LEARNING SEARCH PATTERN FOR CONSTRUCTION PROCUREMENT USING KEYWORD NET

Ren Jye Dzeng, Shyh Shiuh Wang
Department of Civil Engineering,National Chiao-Tung University, Taiwan

Abstract: As more and more procurement websites become available on the Internet, seeking information from websites has become an essential part of a contractor's procurement undertaking. Several e-markets, specifically for construction, have also been established, including bLiquid.com and ProcureZone. However, most websites provide only two primary ways of searching for information, namely by index/menu or by keyword. Instead of relying on the primitive search engines found in most procurement websites, a search guide system could help a user's keyword search by reducing the number of keywords required to find the desired information. Our research recognized that professional procurement experience helped users more effectively carry out website information searches, by using fewer keywords. We planned to capture such experience in order to guide inexperienced users in their search. The research goal was to improve search effectiveness by guiding the user's search using three approaches; namely correction, specification and extension. Based on these three approaches, this research applied the following guides: correction; specification-by-equivalence; specification-by-detail; extension-by-time; extension-by-location; extension-by-team; and extension-by-component. The paper will describe how we classified users for learning credibility, and the learning framework for recording expert users' search patterns. Twelve professionals, using 14 procurement packages, with 64 items in total, evaluated the proposed framework. It will be demonstrated that the proposed learning keyword guide facilitated a dynamic, customized menu and indexing system, and reduced the number of keywords required for the professionals to find the information they desired.

Key words: Construction procurement, information search, machine learning, e-commerce, knowledge acquisition

1. INTRODUCTION

Traditional contractors frequently maintain and send RFQs only to a few qualified suppliers for each type of procurement item by fax or telephone. Since the proliferation of the Internet in the second half of the 1990s, almost all electronic project document management systems have migrated to using the general Internet as their physical network, project-specific Web servers as their storage media and Web browsers as the main platform for buyer interfaces [1]. Several e-markets specifically for construction have also been established, including AEC Info [2], BuildPoint [3], bLiquid.com [4], Citadon, Inc. [5]. The proliferation of Web-based project management platforms and e-markets has given contractors more business opportunities and a wider selection of suppliers, but has also challenged contractors in managing the flood of electronic information.

Commercial construction procurement websites have aimed to provide all relevant procurement information on a single website to attract buyers with one-stop shopping. However, most websites provide only two primary ways of seeking information - by index/menu and by keyword. Most users who are unfamiliar with the content or indexing scheme of a website prefer keyword search. Keyword search may also be more efficient when too many indexes exist, the website hosts a large body of information, or a buyer does not know the specific procurement terms used by the website. Nevertheless, most buyers may have to use a trial-and-error process of inputting keywords to narrow down the search results to the desired information, even though they are professionally experienced.

General-purpose websites, such as portals like AltaVista [6] or Google [7] also offer information indexing and keyword search functionality. However, these sites deal with general-purpose Web pages, with unstructured data instead of domain-specific, structured data as found in the procurement websites, and perform a full-text search based on input keywords. They sort search results based on a pre-specified algorithm, and do not enable buyers to choose how to sort the results.

A construction procurement website must host various structured procurement information and also unstructured data like that found on general-purpose websites, to become a one-stop website that enables buyers to complete a procurement task. Hence, a predetermined indexing system is typically impractical. Also, each type of procurement item may require different specification fields, so establishing a generic data table and covering all items is difficult. The usefulness of applying several predetermined multiple fields becomes limited when on-line suppliers offer various construction items.

Until now, researchers have been developing intelligent guides and agents to help buyers navigate Web pages or find information concerning consumer-product e-marketplaces. However, even for the best knowledge of the authors, no intelligent guide has been developed specifically for the construction procurement website domain. Thus, this work presents an innovative guide system that captures the domain-specific keyword search patterns of expert buyers via a learning framework. When an inexperienced buyer inputs a search keyword, the guide may replace the keyword with correct or more specific ones, or add keywords that target related procurement items. An evaluation that involves 12 professionals and 64 procurement items demonstrates that the proposed guide system enables a dynamic, customized menu/index and reduces the number of keywords required by the professionals to find desired information in conventional websites that have no guide.

2. INFORMATION SETS

A search process generally comprises a series of search sessions. In each session, the buyer inputs a series of keywords to locate the information of interest, termed the target information set (I_{target}). The search result expected by the buyer when he inputs a keyword is termed the anticipated information set ($I_{anticipated}$). I_{target} differs from $I_{anticipated}$ because the buyer lacks confidence in controlling the search; this difference is governed primarily by the buyer's familiarity with the website. After a keyword is entered, the actual search result of the website is termed the result information set (I_{result}). $I_{anticipated}$ differs from I_{result} because the search is not executed as the buyer expects; this difference is influenced mainly by the design of the website. I_{result} differs from I_{target} when the search engine's interpretation of the input keyword differs from that of the buyer; this difference thus depends mainly on the buyer's experience of procurement, assuming that the website is professionally designed. Figure 1 presents the difference among information sets.

