pratt, Hearst, & Fagan, 1999; Pratt, 1999). On the other hand, document clustering has been studied mostly in the context of batch retrieval (Willett, 1988) or as a tool for organizing search results (Hearst & Pedersen, 1996; Leuski, 2001; Zamir & Etzioni, 1999; Zamir, Etzioni, Madani, & Karp, 1997). There is little understanding of the power of clustering to reveal the topical structure of a document collection and to guide exploration in an interactive setting. To fill this gap, we focused on the use of document clustering in an interactive information retrieval system for exploratory tasks and have explored the potential of document clustering for structuring source collections for mediated retrieval.

The expected usefulness of document clustering is based on the *cluster hypothesis*, i.e., on the expectation that “closely associated documents tend to be relevant to the same requests” (Jardine & van Rijsbergen, 1971). Most researchers whose work was based on this hypothesis and who tried to evaluate its validity assumed a reciprocal (bi-directional) relationship between similarity and relevance, i.e., similar documents are expected to be relevant to the same queries and documents relevant to the same queries are expected to be similar. This assumption is implicit in the overlap test proposed by van Rijsbergen and Sparck Jones (1973) and is made explicit by El-Hamdouchi and Willett (1987), (“dissimilar documents are unlikely to be relevant to the same requests”), as well as by Hearst and Pedersen (1996), (“relevant documents tend to be more similar to each other than to non-relevant documents”).

Consequently, there has been little or no distinction made between experiments attempting to show that similar documents tend to be relevant to the same topics and experiments testing whether documents relevant to the same topics are highly similar. Moreover, most experiments on cluster-based retrieval have looked at the distribution of topical documents (i.e., documents that are about a certain topic) over the cluster structure, with no or little distinction between:

- the cluster hypothesis, i.e., the relationship between document similarity and relevance to the same topic, and
- the capacity of a clustering algorithm to group similar documents.

Some researchers wrongly perceive the cluster hypothesis and document clustering as one issue. If a clustering algorithm does not group together the documents relevant to a topic, they conclude that the cluster hypothesis does not hold, without further investigating whether the particular clustering algorithm was unable to group together similar documents or whether for that particular document collection and set of topics the topical documents are not highly similar. There is little surprise, therefore, that such experiments have produced inconsistent results and have undermined the usefulness of document clustering (Shaw, Burgin, & Howell, 1997).

Informal experiments with our proof-of-concept system, WebCluster, on the Reuters collection suggested a nonreciprocal relationship and led us to propose a relaxed version of the cluster hypothesis, the *aspectual cluster hypothesis*:

Similar documents tend to be relevant to the same requests, but documents relevant to the same requests are not necessarily similar. They tend to be dissimilar if they cover different aspects of the same complex topic.

Our experiments also paint a slightly different picture of document retrieval compared to the one generally found in IR literature. Let us call *features* the sets of terms (or keywords) representative for a certain *topic* (or *aspect* of a topic, for complex topics). Documents are represented by features, the interdocument similarity is computed based on features, and the clustering is generated based on features. Based on their contribution (or weight), one can distinguish between major features, which determine the major axes or higher level clusters of a hierarchical structure, and minor features, which determine the minor axes or bottom level clusters of the hierarchy. For example, clustering a subcollection of Reuters news articles produces two major clusters of documents that refer to the former U.S. President Reagan: one about the Iranian arms deal, and one about U.S. agricultural policies. In both cases, “President Reagan” is just a minor feature.

A query that matches a major feature (“Iranian arms deal”) is very likely to hit a major cluster, in which most relevant documents are grouped together, so cluster-based

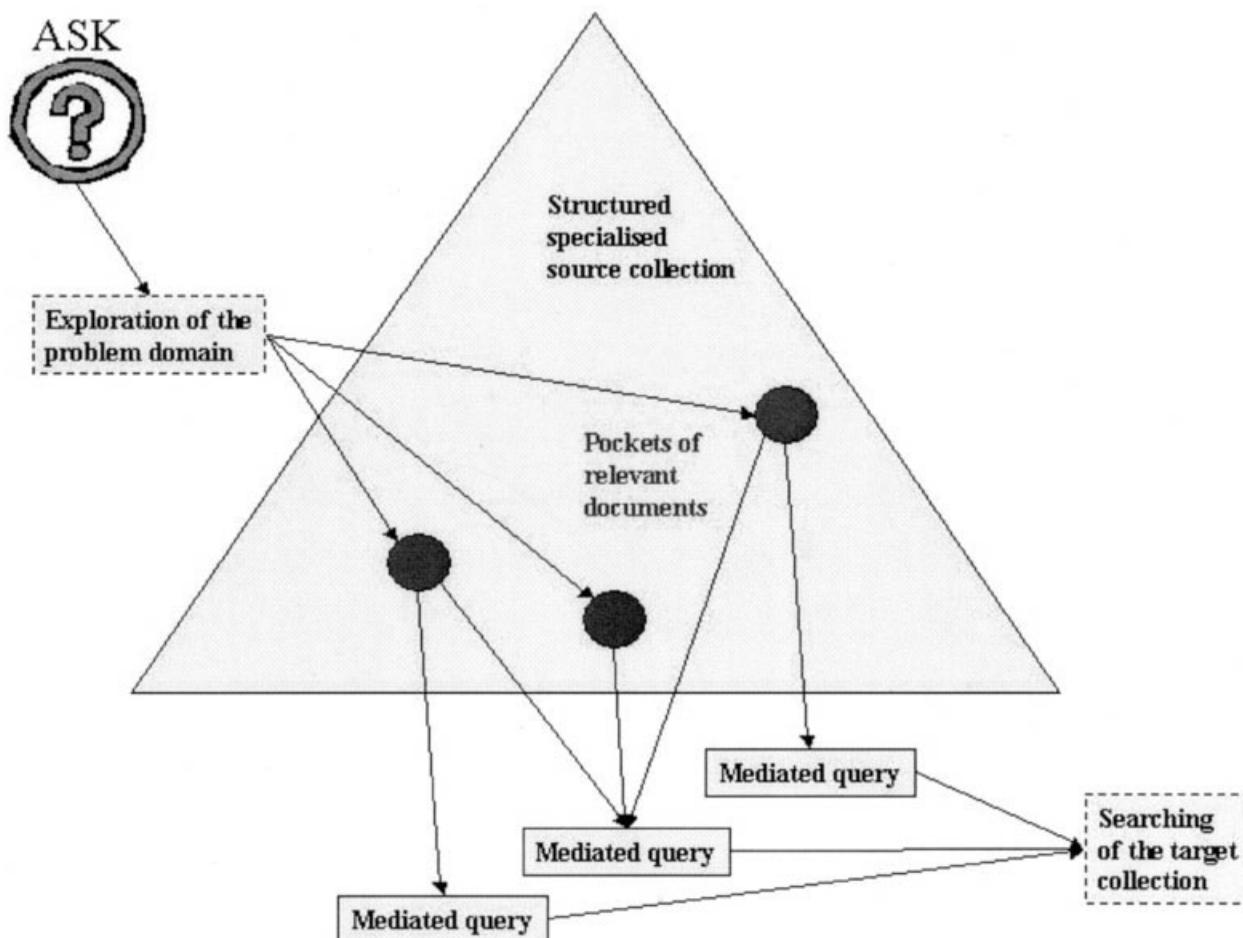


FIG. 3. Aspects of relevance in the mediated access process.

retrieval gives good results. On the other hand, if the query matches a minor feature (“President Reagan”), then the relevant documents are indeed grouped into small subclusters but spread over the collection. The documents about President Reagan in the “Iranian arms deal” cluster are quite similar to each other, but dissimilar to the ones in the “U.S. agriculture” cluster. In this case, classic cluster-based retrieval, which retrieves just one cluster, would do badly. A relaxed version of cluster-based retrieval, more appropriate in an interactive context, would rank the clusters of the structure based on their estimated quality and allow the user to explore them.

Our relaxed (aspectual) cluster hypothesis is consistent with both positive and negative results of traditional experiments on the original cluster hypothesis. Early experiments used small test collections and very simple and focused requests. In such cases, with topics that had only one or a small number of aspects, the cluster hypothesis experiments were successful. Later experiments, with larger collections and more complex topics, were less successful. This is exactly what our hypothesis would predict.

