HELPING PROFESSIONALS MAKE GOOD USE OF THE INTERNET

David Scott and Chung-kit Pang

Department of Building and Real Estate, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

Searching for data using Internet search engines allows identification and collection of single items of data without difficulty. However, a multi-facet, multi-source search for professional (and perhaps business) purposes is difficult to co-ordinate and often frustrated by either too many possible sources or too few. In many ways this is a problem similar to data/records kept in an old-fashioned filing system, finding a single item with a particular attribute is quite straightforward but identifying groups with a variety of pre-specified characteristics often proves more difficult or impossible.

This paper reports work which has been done to provide services/web-site that collects and stores information and the search systems that are being developed using the Concept search philosophy.

The paper also describes and demonstrates the integration and application of an intelligent Concept search into the Construction and Real Estate Information Service system so that the amount and quality of data is of actual practical use to construction professionals searching for reliable information which is relevant to their work but which is needed within a short period of time.

Keywords: concept search, database, information, Internet, search.

BACKGROUND

Surfing the web is now a well-established sport. However for anyone needing to collect a reliable set of data (e.g. for initial studies/feasibility of a construction project) 'surfing' is neither efficient nor effective.

Construction professionals need fast access to reliable information. Although "random" or "one-off" searches on using the Internet may provide useful and valuable information it can sometimes be very difficult to identify and isolate specific data which is needed to perform a particular task. The problem of finding appropriate information on the Internet is not one of availability because in fact there is now an enormous amount of information available. In fact the problem is one of searching.

Typically a designer or an estimator will have reference and trade handbooks as well as cost data and internal records available. Access and availability of Internet data needs to be just as readily available i.e. stored and accessible in a way which the user can use without having to think about the retrieval method. Results of a search must be reliable and predictable, Internet search engines produce variable results depending on word order, search syntax and keywords. In other words, the collection of data needs to be an automatic process which requires no thought in itself and the searcher can concentrate on the particular information rather than how it is collected.

Scott, D and Pang, C-K (1999) Helping professionals make good use of the Internet. *In:* Hughes, W (Ed.), *15th Annual ARCOM Conference*, 15-17 September 1999, Liverpool John Moores University. Association of Researchers in Construction Management, Vol. 1, 75-83.

It is with these requirements and needs that the Construction and Real Estate information service (CARE-IS) has been designed. It is not intend to be a super-sophisticated system aimed at increasing speed or scope but rather it is designed to be a "one-stop shop" in which users will not be required to use a multitude of different search engines, keywords or search syntax to find the data that they need.

Another difficulty in searching for information on the Internet is that typically a user will find either too much or too little data. In the first case, identifying the most relevant data from the multitude of references needs to allow a systemic reduction by eliminating less relevant material.

A more difficult problem however is a low yield of information. In this situation a "concept" search technique is required which will allow the user to expand their search in a sensible way. Frequently in terms of construction and real estate data this means extending the research to include items which are close to the original search. For example, "closeness" may be geographical (extending search from a particular city to the area surrounding it) or to an area closely related technologically (e.g. including a variety of types of form work rather than a proprietary brand which has been specified.

One feature which is not needed for an effective Internet based information system is more data. Rather than providing yet another information database the system needs to hold information on the location of existing data sources. In this way the data which is held by the information service does not have the problem of updating (which is done by the primary supplier of the information, at source). Rather the problem is to maintain an up-to-date database of data sources which can be then accessed through the system/service.

REVIEW OF REPORTED WORK

Work done and reported on construction and real estate information system has two aspects. The first concerns mainly the modelling of construction related information and use of this information to facilitate the construction process. The second aspect is automatic processing of construction related information.

Rezgui *et al.* (1998) described the central models underlying the COMMIT (construction modelling and methodologies for intelligent integration of information) project. COMMIT addressed six primary issues that are central to information management. These are ownership, rights and responsibilities; versioning of information; schema evolution; recording the intent behind the decisions leading to information; tracking the interdependence between pieces of information; and notification and propagation of changes.

