WebExplorer: A TOOL FOR ONTOLOGY-BASED INFORMATION EXPLORATION

Feza Baskaya, Anne Keskimaa, Jaana Kekäläinen, Kalervo Järvelin
Department of Information Studies and Interactive Media
University of Tampere, Finland

ABSTRACT
Vocabulary mismatch and muddled search goals plague users of traditional information retrieval (IR) systems. The former problem affects both monolingual IR and cross-language IR in particular, even if the user is able to express her information need clearly. This is due to the number of ways the search concepts may be expressed in texts and also due to the user’s incomplete skills in using foreign languages. The latter problem is inherent in information seeking and retrieval: information needs evolve as users learn about their information problems. Traditional approaches to vocabulary mismatch have been query expansion, and document indexing through thesauri, or recently, ontologies. The latter two assume document indexing (or annotation) and suffer from costs, inconsistencies and the great variety of ontologies in use. For muddled search goals there has not been proper support: users need to examine retrieved documents and reformulate their need statements. In the present paper we describe and evaluate a novel tool, WebExplorer, for ontology-based multilingual document exploration. It provides an aid for multiple-source searching and document analysis through summarization, clustering, and classification, among others.

KEYWORDS
Ontologies, cross-language information retrieval, summarization, classification, user interface, formative evaluation

1. INTRODUCTION
In modern Web information environments, billions of documents are available for searching. The documents are written and retrieved in many different languages. Because document production and publication are distributed, there are no generally agreed rules regarding layout or content formation. Consequently, the occasional user faces an old problem, vocabulary mismatch, meaning the difficulty of guessing the exact textual content of the documents (s)he would like to access. The problem is not made easier by huge amounts of spam and other unwanted material the user must sift through. Text searching augmented by link-based ranking is the search engine approach to relieve the access problems. The problem of expressing her/his information needs in one or more (foreign) languages nevertheless remains.

Web searchers’ needs may be classified into informational, navigational or transactional (Broder 2002). In the last two cases a searcher may be quite aware of the target (s)he wants to reach. We focus on informational needs. The searcher’s situation can be characterized as follows: (1) (s)he is interested in a tiny fraction of the documents but that fraction is most often unknown initially; (2) the information need may not be clear and stable; (3) searcher does not know which ontologies or folksonomies may have been used in annotation and is unwilling find out these; (4) even when knowing the applied ontologies, the searcher is unwilling to consult them, may not understand their semantics correctly, and may not find suitable tags.

Cleverdon (1984) summarized many studies on indexing consistency by stating that the indexing results of the same documents, or query formulations on the same topic, produced by two competent professionals using the same index language (ontology) are only some 30 % consistent. On the other hand, searching by full-text words with simple coordination and best match provides the full specificity and exhaustivity of the original text and flexibility of searching (Sparck Jones 2004).

In the present paper we set forth from the conviction that there is no alternative for text searching in modern large document collections: most documents remain unannotated, and the remaining ones are annotated inconsistently perhaps using an ontology which the user is not likely to examine. Automatic annotation does
not solve the problem either because the semantics of the automatic tags cannot be guaranteed. Nevertheless, we do not want to entirely dispose ontologies. In the present paper we suggest using ontologies as search tools – and, through them, search collections which may have some tagging but need not have any. We seek to remedy the challenges as follows: (1) the searcher uses her/his own ontology that represents her/his interests – other onto-folksonomies need not be known. (2) Inconsistency is managed by having relevant expressions as synonyms (and translations) in the search ontology. (3) The searcher would learn / understand the personal ontology. (4) All text with possible tags is searchable. In addition, we aim to support users, who cannot properly explicate their needs. This is achieved in part by having the search ontology as a learning and reminding tool. Further, we integrate several techniques supporting information exploration with the search tool: online document summarization, clustering, ontology-based classification and ontology-based bookmark management. These support evolutionary query reformulation over a session, e.g. Batesian berry-picking (Bates 1989).

The rest of the paper is organized as follows. Section 2 opens by requirements analysis for tools supporting ontology-based information exploration in unannotated text collections and then presents our prototype tool, WebExplorer. Section 3 presents the results of a formative user evaluation and Section 4 concludes the paper.

2. THE DESIGN OF WEBEXPLORER

2.1 Requirements on the Tool

Due to the characteristics of modern search environments, document annotations are welcome but best treated as additional keys for search. Ontologies may however be useful for searching if they meet the following requirements:

- Personal: ontologies are personal or close-enough so that the searcher knows them
- Small-scale: ontologies are small-enough so that the searcher can manage their content
- Mapping: ontologies map searchers’ concepts to reasonable synonymous and associated keywords
- Multilingual: ontologies map searchers’ concepts to keywords in desired searching languages
- Editable: changes in searchers’ interests (concepts) may easily be encoded in the ontologies without revising document annotations.

