Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2005 Proceedings

Americas Conference on Information Systems (AMCIS)

2005

Collaborative Indexing as a Framework for Search and Knowledge Management

John Owens
Virginia Commonwealth University, john@jowens.net

Collaborative indexing as a framework for search and knowledge management

John Owens
Department of Information Systems
Virginia Commonwealth University
john@jowens.net

ABSTRACT

Information search relevance is a key challenge for information systems researchers. We propose a framework for development of a new architecture for search, collaborative indexing, employing the collaborative thinking of informed individuals to develop indices. Creator circles of individuals will mark pages of relevance to a topic, or pages that usefully answer questions, through interaction with their web browser. These marked pages will create and continuously add to a domain specific search index. Through searches of this index by members of the creator circles, other members of the organization, or the public, we propose the results returned will have greater relevance and usefulness in finding information, collaborative learning, and collaborative problem solving.

Keywords

Search, collaboration, knowledge management, human-computer interaction

INTRODUCTION

Finding useful information to answer a problem is a challenge we all encounter. Users often create bookmarks to remember sites or pages they have visited on the Internet with useful information. Search technologies have been developed to mine the Web for content to answer queries. Google Search, MSN Search, and Yahoo Search are all examples of systems that utilize software bots to index content and utilize varying page ranking algorithms to determine relevant documents to return as answers to queries. Google, for instance, indexes over 8 trillion web pages—page ranking technology thus becomes critical in returning the best choices for users. Many vendors produce search software for organizations to index and search the organizations' local web and document content.

Kleinberg [1] speaks of the Abundance Problem, "the number of pages that could reasonably be returned as relevant *[in search results]* is far too large for a human user to digest." Kleinberg also speaks of the notion of authority relating to broad-based queries where some pages can be identified as authoritative on a subject.

Focused crawling has been suggested as a method for topic-specific resource discovery [2]. In focused crawling, a web crawling engine compares linked pages to a training dataset of exemplary documents to selectively index pages of relevance to pre-determined topics.

Research Goals

Mai [3] advocates a domain-centered approach to indexing content.

> The domain-centered approach to indexing takes the domain as the focal point of analysis and uses knowledge about the domain and the users to determine the subject matter of documents. The benefit of this approach is that indexers have a clear frame of reference for making decisions when indexing, it ensures that indexing is consistent with users' use of information, and it provides effective results. (Mai 2004)

We suggest a methodology for creation of domain-centered indexes for knowledge management and Internet content.

There are three propositions we advance while in the design theory stage:

1) Users often need answers to previous questions they have asked, or questions that others in their domain have asked.

2) A subset of all web content is relevant to answer a particular question, and a further subset of that content is actually useful to answer the question.

3) Present computer algorithms for determining relevancy do not surpass the capacity of a knowledgeable human determining relevancy.

Some relevant technologies

An interesting extension to the concept of bookmarking is social bookmarks, like those provided at http://del.icio.us. Individuals can share their bookmarks with others and searches can be done on the keywords of the links. This feature may be built in to some web browsers soon. Social bookmarks provide a method for users to determine relevance and usefulness of pages as determined by another individual in the community.

Eurekster (http://www.eurekster.com) has developed technology to note visited links from its search pages (generated from results by Yahoo), and then uses a measure of usefulness (a certain time spent on the linked site) to drive up relevance for the page in future searches. They also have a collaborative searching method called SearchParties, which allows fellow linked users' visits to sites to influence future search results for the whole group of linked users [4]. Table 1 shows a comparison of Eurekster attributes compared to this paper's methodology.

Google Desktop Search can search a local web cache on a user's computer for search terms, returning viewed pages. Google Desktop search keeps previous versions of pages as well, so past versions can be searched for changed content [5].

During our literature review, we also discovered architecture specifications for commKnowledge, an online community system [6]. The system allows users to submit URLs and user-created summaries of sites to a central database through a web POST form. These submissions can be located through a "Recent Additions" section, categories generated from the submitting user's categorization of the item, and a most popular section. Rankings of results can be influenced by user votes on an entry's usefulness through the website and computed measures of page connectivity. It was not clear if this system was ever operationalized.