WebCluster takes into account this new view of clustering by allowing the user to collect (or berrypick) (Bates, 1989) and bookmark subclusters of interest from pockets of

relevance. We will call this *aspectual retrieval*, as it usually identifies aspects of a topic or of a request.

Figure 3, detailing the first stage of mediated access, i.e., the exploration of the source collection, illustrates this issue: Relevance is often aspectual, in the sense that a document can be relevant to a query for a variety of reasons, potentially responding to various aspects of an information need. For example, documents about ethnic violence in Southeast Asia, documents about the Indian Ocean climate, and documents about the Passport Agency’s problems, although very dissimilar, may all be very relevant for someone planning their vacation.

This issue applies to the generation of the mediated query and also to the design of the user interface. Imagine the user who wants to go on holiday. She may not be aware of problems at the Passport Agency, of safety issues (either because of violence or because of some disease outbreak), of the monsoon period, or of some cheap deals. However, a general query such as “holiday Asia” should point the user to all these various aspects that should be of interest to her, so that she explores all potentially useful pockets of information in the source collection.

When a set of documents or clusters is marked as relevant by the user, these selected items may be in the same

area of the cluster hierarchy, presumably referring to the same concepts, in which case a unique mediated query needs to be generated. However, the generation of the mediated queries is problematic if the exemplars are spread over several areas of the hierarchy, referring to different concepts or aspects. It is debatable in this case whether the topic would be better captured by a general query, linking all the different aspects of the topic, or by a set of precise queries, each referring to a certain aspect of the topic. We would intuitively expect the former approach to reveal relationships between various aspects of a topic and to generate higher recall, and the latter to better focus on individual aspects and to generate better precision of retrieval. This expectation needs to be tested. The experimental results, presented later in this report, will influence the design of the user interface and the underlying model to be used in the operational mediation system, as well as the design of further mediation experiments.

A Discussion of the Mediated Access Paradigm

The particular design decisions taken for implementing WebCluster may wrongly convey the idea that these decisions are inherent constraints of mediation and may create an incorrect, narrow view of the mediation paradigm. We address this potential misconception below.

The core of the mediated access model is simple: The user explores a structured, specialized source collection indicating, in the process, relevant documents and clusters. One outcome is that the user becomes more familiar with the problem domain and more aware of the details of her information need. The other outcome is that the system analyzes the exemplars identified by the searcher and proposes a mediated query, which she can use to search a target collection. Based on this simple model, a variety of semi-independent choices can be made with regard to different aspects of mediation:

Structuring the source collection: We chose document clustering because it is fully automatic and domain independent, and it allows flexibility in choosing parameters (such as the indexing method, the weighting scheme, and the clustering algorithm) that meet particular needs. Moreover, it gave us the opportunity to conduct experiments on the cluster hypothesis. However, any classification method, manual, semi-automatic (supervised) or fully automatic (unsupervised), can be employed as long as it reveals the topical structure of the source collection and thus supports the exploration of the problem domain.

Exploring the structured source collection: We chose the folder metaphor for representing the structured source collection in the user interface because of its simple implementation and the computer users' familiarity with it. We also combined this structured view of the domain with a linear view, obtained through query-based ranked searching, in order to offer a rich set of retrieval strategies. However, alternative visualization tools such as hyperbolic trees, thematic maps, tree maps, or cone trees

can be used instead, and support for other search strategies can be offered. The constraint is that the user interface should support concept learning and the identification of exemplary relevant documents.

Explicit vs. implicit relevance judgments: We chose explicit relevance judgments because, despite the user interface complexity that it introduces and the need for user cooperation, it indicates the user's preferences more clearly. Implicit feedback, which relies on cues from the user's behavior and actions, is currently unreliable in conveying the user's interest (Kelly & Belkin, 2001). However, when technology that supports it matures, that approach will also be applicable to mediation.

The generation of the mediated query: Language models were our choice due to their power, flexibility, and uniform treatment of documents and clusters of documents. Alternatively, any techniques developed for query expansion based on relevance feedback or any machine learning techniques can be used just as well.

The interaction model and metaphor: The traditional library was used as model and metaphor for our implementation of mediated access because of people's familiarity with it and also because our system's mediation role is similar to that of the librarian. For the new generation of information seekers, arguably more familiar with digital libraries and hypermedia than with the traditional library, a different metaphor may be more appropriate. For example, an electronic encyclopedia could support a wide variety of information retrieval strategies and the mediated query could be used to expand the quest for information to the Web.

This discussion should reveal the clear distinction between mediated access, the novel paradigm that we propose, and relevance feedback (RF). RF is a technique for iteratively improving the query and, implicitly, the retrieved set of documents during an interactive search session on a target collection. It relies on the user having a minimum knowledge of the domain investigated and on her being able to generate a reasonably good initial query.

In contrast, mediated access proposes a new interaction model. It relies on the user interacting with a structured specialized source collection, representative of the user's problem domain and rich in documents that can support the user's learning of the terminology, concepts, and topics of the domain. It also relies on the user finding sufficient exemplary documents to solve her problem or to clearly convey her information need so that a high-quality mediated query, which balances representation and discrimination, can be generated for searching the target collection.

Topic Models

The Generic Approach

The conceptual model of mediation access relies on the user making relevance judgments with regards to documents and clusters of the source collection, and thus conveying her topic of interest. As depicted in Figure 4, the user

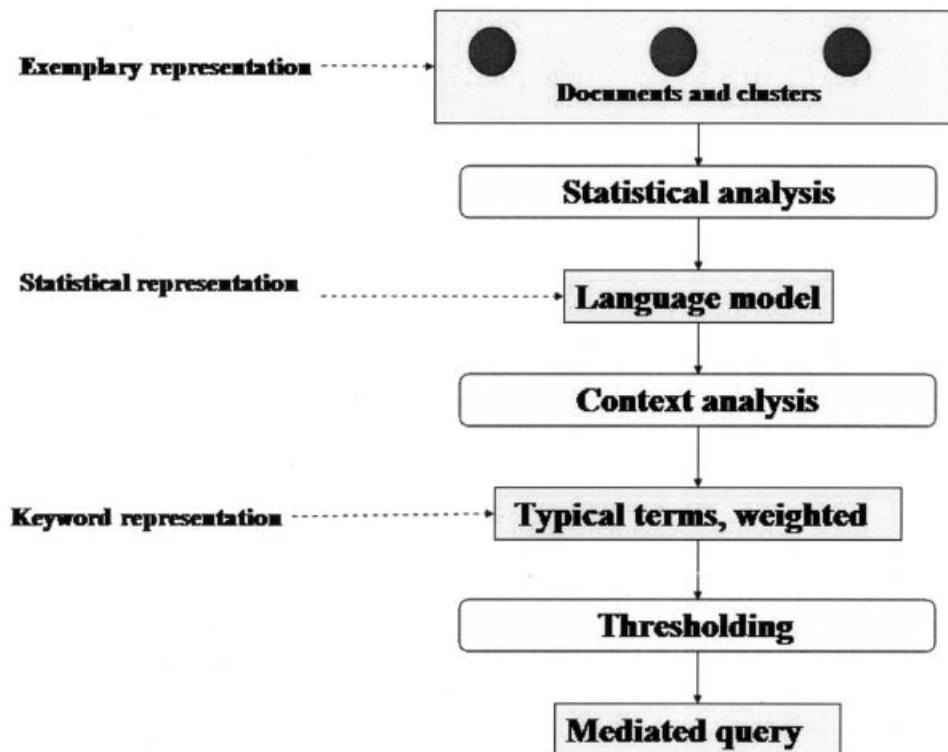


FIG. 4. Topic model representations.

provides an exemplary representation of the topic of interest, which consists of documents and clusters of documents that are typical for the topic investigated. Based on these exemplars, the system derives the mediated query. Let us discuss in some detail the steps of our approach. More complete explanations of this model are available in Muresan (2002).

Statistical analysis: It means counting the appearances of each term in the documents and clusters selected by the user, as well as in the documents and clusters of the source collection; the choice of the data structures and the algorithm employed are not relevant. As statistical representation, the language model consists of frequency distributions of each term in the topical documents and clusters.

