INTELLIGENT BROKER FOR COLLABORATIVE SEARCH AND RETRIEVAL OF CONSTRUCTION INFORMATION ON THE WWW

N. Bakis, M. Sun
School of Construction and Property Management
University of Salford
Salford, M7 9NU, UK

ABSTRACT: The construction industry is beginning to use the World Wide Web (WWW) as an information dissemination vehicle. The existing information retrieval tools, such as structured classification systems and Internet search engines are inadequate in supporting ordinary construction professionals to locate and retrieve the right information in the ever expand information cyber space. This paper presents a Construction Information Broker (CIB) which improves the discovery of construction related information on the WWW through user collaboration and incorporating construction domain knowledge in the information retrieval process.

KEYWORDS: Construction information, World Wide Web, Collaborative Information Retrieval, Information Filtering

1. INTRODUCTION

In the last few years, the World Wide Web has become the most important vehicle for information dissemination. Its ability to deliver multiple formats information to large number of users over geographic distances makes it an ideal choice for individual and corporate information providers. However, because of its tremendous size and anarchic structure, the discovery of information on the Web would be very difficult without tools, like the search engines or online structured information gateways. Construction like other industries begins to use the World Wide Web to disseminate and access information (Dixon 1998). Although the general purpose search engines, such as AltaVista or Infoseek, and general purpose online information gateways, like Yahoo, can be used to discover construction information, there are serious limitations and weaknesses of these information discovery solutions. There is a need for purpose-built information discovery systems specifically for construction. This paper describes the development of such a system, the Construction Information Broker (CIB).

2. THE INFORMATION DISCOVERY PROCESS

In this paper, the term "information discovery process" or "information seeking" is used as an overarching term to describe any process by which users seek to obtain information from automated information systems. The two main types of information seeking are information retrieval and information filtering. In information retrieval the user seeks in a relatively static information environment to obtain specific information which addresses a short-term interest. In information filtering the user seeks in a relatively dynamic information environment to obtain information which addresses a long-term interest, i.e. seeks to stay informed about this interest. A support process to information retrieval and information filtering is the storage of found information to a personal information space for the purpose of its future access. An

example of the personal information storage is the bookmark function of the WWW browsers.

2.1 Information Retrieval Systems

When it comes to the development of systems that assist the retrieval of information there are two possible options:

- Automatic acquisition and indexing of information
- Manual acquisition and organisation of information

The Web search engines are examples of systems that automatically acquire and index information. These search engines work by building up an index database for the documents available on the WWW. Starting from a web page, they follow all the hyperlinks found in the page. In this fashion they cover a considerable part of the World Wide Web, depending on the number of starting pages they know and the number of links found in the starting and successive pages. The automatic indexing and retrieval of information in response to the user's queries is based on strategies found in the research area of Information Retrieval.

Information Retrieval strategies assign a measure of similarity between a query, which is a set of terms expressing the user's information needs, and a document. These strategies are based on the common notion that the more often terms are found in both document and the query, the more relevant the document is deemed to be to the query. The main categories of methods for the calculation of the relevance of a document to a query include Probabilistic Retrieval, Boolean Indexing, Latent Semantic Indexing, Fuzzy Set Retrieval, and Retrieval based on the Vector Space Model, Inference Networks, Neural Networks, or Genetic Algorithms (Grossman 1998). Although the techniques of each category vary considerably, all of them suffer from the vocabulary mismatch problem. This problem arises from the fact that the same word may have more than one meaning and that for the same meaning may exist more than one words. This means that a document relevant to the users' information needs will not be selected by the system if the user uses a different word from the author of the document to refer to the same concept. It also means that a document including a word found in the user's query will be selected by the system even if it is not referred to the same concept. The performance of an Information Retrieval system is measured in terms of precision and recall. Precision is defined as the percentage of the selected documents that are relevant to the user's information needs while recall is defined as the percentage of the relevant documents in the collection that have been selected by the system (Grossman 1998). Precision and recall depend on the collection of documents indexed by the system. However, we can say that in general - and especially on the World Wide Web - precision and recall is often quite low.

With the manual acquisition and organisation of information the performance of an information retrieval system can be considerably improved. The available information can be grouped into categories for the purpose of its easiest retrieval. For each category rich meta-information can be used to further assist the retrieval of information. Depending on the meta-information we may even use a Database Management system to select those documents that suit our information needs. Our information needs can be expressed in a more specific and intelligent way than simply using keywords. Although some of the above actions could be automated, human intervention will always produce better results, at least in the foreseeable future.