Miranda and Park (1998) developed a prototype learning system for architects. The computer-assisted lesson system was designed to represent architectural concepts related to spatial composition in design by using graphic images and text. The system was evaluated according to observations of the researchers and design studio instructors, test scores from the lessons and student surveys conducted after the lesson was taken. The authors concluded that the access to relevant " information-on-demand" of these lesson systems fills a long felt need of students in design studio. Aouad *et al.* (1995) attempted to provide a framework of information into which the modelling of construction management information fits. This project modelled the construction activities and then integrating planning and estimating activities through object types sharing common subtypes. The behavioural aspect of the objects were

reflected in process models implemented as methods of the objects. The authors concluded that the use of an appropriate methodology helped in defining usable information models for the construction information domain.
The reports cited above are conceptual in nature and do not present concrete implementation details. Sacks and Warszawski (1997) had developed an automated building system, taking user's requirements as input. The system then generates the design brief, followed by preliminary design of the building and its floors, layout of each floor, selection of main work assemblies. A detailed design was also generated covering structure, mechanical systems, electrical systems and various finish works separately up to the level of production details. Finally a construction plan was produced. This consisted of the construction schedule and cost estimate and budget. The automated system relied on a knowledge base. Knowledge was represented in a tri-hierarchical fashion with hierarchies of spaces , work assemblies and activities that are interrelated at each level. According to the authors, the main limitation of this method is that although it produces feasible solutions conforming to general data, codes and design rules, it cannot take into account the multitude of non-explicit or project-specific requirements inherent in any building venture.

It appears that an adaptive knowledge base capable of adjusting to situation-specific considerations will be a useful supplement to any automated building system. It is likely such a knowledge base will be composed of cases of construction and real estate. Chinowsky P. (1998) described the civil engineering resource library. The resource library were defined with cases (a comprehensive collection of documents and video clips covering individual projects through the project life cycle), processes (a collection of indexed entries following CSI format that illustrate the steps, applications, and variations associated with common construction process), materials (a collection of indexed entries that showcase the use and installation of common construction materials including cast-in-place and pre-cast concrete). Initial assessment found the multimedia library concept helpful. However the element of system adaptability was still missing.

Chu *et al.* (1996) proposed Cobase system which was an attempt to formulate the rules for maintaining a knowledge base. Generalization (moving up in the hierarchy), specialization (moving down the hierarchy) and association (moving between hierarchies) are the three key operations in deriving co-operative query answer for the users. Data are organized into Type Abstraction Hierarchies (TAH). TAH provided multi-level knowledge representation and can be generated automatically from data sources. Further, TAHs are user and context sensitive, and easy to customize and maintain. According to the authors, CoBase stemmed from the transportation planning application. However technology transfer to other domains, such as construction, is possible. It seems that the concepts presented in CoBase are useful for the CARE-IS but this is dependent upon the availability of expert rules in the construction and real estate information domain for the automatic indexing and maintenance of the knowledge base.

In fact a review of recent researches on knowledge base applications reveals that many of them concentrated on general clustering algorithms. Tam (1993) discussed conceptual clustering approach to automatically generating a concept hierarchy. Kelly *et al.* (1996) proposed a new scatter-search-based learning to train feed-forward neural networks. Zhong and Ohsuga (1996) described a hierarchical model learning approach for refining/managing concept clusters discovered from database which in turn were represented as Multi Layer Logic formulae. Similarly, Ioannidis *et al.* (1992)

examined the idea of incorporating machine learning algorithms into a database system for monitoring its stream of incoming queries and generating hierarchies with the important concepts expressed in those queries. Lee and Tung (1997) presented a new method for clustering the words in a dictionary into word groups. The word groups were used in a language model to improve Chinese character recognition accuracy. Chang (1996) further developed the simulated annealing clustering algorithms in discovering word classes in a Chinese class n-gram model which could be used for contextual post-processing of hand written Chinese character recognition. There is little touching on the specific domain of construction and real estate information.

With the advance and growing popularity of internet services, a third dimension can be added. Given the distributed nature of resources and services on the internet, special techniques are required to enable efficient and effective delivery of information to the target users. Chandler *et al.* (1997) proposed an expert system with a knowledge base and a subject thesaurus, a model of reference librarian's behaviour for cataloguing and searching, and an intelligent user interface to provide context sensitive help to the user. The system was intended for cataloguing and searching digital libraries on the World Wide Web. Gaines and Shaw (1997) reported the development of knowledge management tool operating through the web to support knowledge acquisition, representation and inference through semantic networks and repertory grids. De Bra and Post (1994) described two generic types of web searching tools: a gateway using a precompiled database often built by an automatic web robot; a client based search tool that does automated navigation, working more less like a browsing user. There were some disadvantages associated with client based navigation - slow speed and high network resource consumption - which could be reduced by combining the fish search algorithm with a cache. Chen and Chang (1997) proposed the use of smart page with associated knowledge to react to pre-defined events and perform appropriate actions. Fielding (1994) mentioned the complex infostructure on the web and described the Multi-Owner Maintenance spider (MOM spider). MOM spider could periodically traverse a list of webs (by owner site, or document tree), check each web for any changes which may require its owner's attention, and build a special index document that lists out the attributes and connections of the web in a form that can itself be traversed as a hypertext document.