Context analysis: The language model of the exemplars is compared to that of the other documents and clusters in the source collection; each term is assigned a weight based on its “specificity” to the topic. The keyword representation consists of the terms that discriminate the topic in the given context, ranked based on their contribution to the topic description. This is the more interesting step of the algorithm, so it is presented in more detail in the rest of this section.

Thresholding for the mediated query: Query size thresholds can be applied based on search engine specifications or on efficiency considerations (search time increases with query length). Thresholds on query term weights can also be imposed, based on the context: Terms that are not highly specific to the topic may improve recall but will probably degrade precision.

Apart from generating the mediated query in response to the user’s selection of exemplars, the mediation system should also support the user’s exploration of the source collection via a combination of browsing and searching. While browsing, the user relies on cluster and document labels (or representatives) to navigate the structured collection and to find promising documents. When a search is submitted, the system ranks documents and/or clusters based on the match between their representation and the query.

It is apparent that we are confronted with the need to build cluster and document representations for browsing and for searching, and topic representations for mediation. Conceptually, the mediated queries that represent topics of interest to the user are similar to document and cluster labels, as they need to represent those documents or clusters. Therefore, we have developed an integrated approach to generating document, cluster, collection, or topic representatives. We view the cluster as the generic content container that can model a document (singleton cluster), a cluster of documents, a topic (a set of exemplary documents), or a collection (root cluster), so we will address the generation of cluster representatives.

Our approach is based on the Kullback-Liebler (KL) *divergence* or *relative entropy*, which indicates how different two probability distributions are (Manning & Schütze, 1999, p. 72). For each term t_i present in the cluster *Clust* we compute the probability of the term being selected by a random sampling of the cluster:

$$P_{i,Clust} = \frac{\text{number of occurrences of term } t_i \text{ in Clust}}{\text{total number of term occurrences in Clust}}$$

and estimate the probability of t_i being selected from the reference collection of documents that represents the context:

$$P_{i,Ref} = \frac{\text{number of occurrences of term } t_i \text{ in Ref}}{\text{total number of term occurrences in Ref}}$$

Then the relative specificity of term t_i in Clust, compared to the reference collection Ref,

$$KL_i = p_{i,Clust} \log \frac{P_{i,Clust}}{P_{i,Ref}}, \quad (1)$$

measures the contribution of term t_i to the Kullback-Liebler divergence between the cluster Clust and its context. The terms with positive KL_i values are more specific to Clust than to Ref: they are included in the cluster representative and are ranked based on their KL_i value in order to identify the most typical terms. It is apparent, therefore, that the cluster representative depends on the context.⁶ The following subsections discuss the choice of the appropriate information context for a certain kind of cluster representative.

Relative Cluster Representative for Browsing

Imagine a user browsing the hierarchic cluster structure. In order to decide which of the subclusters of the current cluster is worth expanding for further exploration, she needs to know what is specific about each subcluster. For that, she relies on the cluster labels displayed by the user interface. Therefore, the browsing or relative label of each cluster needs to indicate in what way the cluster differs from its parent. This suggests the use of the parent cluster *Parent* as the reference collection in equation 1:

$$R_i = p_{i,Clust} \log \frac{P_{i,Clust}}{P_{i,Parent}}. \quad (2)$$

This weight indicates the relative specificity of each term in the cluster, compared to the parent cluster. The terms with negative weight are ignored (they are not specific) and the remaining terms are ranked according to their R_i weights in order to generate the browsing label, or relative representative.

Absolute Cluster Representative for Searching

When searching the source collection, based on a query submitted by the user, the system needs to find the cluster

⁶Context is typically a collection or set of clusters from which a certain cluster needs to be distinguished, based on its representative.

that best matches the query and, therefore, the representative needs to distinguish each cluster from the rest of the collection. This suggests the use of the full source collection *Source* as the reference collection in equation 1 in order to generate the *searching* or *absolute representative* of a cluster:

$$A_i = p_{i,Clust} \log \frac{P_{i,Clust}}{P_{i,Source}}. \quad (3)$$

Additionally, we introduce the option of taking into account the uniformity of the term distribution in a cluster by multiplying the above formula with a *uniformity factor*. For example, consider a cluster with 10 documents. A term t_1 may appear once in every document, while another term t_2 may appear 10 times in one document and not at all in the other documents. While both terms have the same frequency, t_1 is more uniformly spread, which may indicate that it is more typical for the cluster. We propose the following uniformity factor:

$$u_i = \frac{1}{1 + k \cdot \sigma_i},$$

where σ_i is the standard deviation of the term frequency over the documents of the cluster and $k \geq 0$ a parameter that can be set to indicate how important uniformity is for specificity (if $k = 0$, then $u_i = 1$, so there is no influence).

The terms with negative weight are ignored and the remaining terms are ranked according to their A_i weights (optionally multiplied by u_i) in order to generate the absolute representative.

Expanded Cluster Representative for Mediation

The source collection used for mediation is hierarchically structured, with smaller and more specific topics included in larger and more general topics. Therefore, in order to accurately convey the user's topic of interest, we build the *mediation* or *expanded representative* of a selected cluster by summing gradually reduced contributions of the absolute representative of the cluster, of its parent, and of all the clusters on the path to the root of the structure. The weight of term t_i in the expanded representative is:

$$E_i = (1 - w) \cdot A_{i,0} + (1 - w) \cdot w \cdot A_{i,1} + \\ (1 - w) \cdot w^2 \cdot A_{i,2} + \dots + \\ (1 - w) \cdot w^{r-1} \cdot A_{i,r-1} + w^r \cdot A_{i,r}$$

where $A_{i,0}, A_{i,1}, \dots, A_{i,r}$ are the weights of t_i in the absolute representative of the chosen cluster, its parent, ..., the root cluster, and $w \in [0, 1]$ is the decay rate of the contribution as the context goes from specific to general. For example, for $w = 0.1$, the contribution of the current cluster to the

term weight is 0.9, of its parent 0.09, and so on. These factors sum up to 1.0.

When the user selects a single cluster as an exemplar, there is a choice of using the A_i weights (for efficiency) or the E_i weights (for a more accurate representation, which performs context disambiguation). If several clusters are selected by the user, taking the context into account may become too complex; the easy option is to merge the selected clusters into a (conceptual) *topic cluster* and to compute A_i weights for its terms.

Comments

Various researchers have used different formulae for generating a “one for all purposes” cluster label for representing a cluster. We have proposed a novel approach based on the fact that the label needs to distinguish a cluster in a certain context. For example, when the user is browsing, the cluster needs to be distinguished from its parent and siblings; when the user employs cluster-based searching, the best cluster needs to be distinguished from the rest of the collection; when the user selects a cluster as an exemplar for mediation, its representative needs to capture its topic and its context. We have, therefore, pioneered the use of multiple cluster representatives, based on their use.

The advantage of using the KL formula is threefold. Firstly, it allows documents and clusters to be treated similarly, as bags of terms, so that 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, which offers the flexibility of a variable representation, dependent on context.

The modeling approach described in this section is expected to be typical for the mediated interaction. Shortcuts are, however, acceptable: as the user explores the source collection and learns the terminology and concepts of the problem domain, she may become confident enough to generate the mediated query herself, rather than rely on the system completely.

Experiments

Experimental Setting

We are proposing a new interaction model for information access and, therefore, reporting a formal user study is a natural expectation. However, a difficulty in evaluating such an interactive system and the underlying conceptual model is the fact that the interactive model proposed is quite flexible and can be “enacted” by a variety of user search strategies. A user experiment would contain, at this stage, 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 look most promising in terms of retrieval effec-

tiveness. These optimal combinations will be used in future user experiments.

The purpose of the simulations reported here was mainly to compute upperbounds of performance for various search strategies in order to estimate the potential of mediation and to establish guidelines as to which strategies are more likely to be successful with real users. The influence of various parameters that are incorporated in a mediation system can also be estimated, so that an operational system can take the optimal values. Upperbound measures, expected to be reached by the “ideal user,” are not the only performance values that can be estimated through simulations; the behavior of the average user in a realistic situation, following certain guidelines, can also be simulated.