The difference in the performance between an automated and a manual system becomes more noticeable when the system is built for a specific area of interest. This is because the more specific the area of interest is the better and more commonly accepted organisation of information we can achieve. Examples of systems for the retrieval of construction information on the World Wide Web that are based on the manual acquisition and organisation of information are the Construction Information Gateway (Lockley, 1998) and CONNET (Turk, 2000).

A system that is based on human intervention for the acquisition and organisation of information can outperform the automatic ones in the effectiveness of the information retrieval. However this type of systems has the disadvantage of the fact that they require extensive human effort to operate which makes the expansion in scope difficult. As a result, systems of this kind usually only cover a very small fraction of the WWW resources. Furthermore there are no widely accepted standards for classification. Usually each information supplier decides the structure and content of its materials governed by chance, occasional decisions and staff responsible for the implementation. The consequence is a lack of consistency and reliability and a lack of independence from the individuals performing the task. Another weakness of information gateway solutions is the difficulty of keeping information updated. Due to the rapid changing nature of the WWW, many resources are quickly becoming non-usable. Because there are no effective automatic updating mechanisms it is very difficult for the gateways to follow the rapid changes of their contents, addresses, appearing and disappearing of documents on the sites they cover.

2.2 Information Filtering Systems

The purpose of an Information Filtering (IF) system is to assist the user in finding information relevant to his or her long-term interests. Although some of the techniques used in Information Filtering systems are based on Information Retrieval techniques, there are a number of differences that make these systems quite distinctive. There are two main types of Information Filtering systems: (1) Content-based filtering systems, and (2) Collaborative information filtering systems.

2.2.1 Content-based Information Filtering Systems

Content-based Information Filtering systems have found application in a number of information environments such as the Usenet News. The Usenet News system provides the capability to a user to post an article in a common bulletin board that can be accessed by any other users. Despite the fact that each bulletin board is related to a specific area of interest, the number of articles posted each day can reach several hundreds making the task of finding and reading the interesting ones quite difficult. A Content-based Information Filtering System tries to solve this problem by automatically selecting those articles that suit a user's long-term interests. This selection is based on a user profile, which is a representation of the user's long-term interests, and on methods that match this profile with the incoming articles. A user profile consists of keywords and Information Retrieval methods can be used to match articles and profiles (Oard 1996).

Although a Content-based Information Filtering system may appear similar to an Information Retrieval one, there are a number of important differences. The first difference arises from the fact that each user can have several long-term interests and a user profile that consists of hundreds of keywords. Because quite often is difficult for the users to construct such big

profiles the system has to provide mechanisms for the automatic construction of them. These mechanisms are based on articles that the user has found interesting in the past or on the user interaction with the system (Stevens 1992). Secondly, because there can be changes in a user's interests over time, the system should be able to detect them and update the profile accordingly. Finally, another difference between Content-based Information Filtering systems and Information Retrieval systems is that in environments such as the Usenet News new articles appear quite often making the indexing process quite difficult thus a retrieval system impractical.

2.2.2 Collaborative Information Filtering Systems

A Collaborative Information Filtering system, instead of comparing the user profile with the information items, compares a user profile with other profiles to identify groups of users having similar interests. Having identified such groups, it predicts the importance of an information item for a particular user on the basis of the degree to which other users of the group have found this document interesting. The user profiles in Collaborative Filtering systems can consist of keywords describing the users information needs or pairs of item-rating that the user gave in the past. These ratings can be sometimes implicitly calculated by the system (Oard 1996).

Collaborative Information Filtering systems have several advantages over the content-based ones. Because they are based on human's evaluation of the information items they can be used for items that are not in a machine readable form, for example sound and photographs, and can select items based on some assessment of quality, style or semantics. In addition, they can select items the knowledge of which the user did not have before (serendipitous finds). However, the disadvantage of these systems is that the number of item ratings given by the users must reach a certain level before the system starts making selections.

3. THE CONSTRUCTION INFORMATION BROKER

The Construction Information Broker (CIB) is a system under development that aims at improving the discovery of construction related information on the World Wide Web. It follows the "automatic acquisition and indexing of information" approach for the retrieval of information. Its main characteristic is that it incorporates construction domain knowledge and supports Information Filtering capabilities as well as a collaborative type of information retrieval to achieve better performance than the general-purpose search engines.