It is apparent from the above review that there is a need for an adaptive knowledge base with both source of data and target sites on the Internet in the Construction and Real Estate (CARE) domain. A searching process is needed in such a system that will develop expert rules for building and searching a knowledge base from sources of data on the Internet. Work on this problem is described in the following sections. Specific attention is paid to

1. application of artificial intelligence to automate the generation of concept hierarchy,
2. in the construction and real estate information domain and
3. the concept of hypermedia usability Smith *et al.* (1997).

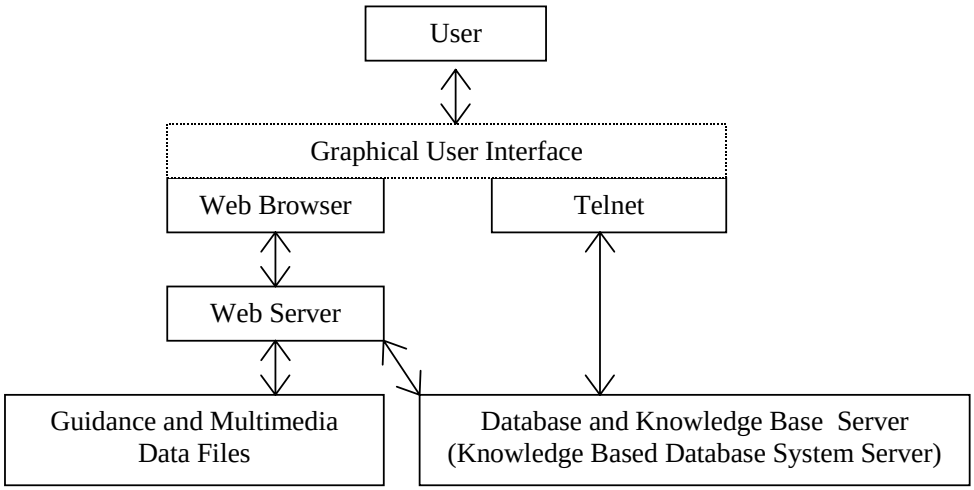


Figure 1: The architecture of the Internet-based and intelligent search system

ARCHITECTURE OF THE INTERNET-BASED AND INTELLIGENT SEARCH SYSTEM

An information system has been designed and built which holds various databases which are integrated to provide a comprehensive information service. The data (Scott 1998) held is in a variety of formats and the medium of transmission and retrieval is the Internet (http://care.bre.polyu.edu.hk)

The conceptual structure of this Internet-based information system, Construction and Real Estate Information Service (CARE-IS), is shown in Figure 1. The system consists of a database, multimedia data files, guidance data files, a knowledge base and a Web server. This system has been developed and put in place on a departmental LAN and can be accessed by users through Internet connections.

1. Database: The database stores data according together with a corresponding knowledge base.
2. Multimedia data files: Multimedia data are not stored in database directly. Only the URLs (Uniform Resource Locators) are stored in the database, the actual files are housed in their home site. Users can get access to the source data through the Web browser.
3. Guidance files: These are used to help to guide the user in their search. This assistance is in the form of either conventional searches or Concept searches (described in following sections).
4. The knowledge base and the knowledge based database system: The knowledge base stores the relevant knowledge of CARE data that has been built up through earlier searches. The knowledge in the knowledge base corresponds to individual attributes or concepts. For example; in the attribute “location”, which holds the name of a city, the knowledge about the relationships between that geographical location and nearby cities, provinces and countries is stored. More detail is presented in later sections.
5. Web server: The Web server links up with the database and knowledge base server, houses the guidance and multimedia files, and transmits the required data and information to the Web browser when requested by users.

THE CONCEPT SEARCH SYSTEM FOR CONSTRUCTION AND REAL ESTATE INFORMATION

At present CARE-IS supports searches using conventional search engines. A more advanced user friendly, professional-use orientated system for the automatic acquisition, organization and retrieval of information is still needed and is being developed. This will help professionals to gather and utilize the current massive amount of construction data now available on the Internet.