	Eurekster.com	**The proposed method**
Source of Search Results	Yahoo search results	Search index created by direct end-user page URL addition to domain index through marking
Methods of influencing relevance	1) Measure of site relevance/ usefulness as judged by user spending a certain amount of time on a clicked result link 2) Moderator flags search result as useful/relevant Only occur when searching results from index	Relevant by the fact it was added to the domain index by the user Additions to search index can occur when viewing any page on Web, via a toolbar button press or other mechanism
Domain specificity	Yahoo search results modified by embedded keywords in search terms as specified at index creation	User specification of page to belong to domain-specific index
Group capabilities	Multiple users can influence relevance through use of search results In "Anarchy" mode, any group member can flag certain search results as moderator	Multiple users can submit page URLs for addition Potential to weigh relevance by number of users submitting page URL

Table 1 – Comparison of approaches to domain search by Eurekster and collaborative indexing

FRAMEWORK

Requirements

Mai [3] points to four steps for domain-centered indexing: analysis of the domain, determination of users' needs, determination of the indexers' perspectives and roles, and finally analysis of the document in terms of the domain and users' needs.

By proposition 3, *present computer algorithms for determining relevancy do not surpass the capacity of a human determining relevancy*, we advocate humans to perform these steps in our method. Yahoo and other search firms produced directories of web sites that were edited by human editors. However, these directories were more like collections of bookmarks rather than groupings of actual page content. We require a method for capturing human input to the system.

It would be difficult for one individual to analyze all needed content for a useful index. We thus require a method for multiplying effort in analysis.

Finally, a software search indexer is required to provide the final steps of the indexing process, analyzing text content and patterns in the pages.

Components

To meet the system requirement for capturing human input into the system, a marking tool should be developed. As the user browses local or Internet content, a button can be provided for adding the particular page URL or location to the index. The index grows as the user marks more content through use. We advocate a button press or other user initiation so as to provide privacy in searching (not all pages visited are automatically cataloged) and allow for possible domain categorization through selection of a particular index via the toolbar to add the entry. This differs from the approaches of Eurekster [4], which tracks selected links followed in the background, and commKnowledge [6], which requires a visit to the commKnowledge website to add a link and summary via web form.

Users can take different strategies as appropriate to adding content. For a particular index, only direct answers to problems might be added. In other cases any relevant and useful content as judged by the user could be added to the index.

Computer algorithms have the advantage of potentially unlimited effort in relevancy determination. To approach the algorithm's ability, a force multiplier is necessary. We advocate collaborative submission of items for indexing. Using a group of individuals numbering 10, 50, or larger all submitting items to a common indexer (with like-minded goals for the index, and with common "descriptive" roles [7] in building the index) a useful index can be created. This group approach offers the benefit of shared experience for the group users and outside users of the index to find answers or information others have found.

How it works

First, a decision is made to create a domain-specific index. The index's purpose and goals should be articulated. Second, a group of individuals should be identified and recruited to form the creator circle for the index. In an organization, this might be as simple as the task work unit, e.g. "Microsoft Windows Desktop Support Staff." As a domain-specific index shared across the Internet, the group of individuals might be those identified as experts in the domain, e.g. ten top immigration lawyers creating an index about immigration law.

These individuals should learn and agree to the purpose and goals and rules (if any) of the collaborative index. These individuals would install an application on their computer system to allow marking (submission of pages to the search indexer). One example of a marking tool could be an add-in toolbar in a web browser like Internet Explorer with buttons for "Add Page" and "Add Site." Refer to Figure 1 for an illustrative diagram.

Coincident with recruitment of the creator circle, a search indexer should be deployed on a networked server. For example, Microsoft's Sharepoint Portal Server software contains a web search indexer component. A programming wrapper needs to be created to receive submission requests from the creator circle and to add entries to the list of pages and sites indexed by the software. The wrapper could take the form of an ASP.NET web service accessible to the creator circle's browser toolbar. See Figure 1 for a diagram of a collaborative indexing configuration.