Building an experimental setting that supports testing our aspectual cluster hypothesis, simulations of mediated searches and, in the future, real user experiments, is not trivial. With these constraints, we resolved to adapt the setting of the interactive TREC-8 experiment⁷ which was designed to investigate the exploration of complex information needs, with a multitude of aspects, i.e., the very situation when mediation is expected to help the user. There are relevance judgments provided for six topics:

1. 408, “tropical storms”
2. 414, “Cuba, sugar, imports”
3. 428, “declining birth rates”
4. 431, “robotic technology”
5. 438, “tourism, increase”
6. 446, “tourists, violence”

Unlike most other test environments, this one has aspectual relevance judgments: Distinct aspects of each topic are identified and the judgments specify for which aspects of each topic a document is relevant. These judgments were invaluable for supporting aspectual cluster hypothesis experiments, and are also adequate for mediation simulations, as they convey the aspectual coverage (or aspectual recall) of retrieval. Unfortunately, the number of topics available for this test collection is rather small, so the results obtained cannot be safely generalized. However, they represent a first step in investigating the potential of system-based mediation.

We used a TREC subcollection, the Financial Times (FT) of London, containing 210,158 news articles from 1991–1994, as the target collection for our mediation experiments. However, no specialized source collection was available to support mediation. We simulated the specialized source collection by artificially building a smaller subcollection that covers all the test topics and contains a relatively high number of relevant documents, but also a number of nonrelevant documents. We included in the source collection half of the 350 documents judged relevant by assessors in order to support the user’s discovery and

⁷<http://www-nlpir.nist.gov/projects/t8i/t8i.html>

selection of relevant exemplary documents. In other words, these 175 relevant documents were used for training the topic models. The testing of the topic models consisted of searching the target collection for the other 175 relevant documents.

We made the mediation task more difficult by “polluting” the source collection with copies of the 572 documents judged nonrelevant by assessors (the originals remained in the target collection in order to provide near-miss hits and to make the retrieval task more difficult). Due to NIST’s pooling method of assessing relevance, we could assume that these 572 documents had some similarity with the topic descriptions and with the relevant documents (otherwise they would not have been submitted by TREC participants). Therefore, our source collection was a realistic simulation of a specialized source collection: Most documents were in the domains of the test topics; some were relevant to the topics, some were not. The existence of the 572 officially nonrelevant documents (but judged relevant by some users) both in the source and in the target collections makes the experimental setting more realistic and the retrieval task more difficult.

The next subsections discuss a number of experiments that we conducted. We start by briefly analyzing some results of our quite extensive cluster hypothesis experiments described in detail in Muresan (2002). Then we look at the results of experiments simulating various mediation strategies.

Clustering Experiments

Clustering works by attempting to group together documents that have some degree of similarity, usually with regards to their term frequency distributions. Intuitively, documents that are highly similar are expected to cover the same topic and, consequently, clustering is expected to reveal the topics of a collection by grouping together similar documents. Therefore, our set of tests had two stages, in which we investigated:

1. The collection classifiability: the potential of the document collection to be structured. We investigated the *aspectual cluster hypothesis*, i.e., we looked for a correlation between documents covering the same topic and being highly similar.
2. The collection clusterability: the *cluster hypothesis consequence*. We investigated the quality of the structure built by concrete clustering algorithms in terms of grouping together topical documents.

Because clustering uses as input interdocument similarities, usually based on lexical content and on statistical analysis of term distribution in the documents, it is quite obvious that the classifiability of a collection is a prerequisite for its clusterability. If there was no correlation between the interdocument topical commonality and their lexical content similarity, clustering could not be expected to group

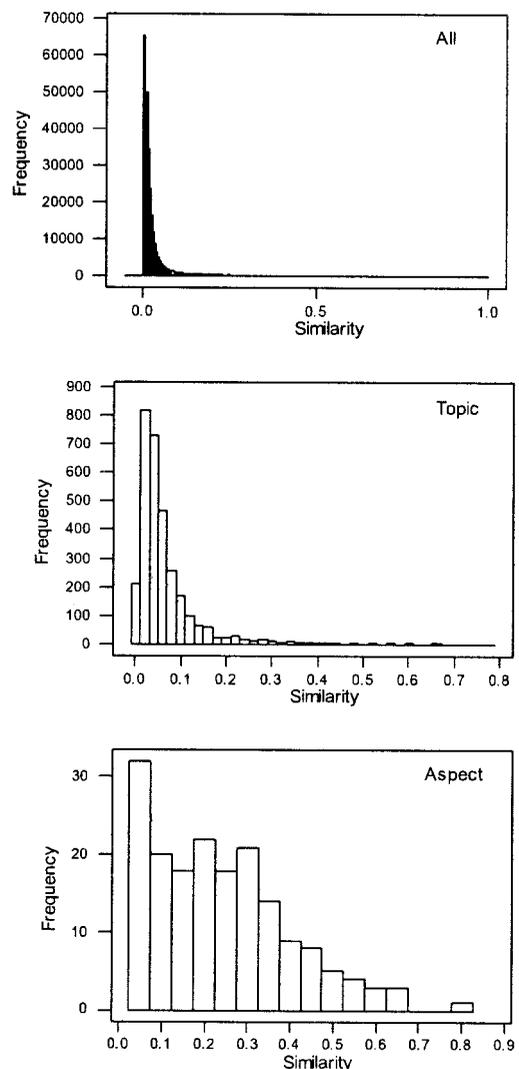


FIG. 5. Comparison in histogram form of the distributions of all similarities, topical similarities, and aspectual similarities.

together semantically similar documents and to identify topics.

Our clustering experiments were conducted on the source collection built for the mediation experiment. First, the collection classifiability was tested based on our aspectual version of the original separation test: We computed the similarities between each pair of documents (“all similarities”), between each pair of documents relevant to the same topic (“topical similarities”), and between each pair of documents relevant to the same aspect of a topic (“aspectual similarities”). The expectation was that aspectual similarities should be, on average, significantly higher than topical similarities, which should be significantly higher than all similarities. The computations were repeated with two common similarity measures, Cosine and Dice (Ellis, Furner-Hines, & Willett, 1993), and three different weighting schemes for document terms: *relative frequency* (term frequency over the number of tokens in the document), *tf-idf* (in the inquiry form) (Callan, Croft, & Harding, 1992), and

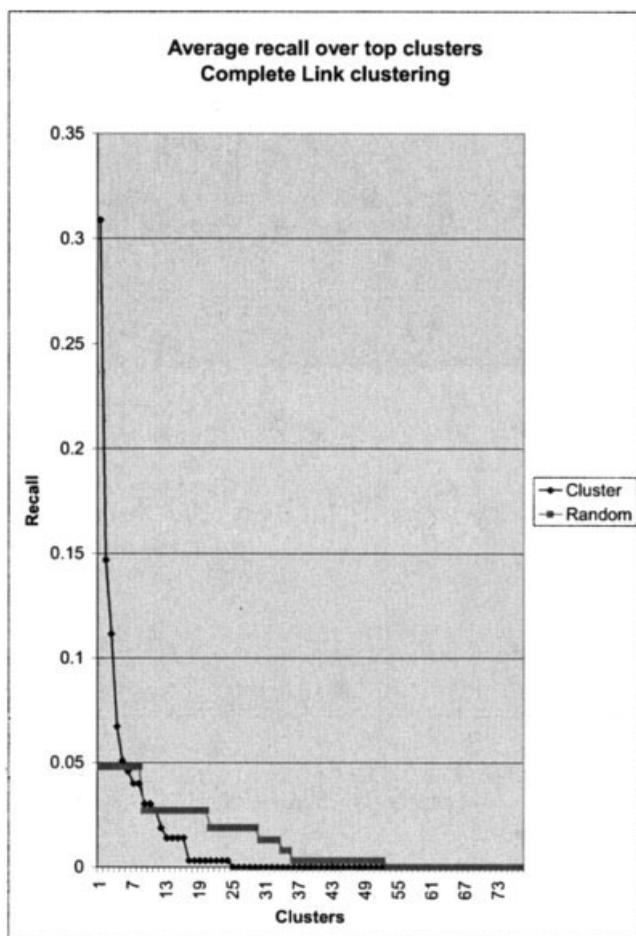


FIG. 6. The distribution of the relevant documents over top-level clusters. Complete-Link compared to random distribution.

the Kullback-Liebler (KL) *measure of divergence* (applied between the document and the source collection probability distributions).