Search ontologies (Bates 1990; Kristensen and Järvelin 1990) can be designed to meet these requirements and are applied in WebExplorer. The mapping requirement is illustrated in Figure 1, which shows three abstraction levels of a search ontology: conceptual, linguistic and string levels (Järvelin et al. 2001). The conceptual level is intended for user viewing and corresponds to the traditional ontologies. Concepts are mapped to alternative synonymous and associated expressions in one or more natural languages or terms used for annotation. Expressions, again, are mapped to search strings, which may be stems, lemmas, string patterns with wild cards, or proximity window expressions for phrases depending on the searching environment and desired search effects. When searching through a search ontology, the selected search concepts are first expanded, and then their linguistic expressions are picked and finally transformed into the corresponding strings for the desired natural target and query language.

Additional functionalities are required for information exploration:
- Keyword search: the searcher may combine ex-ontology words to ontology concepts in searching
• Query-biased summaries: the searcher may quickly analyze retrieved document contents through query-biased summaries; these summaries can be produced also across documents and stored
• Classification: retrieved documents can be classified against chosen ontologies (not necessarily the one used for searching), also across languages, to identify concepts of interest in the document
• Clustering: rapid clustering of retrieved documents and presentation of cluster label terms for identification of significant novel concepts (which may be ex-ontology)
• Bookmarking: the searcher may store links in an ontology for later consultation and analysis
• Smooth integration of functions: text words, cluster labels, identified concepts can all be used flexibly for the next iteration of search and analysis for an evolving information need.

2.2 WebExplorer Interface

The user interface of WebExplorer is designed to work effectively and dynamically with ontology support. The interface is organized in two main panels: the left panel is for WebExplorer functions and the right one for presenting results (Figure 2). The left panel consists of the following tabs for corresponding functions: Search, Summary, Cluster, Classify, Bookmark, and RSS (Line 3 in Figure 2). WebExplorer starts with Search tab as the default. Before start-up, logging in is needed to be able to save personal bookmarks, RSS feeds and summaries. The user can switch between the tabs flexibly all the time.

Figure 2. The WebExplorer user interface in English to Italian cross-language retrieval

Search concepts can be selected from an ontology tree (Line 7). Keys for ex-ontology concepts can be given separately (Line 1). Various search ontologies can be downloaded on the fly (Line 6). Search parameters can be adjusted in an option window, where features like search engine, database, target language, query expansion level of the concepts in the ontology and search result count can be selected from pull-down menus (Line 5). The interface offers a progressive search in the displayed ontology, and highlights the matching concepts as the user is typing (Line 4). Selected concepts are highlighted with a different color. In addition to highlighting, the selected concepts are taken to the Selected Concepts field (Line 2). Moving the mouse pointer over a concept in the ontology tree shows its expressions in an overlay window. After selecting the
concepts, the user can initiate an expanded search via the Search button (Line 1 and 2). The query for the selected search engine is built and executed on the server side. The concepts and their expressions will be used for query building. In case of a different target language, they will be translated by the system. Search results are presented on the right panel.

On top of the result page one finds Summarization Query and tick boxes for activating the single link summarization and Show on Map, Show on Date Chart buttons (Line 8 and 9). If the tick box for single link summarization activation is checked, placing the mouse pointer over a search result for more than a second will bring the summarization inlay window for the current link (Line 11). Clicking Show on Map button (Line 9) shows the results on the Google map (see http://maps.google.com) application as an overlay. Titles and snippets are searched for place names, which are then shown on a map with a marker and link information box, so that the user can access the links shown on the map. Clicking Show on Date Chart button (Line 9) shows the links in chronological order, which is useful in contexts where the document dates are reliable.

Switching to the Clustering tab (Line 3) starts immediately to cluster the search results, and shows the cluster hierarchy. Clicking on a cluster label highlights the search results which are related to the label. The cluster label is taken to the next search query cycle as an additional keyword (on Line 1).

Classification of a document can be done either on the same ontology used for searching or on a different ontology, which can be downloaded upon request. Bookmarking allows storing bookmarks in ontology. The user can store a title, link, own description and tags for a link. Having stored the bookmarks for a concept, moving the mouse over the pertinent concept in the ontology tree shows the titles of the bookmarked documents. Under RSS feeds the user can save the RSS feeds, which (s)he follows. This enables the users to have the same functionality as for the search results, namely summarizing, clustering and classification. Figure 3 depicts the overall system architecture.

3. A FORMATIVE EVALUATION

3.1 Purpose, Methods and Data

The purpose of the formative evaluation was to test the feasibility of WebExplorer. The research questions were the following: What are the typical ways in which WebExplorer is used? How useful do the test users consider WebExplorer and its functions? In this paper we report the findings for the latter question. Eight people working in the library and Information Service of Finnish Parliament were recruited to test WebExplorer. As the system encompasses several functions, most of which are foreign to common web retrieval systems, users might not master it in the beginning, and a new interface may cause disorientation. To overcome such problems we ran our study in two phases: (1) an trial period of five days for four subjects (Group 1); (2) an observed user test (1 hour) for four subjects of the first phase and four new subjects (Group 2). The first phase took place in May 2009 and the second one, a month later.
In the second evaluation phase all eight subjects were given simulated work (information service) tasks (see Borlund 2000). The first task was about the international sanctions against North Korea. This task was inherently cross-lingual because the language of the search ontology (International relations) was Finnish and the query language was English. The task included also summarization of search results and bookmarking of the relevant links. The second task dealt with the effects of the climate change on the North Pole. The task integrated monolingual search (in Finnish) with the help of an environmental ontology, summarization of the search results, bookmarking of the relevant links, classification of some relevant links and modifying the query with concepts from the classification. The new relevant links were to be bookmarked. Group 1 did both tasks and Group 2 (novices) just the first one with the exception of the first subject of Group 2 who did both tasks.