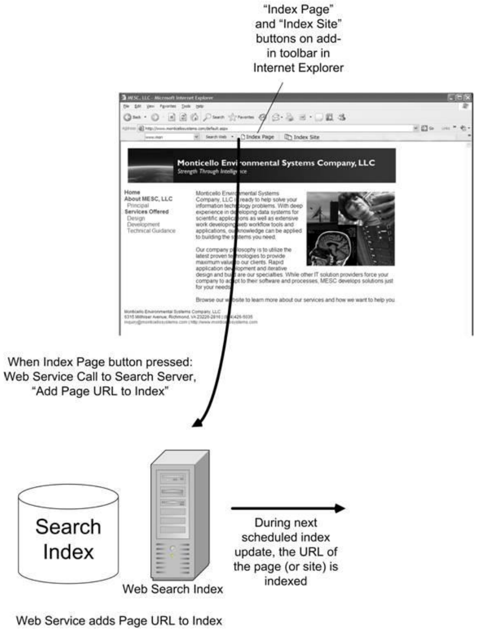
"Index Page" and "Index Site" buttons on add-in toolbar in Internet Explorer
MESC, LLC - Microsoft Internet Explorer
Index Page
Index Site
Monticello Environmental Systems Company, LLC
When Index Page button pressed: Web Service Call to Search Server, "Add Page URL to Index"
Search Index
Web Search Index
During next scheduled index update, the URL of the page (or site) is indexed
Web Service adds Page URL to Index

Figure 1—Diagram of collaborative index configuration and operation

Once the client software has been distributed and the server configured, users can begin to submit pages to the index. The index should be configured to update itself with changed content and to drop persistently non-working links from the index.

For access to search on the index, a search box can be provided for the creator circle through their toolbar. If individuals outside the circle are allowed to search the index, then a web page can provide this functionality.

Limitations

There are several limitations to collaborative indexing in this form. Probably the most obvious is the collaborative index will never provide information that has not been known and marked by at least one member of the creator circle for the index. The results returned from the collaborative index server could address this by providing results both from the collaborative index and another search source like Google search.

The index will not be useful immediately upon creation. The creator circle must add a critical mass of pages to the index before questions will be answered. This time is shortened by having the group of the creator circle adding pages together, but those expecting immediate results from the index may be disappointed.

EXTENSIONS

From the thinking process regarding collaborative indexing, several potential extensions were recognized. One such extension is development of a cumulative relevance for an indexed item. This could be computed from the number of creator circle members submitting a particular item for indexing. Another extension could be cross-domain searches when the domains are categorized together manually or in a automated fashion. Mentioned previously in the limitations section, searches could potentially be made more useful by combining results from the collaborative index with traditional page-ranked results from another source, like Google search.

FURTHER RESEARCH

One obvious goal for this research would be construction of an instantiation of a collaborative index. This will provide an artifact to test and evaluate in comparison with other methods for knowledge management and search.

Some interesting questions for the research agenda are: 1) Do more pages indexed mean more search relevance? 2) Does a very small subset of all Internet web pages hold all required answers for a specific domain? 3) Are most queries recurrences of previous queries? 4) Can algorithm-based determination of relevance match human determination of relevance?

REFERENCES

1. Chakrabarti, S., van der Berg, M., and Dom, B. (1999) Focused crawling: a new approach to topic-specific web resource discovery, in Proc. of the 8th International World-Wide Web Conference (WWW8).
2. Kleinberg, J. M. (1999) Authoritative sources in a hyperlinked environment, *Journal of the ACM*, 46, 5, 604-632.
3. Mai, J. (2005) Analysis in indexing: document and domain centered approaches, *Information Processing & Management*, 41, 3, 599-611.
4. Eurekster (2004). Eurekster Web Site. http://home.eurekster.com/howitworks.htm. Accessed 2/11/05.
5. Google (2004). Google Web Site. http://desktop.google.com/features.html#cachedweb. Accessed 2/11/05.
6. Gordon, M., Fan, W., Rafaeli, S., Wu, H. and Farag, N. (2003) The architecture of commKnowledge: combining link structure and user actions to support an online community, *Int. J. Electronic Business*, 1, 1, 69-82.
7. Wilson, P., (1968). Two kinds of power: an essay on bibliographic control. University of California Press, Berkeley. As cited by Mai (2004).