The histograms in Figure 5 show the distribution of the 278,631 “all similarities,” of the 3,073 “topical similarities,” and of the 178 “aspectual similarities” for the Cosine-KL combination (but similar results were obtained for the other combinations of parameters). They are all positively skewed, showing that relatively few pairs of documents are highly similar. It is clear, based on an ANOVA statistical analysis ($p < .01$), that the experimental results confirm our predictions: documents that cover the same aspect of a topic tend to be more similar to each other than documents that cover the same topic, which tend to be more similar than randomly selected documents in the collection.

Moreover, our aspectual cluster hypothesis seems to be confirmed by the significant number of pairs of topical documents that have the similarity close to zero (or under a rather low threshold): Documents that share a common topic may be very dissimilar if they cover different aspects of the topic.

We then looked at how well clustering algorithms can recognize topics and separate them into clusters. Figure 6

shows the distribution of relevant documents over top-level clusters obtained with the Complete-Link algorithm, compared with a random allocation of relevant documents to clusters (clusters are ranked in decreasing order of the number of relevant document they contain). Figure 6 suggests a result that was consistent when the clustering methods and the similarity measures employed were varied: most relevant documents tend to be grouped in a small number of “good” clusters. Based on the aspectual relevance feedback associated with the test collection, a closer examination of the distribution of relevant documents in the cluster structure reveals that the documents that are grouped together in high quality clusters (i.e., highly topical) appear to be documents that share at least one aspect of the topic. They often tend to be documents that cover more than one aspect, which makes them more likely to be similar to other topical documents.

In conclusion, our experiments provided experimental support for the aspectual cluster hypothesis:

Highly similar documents tend to be relevant to the same topic. Documents relevant to the same topic may be quite dissimilar if they cover distinct aspects of the topic.

Its consequence:

Clustering algorithms tend to group together documents that cover highly focused topics, or aspects of complex topics. Documents covering distinct aspects of complex topics tend to be spread over the cluster structure.

This result has important implications for research in document clustering. Over the years, results of clustering experiments relying on the cluster hypothesis have been inconsistent. The only investigation known to date that tried to explain such inconsistencies is Sparck Jones’s (1973), which looked at statistical differences between document collections on which the cluster hypothesis tests were conflicting. While her experiments failed to find any correlation between collection statistics (such as number of terms per document, per collection, and per test query) and classifiability, perhaps investigating the semantic complexity of test topics and of documents would prove more successful.

There are even farther-reaching implications for IR in general and in particular for the design of experiments. Some experiments such as ad-hoc TREC have attempted to use relevance feedback and other query reformulation techniques in order to build the one query that achieves the best retrieval performance. For complex topics, with distinct aspects, this may be the wrong approach. No single query may be able to cover all the aspects of the topic and also achieve good precision. It may better to generate a set of queries, one for each aspect.

There are also implications for mediation: we achieved a better understanding of the way relevant documents are spread over the cluster structure and of the potential of clustering for exploring a document collection. A practical

TABLE 1. Residual effectiveness of searching the target collection, baseline for mediation experiments.

Topic form	Recall	RelAspR	R-Precision	AUP
Title	0.895873	0.949383	0.128600	0.098780
Description	0.972222	0.993827	0.098628	0.078423
Full	0.972222	0.993827	0.105492	0.099499

result of the experiments conducted is that, for the collection explored, we have an idea of the typical number and distribution of pockets of relevance that need to be found in order to assure a good coverage of a topic.

Baseline for Mediation Simulations

Mediation is proposed as an approach to improving retrieval effectiveness based on improving the query submitted to the search engine rather than the search algorithm. Our approach is successful if the queries generated through mediation produce better search effectiveness on the target collection than the queries generated by the user when no mediation is employed. In other words, we need to compare the effectiveness of mediated searches with a baseline search.

During informal experiments with our mediation system, users almost invariably simply extracted terms from the title or the description of the topics in order to formulate the initial query. Many of them did not reformulate the query. Therefore, the titles and the descriptions of the topics constitute a reasonable baseline for measuring query quality in comparison to various mediated queries.

The search of the target collection consisted in ranking the documents in the target collection based on the simple dot product between their vectorial representation and that of the query. Two independent variables were used:

1. The form of the topic used for deriving a query. The three cases considered were when only the topic title was used (“Title”), when only the topic description was used (“Description”), and when their combination was considered (“Full”).
2. The weighting scheme used for generating the term weights of the document representations. The three different schemes tested were relative frequency (“Rel-Freq”), tf-idf in the form used by Inquery (“TfIdf”), and Kullback-Liebler (“KL”).

TfIdf and KL did not produce a statistically significant difference in retrieval effectiveness, but were both significantly better than RelFreq. In order to save space in this report, the result tables only show effectiveness measures when the TfIdf weighting scheme was used; however, the reported statistical results take into account searches based on all three weighting schemes.

Table 1 shows the baseline values, obtained by searching the target collection based on queries derived from the topic description; the term “residual effectiveness” is used be-

cause the relevant documents also present in the source, used for training the model, were not counted as valid hits. No cutoff was applied; all documents with a nonzero score were retrieved. We computed effectiveness measures common in TREC: recall (R), aspectual recall relative to the number of aspects covered by the target relevant documents (RelAspR), R-precision (precision at a cut-off equal to the number of relevant documents), and average uninterpolated precision (AUP).

An analysis of variance⁸ shows a highly significant influence of the topic form on R and RelAspR ($p < .01$), indicating that some relevant documents match the context description, but not the title of the topic. This result confirms one of the intuitive ideas behind mediation: Enriching the query with some context can significantly improve recall (both absolute and aspectual). This is important for tasks where recall is essential (for example finding previous court cases) and also for exploratory searches, when a user unfamiliar with the domain may find it difficult to produce a query that comprehensively conveys her information need.

It is worth noting that R and RelAspR are quite high on both source and target searches, indicating that most relevant documents contain the keywords used for describing the topic. However, the values are < 1 , which indicates that there are relevant documents that do not contain any of the terms used for describing the topic. For example, in the case of the first topic, the results show that there are documents about tropical storms that have caused property damage or loss of life, but do not contain any of the words “tropical,” “storm,” “property,” “damage,” “loss,” and “life.” We would expect that through mediation even such high levels of recall can be improved.

A somewhat surprising result was obtained when examining the effect of the topic form on precision. We had expected that using an extended description of the topic for producing a query would generate better precision (because more context was being provided) and that using the full description (combination of title and extended description) would improve precision even more (because context terms were being used, but the title terms were being given higher weight). The analysis of variance indicates a consistent, although not quite significant, advantage of using just the title. The full description is slightly worse, while the simple description comes a more distant third. An explanation for this result is apparent from more closely examining the topics and the relevant documents: the context offered by the topic description is worded so that it would help a human decide whether a certain document is relevant or not, but it does not offer, in general, terms that are expected to be found in the relevant documents. In other words, some of the descriptive terms are not content-bearing. So, the challenge for mediation is to produce a query that has a high

⁸Throughout this report, we used an ANOVA test for the statistical analysis of variance when the data displayed normal distribution and a Kruskal-Wallis test otherwise.

TABLE 2. Effectiveness through nearest-neighbor mediation.

Sim	Recall	RelAspR	R-Precision	AUP
Cosine	1.000000	1.000000	0.200330	0.161563
Dice	1.000000	1.000000	0.142527	0.119690

number of terms that are specific for relevant documents, but not for nonrelevant documents.

Nearest Neighbors Mediation

Mediation is a generic interaction model that does not specify how the user should organize her exploration of the source collection, selection of exemplary documents, and request for the generation of a topic model and of a mediated query. There can be different approaches to mediation in terms of the number of documents used for generating each mediated query and the number of distinct queries submitted in a search session (for retrieving documents relevant to a certain topic). The experiment reported in this subsection simulated one extreme case, when each of the documents judged relevant in the source collection was used to generate a distinct query which was submitted to the target collection for a nearest-neighbor search. Cosine and Dice were the two similarity measures used for computing interdocument similarities in order to find nearest neighbors.