The following sections describe work done in the area of Internet information acquisition, organization and retrieval in the Construction and Real Estate domain. The concepts derived from this development will be applicable for CARE information stored in other formats, e.g. Relational Database Management System (RDBMS).

Information on the Internet is always changing as the information and data sites are constantly being updated and new data added. One of the most common services provided on the Internet is Hypertext Transfer Protocol (http). Using http information data streams can be easily transmitted around the world-wide network. These information data streams are characterized by a list of keywords.

This characteristic is used in the searching system reported here. Each data stream acquired through http can be placed in a single "concept". Each concept can be described fully by a list of keywords. For example, the concept "materials" in the CARE domain can be expanded into a list of keywords (cement, sand, water, aggregate, mortar, etc.). In this way, concept manipulation is equivalent to key word list processing. Relevance can be determined by the number of key word matches. The higher the number of key word matches, the more relevant is a data stream to a concept.

Concepts can themselves be linked to other concepts which are similar or related in content or meaning. These clusters of concepts form an inter-linked or overlapping network of coverage of a topic. Clustering the concepts is the key to successful searches. The appropriate linkages of concepts are built up in a 'training' process of searching and updating. In this study, concept clustering is done by manipulating lists of keywords. Clusters are formed by optimizing an objective function so that each concept cluster contains similar concepts. In this case, similar concepts are taken to mean that data streams containing similar (but not identical) key word matches.

CARE information retrieval can be done by matching the concept or its equivalent list of keywords. Unlike traditional key word search, however, concept retrieval has more tolerance to variation in meanings of the concept. Rather than requiring exact match of all the keywords in the list for a particular concept, results can be obtained by partial match as defined by a threshold. The threshold is the amount of difference which will be tolerated.

PRELIMINARY TESTS OF THE SEARCHING SYSTEM

The list of keywords as is denoted as a feature vector. To assess the usefulness of feature vector in concept processing, the following procedure has been used.

(i) start with an initial feature vector to search the internet
(ii) obtain the first 100 search results
(iii) assess the relevance of the search results by inspection
(iv) count the number of key word matches in all the relevant search results

Table 1: Results of prototype search

Keywords in feature vector	Word count in step (iv)	Word count in step (viii)	Word count in step (ix)
Cement	19	14	17
Sand	11	4	4
Water	9	9	8
Material	3	1	None found
Damp	1	3	3
Proofing	2	3	3
Chunam	0	0	None found
Mortar	None found	12	16
Lime	None found	6	9
Concrete	None found	13	17
Membrane	None found	0	None found
Total Count	**45**	**69**	**77**

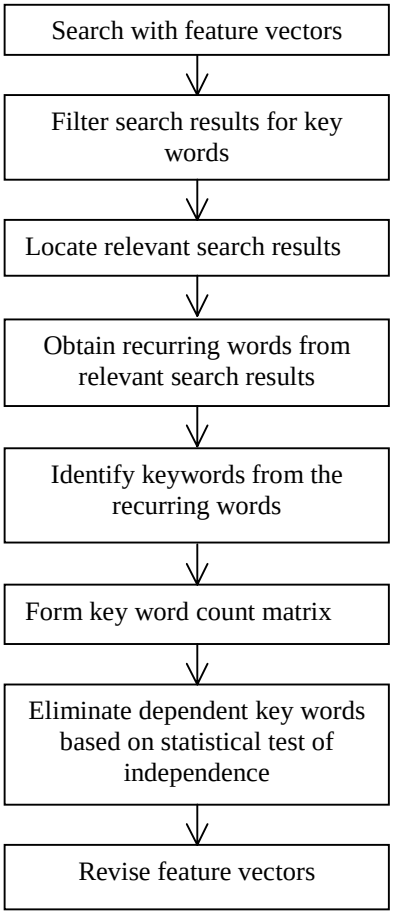


Figure 2: Prototype system for concept construction

(v) revise the feature vector with recurring keywords in the relevant search results
(vi) search the internet again with the revised feature vector
(vii) assess the relevance of the search results manually
(viii) count the number of key word matches in all the relevant search results
(ix) repeat steps (v) - (viii)

RESULTS AND EFFECTIVENESS OF THE PROCEDURE

A typical set of the results of the procedure described in the previous section is shown in Table 1

From Table 1 these results, it is observed that number of key word matches increases as more and more relevant search results are presented to the procedure. In other words, the feature vector adapts with presentation of more and more training material.