The following data were collected in the second testing phase: First, the subjects filled a consent form giving their permission for voice recording, next they were given an initial questionnaire and after each task a post-task questionnaire, and a final questionnaire at the end. All transactions with WebExplorer were logged and the search sessions were voice-recorded. Additional questions for clarification were asked if needed. The search sessions were observed by one of the researchers.

On the final questionnaire the subjects were asked to evaluate the usefulness of WebExplorer’s functions (ontology-based IR, cross-language IR, summarization, classification, bookmarking and RSS-feeds) on scale: “no benefit at all”, “only little benefit”, “I can’t say”, “some benefit” and “much benefit”. The options were afterwards transformed into corresponding numbers (from -2 to +2) and the median was counted for each function amongst the answers of each group. As the only task of Group 2 did not include the classification function or the RSS feeds, their evaluations of the mentioned functions are not taken into account here.

### 3.2 Results

Group 1 considered the ontology-based IR and the cross-language IR as the most useful functions of WebExplorer. The median for both functions was +2 (“much benefit”). Similarly, Group 2 considered the cross-language IR to be the most useful function (median 1.5: between “some benefit” and “much benefit”). The second best function, what comes to usefulness, was the ontology-based IR (median 1: “some benefit”). Thus Group 2 (novices) was more critical than Group 1 but overall the opinions of both groups on the most useful function were quite the same.

What comes to the least useful function measured by the median, it was the summarization for Group 1. The probable reason for the judgment could be the quality of summaries which was criticized by the subjects. The known challenge is that many web documents are difficult to summarize due to their varying structure and other features. Nevertheless, the distribution of answers covered all answer options from “no benefit at all” to “much benefit”. The median was 0 (“I can’t say”). Group 2 evaluated the function more positively (median 1: “some benefit”). In contrast, Group 2 considered the bookmarking function to be the least useful one, the median being 0.5 (between “I can’t say” and “some benefit”). Several subjects were of the opinion that learning the bookmarking function required time. The restricted time for the system demonstration and the task could have affected their evaluation, especially if they felt that they could not make a proper use of the function. What comes to the classification function, the median was 0.5 (between “I can’t say” and “some benefit”) for Group 1. The comments made by the subjects showed that the function was considered quite useful on condition that one knows how to use it fluently.

The subjects were asked to give WebExplorer a school grade from scale 4 to 10. The grades ranged from 6 to 8.5, and the mean grade was 7. When asked if the subjects would be willing to use a tool like WebExplorer, 2/8 subjects chose the option “Yes”, 1/8 chose the option “No” and the rest 5/8 chose the option “I can’t say”. The relatively high number of subjects who were uncertain if they liked to use such a tool might be at least partly explained by the fact that they had so little time for testing the system.

### 4. DISCUSSION AND CONCLUSION

We have presented and evaluated a novel interface, WebExplorer, for ontology-based multilingual document retrieval and analysis. It provides versatile search functions with multi-source multi-lingual searching through ontologies and through searchers’ own keywords. Searching is possible in the Web and dedicated
collections, and RSS feeds can be received. A range of ontologies can be used without assuming annotation of documents. WebExplorer offers a document analysis function through online query biased summarization of individual documents and document sets. The summarization viewpoint can be flexibly changed. Further analysis is possible through clustering of documents and identification of cluster focus concepts which can be used in query modification for the next round. Also ontology-based cross-lingual classification, and bookmark management are available. Classification and clustering can be used to automatically feed into query reformulations, and summaries manually. These functions effectively support document analysis and user learning in case of unclear, muddled search goals. Finally, maps and timelines can be used to study the geographic and temporal distribution of the found items.

In the formative evaluation the subjects commented both (a) the system’s usability and (b) ontologies or searching with them in general. The most preferred functions were ontology-based and cross-language IR achieved through the three level architecture of the ontology. This result is reasonable because search key selection is a focal effort in searching; even more so if the target language is foreign. Among the functions evaluated in the second testing phase, summarization and bookmarking were considered to be the least useful ones. The result could be explained by the mixed quality of summaries of www pages. Further, these functions were not the most essential for the given test tasks. All the subjects were able to find search concepts in the test ontologies although subjects were not familiar with them. However, the subjects commented that knowing the ontology would facilitate searching. The formative evaluation indicated that digesting a new information exploration method takes some time; the subjects were not capable of exploiting all functions because of the short testing period. The results of the evaluation give rise to further development of the system, and further evaluation in a more specified work and search environment.

ACKNOWLEDGEMENT

This research was supported, in part, by the Academy of Finland grants #120996, #124131, #130482 and by the Finnish Funding Agency for Technology and Innovation grant #21239.

REFERENCES