Table 2 shows the results of the nearest neighbor mediation obtained by score-fusing the lists of nearest neighbors associated with each relevant document found in the source. For documents that appeared in multiple lists, only the higher score was considered. An immediate result is that recall and relative aspectual recall are 1 for all combinations of weighting schemes and similarity measures, indicating that all the relevant documents were retrieved. A statistical analysis of variance shows that, compared to the baseline search, the recall has improved highly significantly ($p < .01$), while the relative aspectual recall is also higher, but the difference is not statistically significant. The increase in precision (both R-precision and AUP) through nearest-neighbor mediation is highly significant ($p < .01$).

Apart from the extreme case simulated above, we also considered a more realistic scenario. The results of our document clustering experiments confirmed the picture depicted in Figure 3: The documents relevant for a certain topic are typically spread over the cluster structure in pockets of relevance, each of them corresponding to a distinct aspect of that topic.

Therefore, we considered the scenario when the user does not do a comprehensive search for relevant documents in the source selection, but is satisfied (possibly because of a time restriction) with finding just one exemplary document in each pocket of relevance. We simulated this scenario by taking just one relevant document for each aspect of each topic and using the obtained set for nearest neighbor

TABLE 3. Effectiveness through nearest-neighbor mediation when only one exemplary document from each aspect is used.

Sim	Recall	RelAspR	R-Precision	AUP
Cosine	1.000000	1.000000	0.184206	0.146965
Dice	1.000000	1.000000	0.114829	0.093455

mediation. The results are in Table 3. The analysis of variance shows that the improvement in precision, compared to the baseline, is still significant ($p < .01$ for R-precision, $p < .05$ for AUP), although slightly lower than when all relevant documents are used. Recall and relative aspectual recall are 1, so even using just one exemplary document for each aspect will support the retrieval of all relevant documents from the target collection.

Upperbound Experiment for Topic Modeling

We investigated the capacity of statistical language models to support topic models, built based on a set of documents judged relevant, and to generate queries for mediation. Before exploring mediation through a real cluster structure, obtained with concrete clustering methods, we started with an upperbound experiment: We considered an “ideal” clustering method that would group together in one cluster, for each topic, all the relevant documents in the source collection and no nonrelevant documents. While in practice this case is unlikely to happen, this experiment allowed us to analyze the effect on retrieval effectiveness and, therefore, on mediation quality of various independent variables.

The experiment consisted of taking, for each topic, the set of all relevant documents in the source collection, generating a topic model for it, and deriving the mediated query to be submitted to the target collection. The independent variables considered are the size of the query derived from the topic model and the weighting scheme used in the searching process. The results are displayed in Table 4. The user is reminded that, although the statistical analysis covers all the weighting schemes, the tables only show the recall and precision values corresponding to TfIDF.

The increase in precision of retrieval due to mediation, compared to the baseline search, expressed both as R-

TABLE 4. Effectiveness of mediated search based on topic models.

QuerySize	Recall	RelAspR	R-Precision	AUP
100	1.000000	1.000000	0.178181	0.143192
75	1.000000	1.000000	0.172434	0.142281
50	1.000000	1.000000	0.172434	0.140619
40	1.000000	1.000000	0.172434	0.141274
30	0.994253	0.983333	0.172434	0.141272
20	0.994253	0.983333	0.166687	0.137744
15	0.994253	0.983333	0.166687	0.136986
10	0.988506	0.983333	0.166687	0.138048
5	0.969987	0.955556	0.151680	0.131001

precision and of AUP, is highly significant ($p < .01$). The increase in absolute recall is also highly significant: a mediated query of as small a size as 5 gives better recall than a query based on the topic title, while longer mediated queries consistently generate better recall than a query based on the full description of the topic. Aspectual recall, already close to 1 in the baseline experiment when the full description of the topics is used as a query, only shows a significant improvement for longer mediated queries, which cover all the relevant aspects.

The Effect of Query Term Weighting

Another question we wanted to answer was whether it is sufficient to identify the most typical terms for a topic or weighting them is also necessary to capture the topic accurately. We studied the effect of the mediated query weights on target search effectiveness by comparing the results obtained in the previous subsection with results obtained by setting all query weights to 1 (after the real weights were used to rank the terms and to choose the most significant ones). Table 5 shows the new result.

As expected, the statistical analysis indicates a significant difference between the two cases ($p < .01$). It is clear that the weights are decisive in generating a substantially improved precision. (The recall is not affected if no output cutoff is applied.) Therefore, topic models should be weighted and search engines that allow query terms weighting should be preferred for searching the target collection.

If weighted search of the target collection is not possible, the increase in precision, compared to the baseline search, is not statistically significant. For recall-oriented information-seeking, mediation is still worth doing, as the increase in recall is highly significant.

Cluster-Based Mediation

While the previous two subsections considered the topic model generated based on the set of all relevant documents for a certain topic, here we are investigating the potential for mediation of a real cluster structure, obtained with real clustering algorithms. While in future experiments we may want to compare the influence of clustering algorithms and

TABLE 5. Effectiveness of mediated search based on topic models when query weights are ignored.

QuerySize	Recall	RelAspR	R-Precision	AUP
100	1.000000	1.000000	0.102077	0.101806
75	1.000000	1.000000	0.113651	0.097485
50	1.000000	1.000000	0.103859	0.092796
40	1.000000	1.000000	0.137969	0.104523
30	0.994253	0.983333	0.169179	0.108548
20	0.994253	0.983333	0.142518	0.104722
15	0.994253	0.983333	0.169762	0.118708
10	0.988506	0.983333	0.174433	0.120020
5	0.969987	0.955556	0.120226	0.102199

TABLE 6. "Best cluster" mediation.

QuerySize	Recall	RelAspR	R-Precision	AUP
100	1.000000	1.000000	0.112761	0.104839
75	0.982759	0.983333	0.112761	0.104845
50	0.982759	0.983333	0.109288	0.104558
40	0.942529	0.966667	0.100955	0.105631
30	0.942529	0.966667	0.112761	0.105139
20	0.898787	0.916667	0.112761	0.104307
15	0.852810	0.866667	0.109288	0.108837
10	0.826229	0.860494	0.100955	0.103146
5	0.810105	0.816049	0.097483	0.107395

their parameters for mediation, now we are content to break the ice in this kind of experiment and to consider just one structure. We chose the hierarchy produced by complete-link clustering, likely to identify topical clusters biased toward precision, with Cosine used as similarity measure and Kullback-Liebler as weighting scheme for computing interdocument similarities. For estimating the best clusters, we used the F measure of cluster quality (Jardine & van Rijsbergen, 1971):

$$F = \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R},$$

where β is a parameter that can be used to bias the measure towards recall (R) or precision (P).⁹ The justification for this measure of cluster quality for mediation is that a "good" cluster should have both reasonable recall (so that the user gets good coverage of her topic of interest) and reasonable precision (so that the user does not get too many nonrelevant documents).

The three scenarios that we simulated are:

"Best cluster": The user identifies the best cluster and uses it for mediation. The simulation finds the cluster with highest F, builds a statistical model, and generates a query for searching the target collection.

"Fuse and search": The user identifies a set of top-ranking clusters that, taken together, address her information need. The simulation identifies the best set of nonoverlapping clusters in the hierarchy, builds the overall topic model, and generates the mediated query.

"Search and fuse": The user investigates various aspects of the topic separately, by identifying the set of top-ranking clusters that cover all relevant aspects and using mediation separately, on each of them. The simulation builds a separate statistical model for each aspect, based on which it generates a query, and searches the target collection, then it score-fuses the results.

For all three scenarios, the mediated queries were submitted repeatedly, truncated at various lengths in order to

⁹R and P are based on the relevance judgments of the test collections.

TABLE 7. "Fuse and search" mediation.

QuerySize	Recall	RelAspR	R-Precision	AUP
100	1.000000	1.000000	0.172434	0.141794
75	1.000000	1.000000	0.167532	0.140122
50	1.000000	1.000000	0.167532	0.139609
40	1.000000	1.000000	0.171004	0.139803
30	0.994253	0.983333	0.184240	0.139452
20	0.994253	0.983333	0.175020	0.134235
15	0.994253	0.983333	0.153724	0.133148
10	0.988506	0.983333	0.160014	0.133289
5	0.969987	0.955556	0.151680	0.126894

asses the influence of the query size in retrieval effectiveness.