Another observation after testing the process is that increasing the number of keywords in a feature vector does not necessarily produce the best search result, measured by the total count of key word matches in all relevant search results.

WORK IN PROGRESS

To test the validity of the theory proposed, a prototype system for integration into the CARE-IS system is under construction. The prototype consists of the components shown in Figure 2.

In this system the key word count is incremented by each coexistence of keywords in a data stream. For example, if “cement” and “sand” are found to be in the same data stream, the key word count at the intersection of the “cement” row and the “sand” column is increased by 1.

The prototype will be used to test the effectiveness of the searching system for the following procedures:

(i) automatic acquisition of internet CARE information through feature vector revision
(ii) automatic clustering of CARE concepts through key word count matrix and statistical test of independence
(iii) assessing the usefulness of CARE information retrieval by concept through key word relaxation.

The functioning of the Internet-based and intelligent system for CARE information

Using the system which has been developed with the concept search system users can review the data and the information of the historical projects by questioning the database with the help of the knowledge stored in the guidance files and in the knowledge base, from any computer which is connected to the Internet.

In this way the user can have access to information stored both in the CARE-IS system and throughout the Internet. The user will not need specialist knowledge of Internet searching or search engines. They will have access to a “one stop shop” which has been trained to identify sites which will provide relevant and related information.

ACKNOWLEDGMENT

The authors would like to thank UGC for their support of this project.

REFERENCES

Aouad, G., Kirkham J., Brandon P., Brown F., Child T., Cooper G., Ford S., Oxman R., Young B., (1995), The conceptual modelling of construction management information. *Automation in Construction*. **3**(4), 267–282.

Chandler, P., Shinghal R., Desai B., Radhakrishnan T., (1997) An expert system to aid cataloguing and searching electronic documents on digital libraries. *Expert Systems with Applications*. **12**(4), 406–416.

Chang, C. (1996) Simulated annealing clustering of Chinese words for contextual text recognition. *Pattern Recognition Letter*. **17**(1), 57–66.

Chinowsky, P. (1998) The civil engineering resource library. *Journal of Computing in Civil Engineering, ASCE*. **12**(4), 208–214.

Chu, W., Yang, H., Chiang K., Minock, M., Chow, G. and Larson, C., (1996) CoBase: A scalable and extensible co-operative information system. *Journal of Intelligent Information Systems*. **6**(2-3), 223–259.

De Bra, P. and Post, R. (1994) Information retrieval in the World Wide Web: making client-based searching feasible. *Computer Networks and ISDN Systems*. **27**(2), 183–192.

Fielding, R. (1994) Maintaining distributed hypertext infostructures: welcome to MOMspider's Web. *Computer Networks and ISDN Systems*. **27**(2), 193–204.

Gaines, B. and Shaw, M. (1997) Knowledge acquisition, modelling and inference through the World Wide Web. *International Journal of Human-Computer Studies*. **46**(6), 729–759.

Kelly, J., Rangaswamy, B. and Xu, J. (1996) A scatter-search-based learning algorithm for neural network training. *Journal of Heuristics*. **2**(2), 129–146.

Miranda, V. and Park, T. (1998) Representation of architectural concepts in the study of precedents: a concept-learning system. *Automation in Construction*. **8**(1), 99–106.

Rezgui, Y., Cooper G. and Brandon P. (1998) Information management in a collaborative multi-factor environment: the COMMIT approach. *Journal of Computing in Civil Engineering ASCE*. **12**(3), 136–144.

Sacks, R. and Warszawski, A. (1997) A project model for an automated building system: design and planning phases. *Automation in Construction*. **7**(1), 21–34.

Scott, D., Mak, S. and Tang, K.W. (1998) An internet-based and intelligent system for training project managers. *In:* Yang, J and Chang, W.P. (eds) *Building Education and Research: Proceedings of Building Education and Research Conference, Queensland University of Technology, Australia, 8-10 July 1998.* London: Spon, 438–448.

Smith, P., Newman I. and Parks, L., (1997) Virtual hierarchies and virtual networks: some lessons from hypermedia usability research applied to the World Wide Web. *International Journal of Human-Computer Studies*. **47**(1), 67–95.

Tam, K. (1993) Applying conceptual clustering to knowledge-bases construction. *Decision Support System*. **10**(2), 173–198.

Zhong, N. and Ohsuga, S. (1996) A hierarchical model learning approach for refining and managing concept clusters discovered from databases. *Data and Knowledge Engineering*. **20**(2), 227–252.