Figure 1 illustrates the systems architecture, which consists of a CIB server and multiple collaborating CIB client agents. The CIB client agent is a plug-in package to standard WWW browsers, such as Microsoft Internet Explorer and Netscape, that facilitates the communication with the CIB server. Once it is installed on a user's computer, the agent will facilitate the communication with the CIB server and provide intelligent information search and retrieval services. To achieve this, the agent needs to gather information of user profile, information query context and information search history. Each user is associated with one or more profiles that describe his or her information interests. These profiles are stored in the client computer which also provides an enhanced storage space where the users can store any important information. Part of this information is communicated to the CIB server and stored in the knowledge base on the server. This accumulated knowledge helps to improve the information search task of the user community as a whole. The CIB server is in essence a construction domain specific WWW search engine. It does not host the original information

documents. Its main function is to match information sources to users search queries. The main components of the CIB server are a meta-search engine, an index agent, a query handler, and at the heart of the system a three-layer information index. The operation of the system and an explanation of each component are given in the following sections.

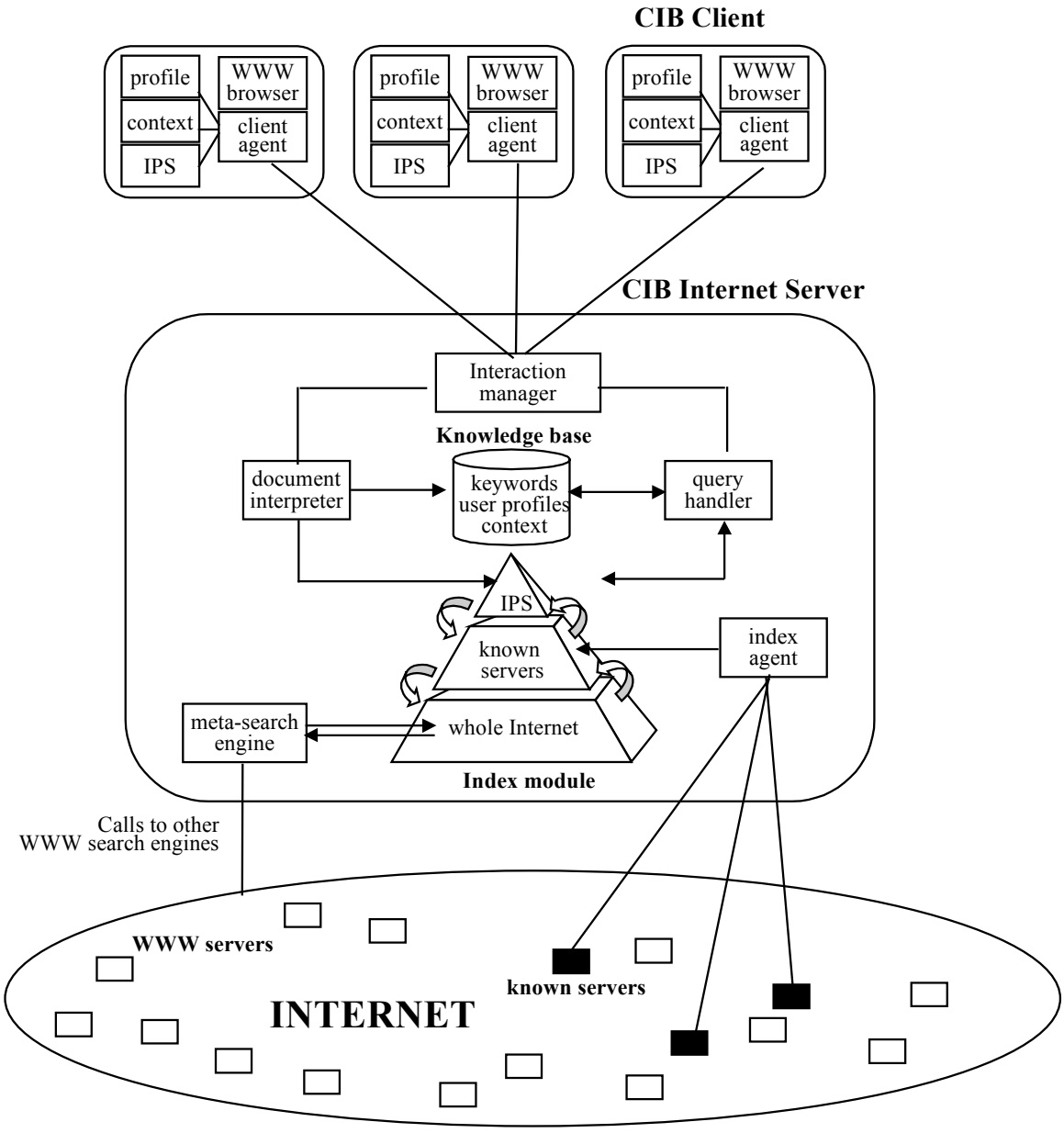

Figure 1. Architecture of the Construction Information Broker

3.1 Personal Information Storage

The CIB client provides a local repository facility for the end user supporting enhanced storage other than the existing bookmarks. It helps the user to organise the downloaded Internet documents for quick recall in the future. The information is stored in a meta-information format similar to Jasper's Intelligent Page Store (IPS) (Davies, 1997). For each WWW page, the system stores at least the following information:

- the document title
- a summary of the content
- a set of keywords
- users' contextual annotations
- universal resource locator (URL)and,
- date and time of storage or update

To avoid unnecessary network traffic, the interaction between the CIB clients and server is not real-time but rather session based. When a user starts the client agent at the beginning of a session, the agent first loads information of the previous interactions with the CIB server. The client then connects to the server and check for any new recommended documents by other clients with the same user profile. Once the connection is established, the user can use an advanced query mechanism to perform information searching. When a document is downloaded and rated by the user as of relevant to his or her interests, the client agent will extract the IPS format meta-information and store it for future use. At the end of each session, the locally stored IPS information is sent to the CIB server with the user's permission so that it can be shared with other users who have the same interests. The sharing mechanism is explained in section 3.3.

3.2 Information Retrieval

3.2.1 Query formulation

The first and most important step in the information searching process is the formulation of a query that best describes the users information needs. Quite often it is difficult for the user to choose the appropriate terms for various reasons. The user may not know what exactly is looking for, or may not use the appropriate term to describe it. He may use a broader term to describe a topic or miss several synonyms. Methods that help the user formulating his query have been studied extensively in the literature. One category of these methods is the expansion of the query to include more terms that may describe the user interests. This expansion can be done using relevance feedback or a thesaurus. The thesaurus can be constructed manually or automatically by the system on the basis of a statistical analysis of the document collection (Kowalski 1997). Kristensen (1993) showed that a manually constructed thesaurus in the domain of economics and environment used for query expansion resulted in a large improvement in overall recall at the expense of a small drop in precision. The query expansion terms can be automatically added by the system, or manually by the user. Magenis (1997) showed that in the manual query expansion, as it might performed by an experienced user, offers a small but significant further improvement in the retrieval performance. CIB uses a construction thesaurus developed using the existing classification standards to help the user in formulating queries. It does not automatically expand the query as it leaves the choice to the user. The query handler in the server is the component that includes the thesaurus and provides this functionality.

3.2.2 Document Selection

The second part in the retrieval process is the selection of documents that are relevant to the users query. CIB improves the retrieval process by using user profiles and a type of collaborative information retrieval.

The CIB server has a collection of built-in user profiles characterising the main type of users related to the construction industry, such as architect, building surveyor, project manager,

cost consultant, academic researcher, etc. Each profile is associated with a set of interests which in turn include a set of keywords and information context. These profiles can be used as independent key for formulating queries. When a user connects to the server, he or she can select one or more standard profiles according to particular interests. The user's profile is further personalised during the process of querying. When a user downloads a document and gives a high rating to the document the client agent on the local computer will extract key concepts from the document and use them to improve the user profile. At end of each session, the updated user profiles will be communicated to the server, of course with the explicit permission of the user. The user can also explicitly modify the user profile by creating new interests, adding keywords and giving weight to existing keywords. The server will update the user profile after collating and analysing inputs from a large number of clients. The update of profile means interests and keywords can be removed as well as added to reflect the user interaction with the server. Through such a learning process, the shared profile knowledge on the server is improving all the time. The server will progressively improve the understanding of the users' information requirements and the accuracy of the information search.

CIB supports a collaborative type of information retrieval. Each user in CIB is associated with one or more user profiles. During the retrieval process he can give a rating to the retrieved web pages. When the user retrieves a document of interest, the client agent installed on the user's computer extracts some information about the document, such as the title and a set of keywords. This information is stored locally in the IPS format described above. At the end of the interaction session the client agent sends this information to the CIB server together with the user profile information and its context. The context of the document is determined on the basis of the user interaction with the query handler. Alternatively, it can be specified by the user. When another user with similar profile raises a query, the documents rated by the first user will be searched before other information sources.