The variation with β of the results was negligible to at least the fifth decimal place, so only the results for $\beta = 0.5$, for precision-oriented clusters, are shown. Table 6 shows the result of mediating through the cluster with best F score for each topic.

When compared to the baseline search, the best cluster mediation tends to generate lower recall, unless relatively long queries are employed. Even worse, aspectual recall is significantly lower ($p < .05$). This result is hardly surprising. As clustering groups relevant document into pockets of relevance that tend to be associated with different aspects of a topic, one cannot expect even the best cluster to cover all relevant aspects. On the other hand, there is an increase in precision, highly significant for the AUP measure ($p < .01$).

A preliminary conclusion is that, for a user employing mediation based on a clustered document collection, a one-cluster strategy is fast and can be used as a precision device if the user is interested in exploring a certain aspect of an information need. If the user is interested in more than one aspect of a topic, more than one cluster should be used for mediation.

Better results were obtained when a "fuse and search" strategy was used, as shown in Table 7. The highly significant increase in recall ($p < .01$) was expected, as the partition used for mediation covers the whole source collection, and implicitly all the aspects of each topic. However, some aspects are better represented than others in the source collection. Therefore, topical terms specific to some aspects may be ranked higher than terms specific to other aspects in the topic model. Consequently, aspectual recall decreases through this type of mediation, unless a high number of query terms is considered. The decrease is not statistically significant and is nonexistent if the user sets the query size high, accepting a trade-off in speed.

In terms of precision, the gain through mediation is highly significant, being, for this strategy, close to that of the upperbound experiment, when the topic was based on all the relevant documents. It appears that the error introduced in the topic models by nonrelevant documents in the selected clusters does not affect precision significantly. It is probably because, although judged nonrelevant, the "residual" documents in each cluster are very similar, in statistical

terms, to the relevant documents so the topic models are not affected greatly.

We expected an even higher increase in precision with the "search and fuse" strategy. The reasoning was that the user would concentrate on distinct aspects of her topic of interest and get help in formulating very precise queries with respect to each aspect. However, the actual results, depicted in Table 8, show an abysmal decrease in precision. An explanation emerges if the log of the queries submitted to the target collection is examined: While occasional terms do suggest the original topic, most terms concentrate on a different topic. This highlights one of the problems with clustering: If a topic does not match a major axis of the hierarchic structure, the documents relevant to it are scattered all over the hierarchy, usually in small groups. For example, documents that refer to violence against tourists are grouped around incidents that involved tourists in Egypt, Florida, Kashmir, Turkey, Mexico, Morocco, Algeria, China, and so on. Apart from having a few common terms such as "violence," and "tourist," these documents and the clusters that encompass them do not display much similarity. Therefore, generating topic models and deriving queries based on each of these clusters fails to capture the common topic. On the other hand, both absolute and relative recall are 1 even with queries as short as 5 terms, showing that this strategy seems to ensure a complete coverage of all aspects of interest.

Note that we only generated short queries. This is mainly because this algorithm is slow and the processing time increases with the number of aspects considered and the size of the mediated query. However, different clusters are expected to cover different aspects of the topic explored, so shorter queries should be sufficient. (According to the result of the best cluster mediation experiment, working with shorter queries should not degrade precision substantially.)

Finding Good Clusters

Mediation through topical clusters is a more realistic scenario than the one underlying the upperbound experiment: rather than assuming that the user can identify all the relevant documents and use them as exemplary documents for mediation, we assumed that the user would be able to identify good clusters, which give a reasonable balance between recall and precision. Future user experiments are expected to reveal whether real users in an operational setting are able to identify good clusters and to use them for mediation successfully. For now, we were content to use

TABLE 8. "Search and fuse" mediation.

QuerySize	Recall	RelAspR	R-Precision	AUP
20	1.000000	1.000000	0.049961	0.020935
15	1.000000	1.000000	0.046489	0.021164
10	1.000000	1.000000	0.046489	0.021095
5	1.000000	1.000000	0.036112	0.013700

simulations. More specifically, we investigated whether a user who had a vague idea of her information need and was able to formulate a basic query could employ cluster-based retrieval to identify clusters with high F scores, shown to be good for mediation.

In our simulation, we computed F scores for each cluster in the structured source collection, with different values for β : 0.5 (biased towards precision), 1.0 (balanced), and 2.0 (balanced towards recall). Separately, we used the Cosine and Dice similarity measures to compute matching scores between each cluster label and queries derived from the topic descriptions available in various forms (“title,” “description,” “all”) for each of the test topics. We must stress the realism of this approach: in user experiments, the users consistently built their queries based on the terms of the topic descriptions.

The results were very satisfying: we obtained significant Pearson correlation between the two sets of scores, consistently over the topics and over the weighting schemes (correlation values varied between 0.166641 and 0.590561, with $p < .1$). Cosine was significantly better than Dice ($p < .01$). There was no significant difference between the various forms of the topic description. This suggests that the user does not need to make a mental effort to supply context terms for the topic; the most specific terms for the topic are sufficient to identify good clusters. Such a result can be viewed as a confirmation that clustering does indeed group together topical documents and that these good clusters can be identified by their most topical terms.

The consequence for mediation is immediate: By using cluster-based searching, based on a reasonable query, the user is able to find clusters that are good for mediation. Once these clusters are identified, the user can follow whatever mediation strategy she chooses.

Discussion

The results from our simulations suggest that nearest-neighbor mediation based on the “more like this” paradigm is potentially more successful than mediation based on explicit topic models. However, the topic model approach is faster (which is important in an operational setting), more flexible (the user can edit the mediated query proposed by the system, in order to change its focus), and easier to bookmark and reuse. Moreover, many search engines do not support nearest-neighbor searching.

The results also suggest what mediation strategies the user should use, according to her specific task:

- “Best cluster” mediation: if the user is interested in quickly investigating a certain aspect of the information need.
- “Fuse and search” mediation: if the user is interested in the overall topic and wants coverage of most relevance aspects, as well as high precision.
- “Search and fuse” mediation: if the user is interested in exploring in more detail the particularities of various aspects of the information need.

A somewhat disappointing aspect of our results is the generally low values of precision. This, corroborated with the rather low similarity values even between documents relevant to the same topic, obtained in clustering experiments with the same source collection, indicates a rather poor quality of the test topics. Consequently, we are limited in the quality of topic models that we can generate. Unfortunately, no better test collection was available to simultaneously support experiments on the aspectual cluster hypothesis, simulations of mediation, and user experiments. One consequence is that results and conclusions reported here should be verified on other collections before being generalized. On the other hand, if mediation experiments are successful on such poor topics, then we are encouraged to believe that they would be much better on a better collection and, moreover, that mediation is likely to be very successful when used with a high-quality specialized collection.

Conclusions

Contributions

Let us summarize some of the contributions of the work reported in this article. One main contribution of our work is proposing the concept of system-based mediated access as a way of emulating the human mediator, by offering the user:

1. Support for clarifying and refining her information need.
2. Support for generating high-quality queries.

An information retrieval system based on mediation is expected to be particularly useful for novice searchers or for users exploring a new domain, with which they are unfamiliar. An automatic mediator could significantly improve searching effectiveness and, implicitly, the user’s satisfaction. Our proposed interaction model contributes to the modern interactive trend in which the user interacts with the system in order to explore a problem domain and to obtain information relevant to a certain task or information need.

The mediated access concept, proposed as a generic interaction model expected to increase retrieval effectiveness and user satisfaction, raises a multitude of theoretical and practical issues, namely:

1. The conceptual and the mathematical model employed for representing topics.
2. The document and cluster representative formulae and the search strategies that are best for identifying relevant documents and clusters in the source collection.
3. The number of exemplary documents required to convey a topic unambiguously and the acceptable error margin.
4. The mediation strategies that offer best retrieval effectiveness.

While not attempting to solve all these issues, we discussed them and proposed an evaluation methodology for

investigating them. An experimental framework has been set up that consists of a set of hypotheses and conjectures, a set of experimental designs, and software to implement these experiments.

This experimental framework, together with the software framework that offers indexing, clustering, and searching, provides the means to extend our experiments and to easily repeat them on other test collections. Future experiments will hopefully confirm our expectations with regards to the potential of mediation to improve the effectiveness of retrieval.

The conceptual model of mediation does not impose restrictions on how the source collection should be structured. However, we investigated the use of clustering as a structuring tool and our simulations of various mediation strategies have proved the feasibility of this approach.

We also proposed the aspectual cluster hypothesis:

Highly similar documents tend to be relevant to the same topic. Documents relevant to the same topic may be quite dissimilar if they cover distinct aspects of the topic.

Finally, while most researchers in document clustering have investigated different formulae for generating a unique cluster label that should balance representation with power of discrimination, we have proposed the use of multiple cluster representatives, based on the context of their use. The Kullback-Liebler divergence allows for flexibility in generating a context-dependent representation.

Experimental Results and Limitations

Our simulations have shown that, at least for the experimental collection and the topics considered, mediated access through a structured source collection has potential to improve the user's query and to increase retrieval effectiveness. Another potential advantage of mediation, expected to be detected in user experiments, is the learning process and clarification and refinement of the information need, which happen when a user explores the source collection.

Due to the small number of topics and to the relatively small size of the source collection, these results cannot be safely generalized. 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.

Future Work

Our future work is envisaged to build on the experience gained from the experiments described here and to follow several directions:

- On the theoretical front, we need to find appropriate smoothing techniques for the statistical models that we employ.
- On the simulation front, we need to extend our experiments

by considering larger collections and a larger number of queries (even if aspectual relevance feedback is not available) with a wider range of topic specificity levels.

- User experiments are necessary in order to confirm that the potential of mediation can be realized by real users, in realistic scenarios.
- In order to extend the applicability of our model, methods for capturing implicit user feedback (derived from interpreting user behavior) should also be investigated.

Future research in mediated retrieval is worth pursuing due to the the fundamental issue that it addresses: the inability of the average searcher to generate clear and precise expressions of his or her information need. Once precise models of the user's topics of interest can be obtained through mediation, and, consequently, user profiles constructed, search algorithms or techniques such as autonomous agents will become much more effective.

Acknowledgments

The WebCluster Project was sponsored by Ubilab, Union Bank of Switzerland, Zurich.

References

- Bates, M.J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5), 407–424.
- Belkin, N.J., Oddy, R.N., & Brooks, H.M. (1982, ASK for June). Information retrieval: Part I, background and theory. *Journal of Documentation*, 38(2), 61–71.
- Blair, D.C., & Kimbrough, S.O. (2002, May). Exemplary documents: A foundation for information retrieval design. *Information Processing and Management*, 38(3), 363–379.
- Callan, J.P., Croft, W.B., & Harding, S.M. (1992). The INQUERY retrieval system. DEXA'92 (pp. 78–83). Valencia, Spain: Springer-Verlag.
- El-Hamdouchi, A., & Willett, P. (1987). Techniques for the measurement of clustering tendency in document retrieval systems. *Journal of Information Science*, 13, 361–365.
- Ellis, D., Furner-Hines, J., & Willett, P. (1993). Measuring the degree of similarity between objects in text retrieval systems. *Perspectives in Information Management*, 3(2), 128–149.
- Harman, D. (1992). Relevance feedback revisited. *Proceedings of SIGIR'92* (pp. 1–15). Copenhagen, Denmark: ACM.
- Harper, D.J., Mechkour, M., & Muresan, G. (1999, April). Document clustering for mediated information access. *Proceedings of the 21st Annual BCS-IRSG Colloquium*, Glasgow, UK.
- Hearst, M.A., & Pedersen, J.O. (1996, August). Reexamining the cluster hypothesis: Scatter/Gather on retrieval results. In H.-P. Frei, D. Harman, P. Schauble, & R. Wilkinson (Eds.), *Proceedings of SIGIR'96* (pp. 76–84). Zurich, Switzerland: ACM.
- Hersh, W., Leone, T.J., & Hickam, D. (1994, July). OHSUMED: An interactive retrieval evaluation and new large test collection for research. *Proceedings of SIGIR'94* (pp. 192–201). Dublin, UK: ACM.
- Jansen, B.J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the Web. *Information Processing and Management*, 36, 207–227.
- Jardine, N., & van Rijsbergen, C.J. (1971). The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, 7, 217–240.
- Jones, K.S. (1973). Collection properties influencing automatic term classification performance. *Information Storage and Retrieval*, 9, 499–513.

- Kelly, D., & Belkin, N.J. (2001, September). Reading time, scrolling and interaction: Exploring implicit sources of user preference for relevance feedback. In W.B. Croft, D.J. Harper, D.H. Kraft, & J. Zobel (Eds.), *Proceedings of SIGIR'01* (pp. 408–409). New Orleans: ACM.
- Leuski, A. (2001, May). *Interactive information organization: Techniques and evaluation*. Unpublished doctoral dissertation, Department of Computer Science, University of Massachusetts, Amherst.
- Manning, C.D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.
- Muresan, G. (2002, January). *Using document clustering and language modelling in mediated information retrieval*. Unpublished doctoral dissertation, School of Computing, Robert Gordon University, Aberdeen, UK.
- Muresan, G., & Harper, D.J. (2001, September). Document clustering and language models for system-mediated information access. In P. Constantopoulos & I.T. Solvberg (Eds.), *Proceedings of ECDL'01* (pp. 438–449). Darmstadt, Germany: Springer.
- Muresan, G., Harper, D.J., & Goker, A. (2001, September). ClusterBook, a tool for system-mediated access via clustered collections. Demo/poster presented at ECDL'01, Darmstadt, Germany.
- Muresan, G., Harper, D.J., & Mechkour, M. (1999, August). WebCluster, a tool for mediated information access. In M. Hearst, F. Gey, & R. Tong (Eds.), *Proceedings of SIGIR'99* (p. 337). Berkeley: ACM.
- Muresan, G., Harper, D.J.H., Goker, A., & Lowit, P. (2000, July). ClusterBook, a tool for dual information access. In N.J. Belkin, P. Ingwersen, & M.-K. Leong (Eds.), *Proceedings of SIGIR 2000* (p. 391). Athens, GA: ACM.
- Nordlie, R. (1996, October). Unmediated and mediated information searching in the public library. In *Proceedings of ASIS'96. Annual Conference*, Baltimore, MD. Retrieved from <http://www.ais.org/annual-96/ElectronicProceedings/nordlie.html>
- Nordlie, R. (1999, August). "User revelation": A comparison of initial queries and ensuing question development in online searching and in human reference interactions. *Proceedings of SIGIR'99* (pp. 11–18). Berkeley, CA: ACM.
- Pollitt, S. (1997, June). Interactive information retrieval based on faceted classification using views. *Knowledge organization for information retrieval, Proceedings of the Sixth International Study Conference on Classification*, University College, London.
- Pollitt, S. (1998). The key role of classification and indexing in view-based searching. *International Cataloguing and Bibliographic Control*, 27(2), 37–40.
- Pratt, W. (1999, March). *Dynamic categorization: A method for decreasing information overload*. Unpublished doctoral dissertation, Stanford Medical Informatics, Stanford University, Stanford, CA.
- Pratt, W., Hearst, M., & Fagan, L. (1999, July). A knowledge-based approach to organizing retrieved documents. *Proceedings of AAAI-99*.
- Shaw, W.M., Jr., Burgin, R., & Howell, P. (1997). Performance standards and evaluations in IR test collections: Cluster-based retrieval models. *Information Processing and Management*, 33(1), 1–14.
- Spink, A., Goodrum, A., & Robins, D. (1998). Elicitation behaviour during mediated information retrieval. *Information Processing and Management*, 34(2/3), 257–273.
- Taylor, R.S. (1968). Question-negotiation and information seeking in libraries. *College and Research Libraries*, 29, 178–194.
- van Rijsbergen, C.J., & Sparck Jones, K. (1993). A test for the separation of relevant and non-relevant documents in experimental retrieval collections. *Journal of Documentation*, 29(3), 251–257.
- Willett, P. (1988). Recent trends in hierarchic document clustering: A critical review. *Information Processing and Management*, 24(5), 577–597.
- Zamir, O., & Etzioni, O. (1999). Grouper: A dynamic clustering interface to Web search results. *Proceedings of WWW8*.
- Zamir, O., Etzioni, O., Madani, O., & Karp, R.M. (1997). Fast and intuitive clustering of Web documents. *The Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*. AAAI.