The CIB information index consists of three hierarchical layers (figure 1). The top layer stores information of documents that have been accessed by users and communicated to the server through the CIB client agents. The middle layer contains an index of documents located on known WWW servers which are more likely relevant to the construction users interests. These servers become known to CIB through two ways. (1) The information providers register their servers explicitly. (2) When the client agent submits a document to the top layer of the index module, the host server becomes known to the gateway implicitly. The CIB Index Agent is a robot that visits all the known servers periodically and gathers information about documents hosted on these servers. The bottom layer of the index module covers most of the accessible web hosts on the Internet. This layer is not actually stored in CIB. CIB uses a meta-search engine to achieve an Internet wide information search.

The hierarchical layering implies that the amount of information accessible increases as one moves down from the top to the middle and the bottom layers, but the average potential relevance to the users' interests decreases. The advantage of the layering approach is that it allows a user to specify the scope and manner of a query. One can raise a query just for the IPS layer where more criteria can be applied apart from keywords. Alternatively, the user can make a general query using enhanced keywords supported by the server's knowledge base to a wider index in the middle layer or even the whole Internet. A document's position in the index module is not fixed. There are migration paths through which a document can be moved from one layer to another as a result of users' search and retrieval actions. For example, when a document in the middle layer is retrieved by a user, the client agent will

communicate that fact to the server. The server will upgrade the document to the top layer in the index module. On the other hand, if a document in the top layer has not been used by any user for a period of time (pre-defined threshold), the server will degrade it to the middle or even bottom layer. The purpose is to ensure the efficiency of the index system.

3.3 Information Filtering

CIB provides a type of information filtering. In CIB the user has the option to 'save' a query. When he saves a query the system remembers which pages the user has examined and which not. Next time the query is executed the system presents the pages in two sections. The first section includes new pages that have not seen before while the second contains pages that have been examined by the user during previous executions of the query. The pages of the second page are sorted according to the user's ratings. The user can specify a minimum rating that a page must has in order to included in the second section. This capability of the saved queries helps also to increase the user's participation in the collaborative information retrieval process. It gives a one more reason to the user to rate the retrieved web pages.

4. CONCLUSIONS

As the WWW becomes a major information dissemination vehicle for the construction industry, there is an urgent need for more effective information search and retrieval tools to help the large number of construction professionals who are often not computer minded. This paper examined the weaknesses of the existing solutions in the form of structured information gateways and Internet search engines. The weaknesses of gateways can be summarised as the lack of universal classification standards, requiring extensive human efforts, limited coverage of information sources, difficulty of dealing with information updates, etc. For the Internet search engines, the main weaknesses include requiring users to be able to formulate query effectively, no quality control over contents, relying on not very accurate search algorithms, and no consideration of construction domain knowledge. For both types of solutions, one of the most important deficiencies is the lack of support for user collaboration and sharing of query techniques and search results.

This paper reported the development of the Construction Information a Broker (CIB), a system that aims at improving the discovery of construction information on the Web. The main innovations are the use of user profiles, user collaboration and the use of construction knowledge during the information discovery process. This paper describes the main technical aspects of the system which is still in its development stage. To ensure the success of such a system, several non-technical issues will need to be addressed such as the user privacy, information ownership and generation of critical mass of users. These issues will be explored in a separate paper in the future.

REFERENCES

Davies J., R. Weeks and M. Revett, 1997, "Jasper: Communicating information agents for WWW", technical report, BT laboratories

Dixon, T.J., 1998, “Building the Web: The Internet and the Property Profession", A College of Estate Management Research Report

Grossman A. David and Frieder Ophir, 1998, "Information Retrieval - Algorithms and Heuristics", Kluwer Academic Publishers, ISBN 0-7923-8271-4

Kowalski Gerald, 1997, "Information Retrieval Systems - Theory and Implementation", Kluwer Academic Publishers, ISBN 0-7923-9926-9

Kristensen J. 1993, "Expanding end-users' query statements for free text searching with a search-aid thesaurus", Information Processing and Management, 29 (6), pp733-744

Lockley R. Stephen and Amor Robert, 1998, "The Construction Information Gateway", Product and Process Modelling in the Construction Industry", pp337-347

Magennis Mark and Cornelis J. van Rijsbergen, 1997, "The potential and actual effectiveness of interactive query expansion", in the proceedings of SIGIR 97, Philadelphia PA, USA

Oard W. Douglas and Marchionini Gary, 1996, "A Conceptual Framework for Text Filtering", technical report, University of Maryland

Stevens Curt, 1992, "Automating the creation of information filters", Communications of the ACM, 35 (12), December 1992

Turk Z. and Amor R., “CONNET – design issues”, in Proceedings of INCITE 2000, Hong Kong, January 2000, pp896-886