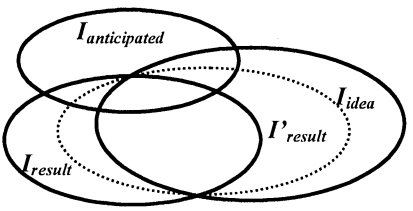

Figure 1 procurement in an open e-market

The following search scenario illustrates the information sets. A buyer wants to find suppliers of H-type steel and enters the search keyword, "H steel" on a procurement website. The search outcome includes all H-type steel suppliers' nationwide and other information on "H steel", including work specifications for H-type steel. In fact, the buyer is looking only for H-type steel manufacturers in the northern area. Therefore, the buyer must perform the search again by entering a specific keyword "H steel *State-A*" as *State-A* is the state of interest in the northern area. Browsing the URL (Unified Resource Locator) references of the search results, the buyer finally finds the suppliers of H-type steel with dimensions 400x400, 13mm and 21mm.

This scenario describes a search session that includes the process of inputting two keywords. I_{target} is "the list of suppliers in the northern area that sell H-type steel". $I_{anticipated}$ is "the list of suppliers of H-type steel" when the first keyword was entered, and "the list of suppliers in *State-A* that sell H-type steel" when the second keyword was entered. I_{result} includes all H-type steel suppliers nationwide along with work specifications, codes, buyer calls and news about H-type steel. I_{idea} is "the list of suppliers in the northern area that sell 400x400-13mm-21mm H steel".

The search effectiveness refers to the hit rate, the proportion of URL references in the search result that provide the target information. The search efficiency refers to the number of keywords that must be input to yield a satisfactory search result. This work focuses on improving search efficiency.

3. KEYWORD NET

The proposed guide system is based on a learning model that builds search guide knowledge by learning from the search patterns of expert buyers. Figure 2 illustrates a typical search for completing specifications for some procurement packages. The buyer inputs a keyword, K_{11}, to find information about a procurement item. The result shows no URL reference, and the buyer then realizes that he has just misspelled a word. He corrects the mistake by inputting another keyword, K_{12}. However, the result now contains too many URL references, so the buyer performs the search again by inputting a more specific keyword, K_{13}, and thus obtains satisfactory references. He inputs another keyword, K_{21} to find another item in the same procurement package. He repeats the process until he inputs keyword K_{24} and finds satisfactory references. The information found so far enables the buyer to complete the procurement documents for the first package. The buyer may still continue the search, starting with keyword K'_{11}, to complete specifications for other packages.

The processes from K_{11} to K_{13} and from K_{21} to K_{24} are two search sessions, each of which seeks information pertaining to an item. The process from K_{11} to K_{24} represents a series of keywords input to complete a procurement package. Such a process is called a learning session because the keywords in this session are related, and their relationships can be learned to support the corresponding proposed guides with appropriate annotation by expert users. Figure 3 presents a net example that connects related keywords using directed links, including *correction*, *equivalence*, *detail*, *time*, *location*, *team* and *component*.

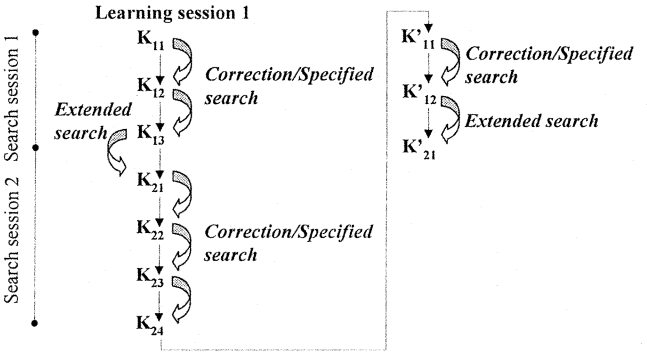

Figure 2 Search session and learning session

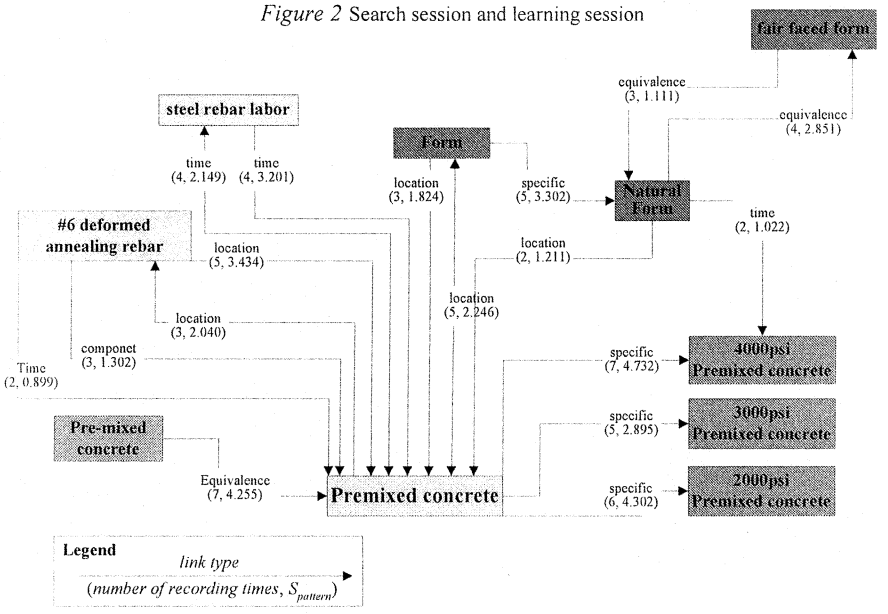

Figure 3 Example of keyword net

4. KEYWORD SEARCH PATTERN

In a learning session, an expert user annotates the relationship between two consecutive recorded keywords that have been input using the links. Each pair of linked keywords constitutes a keyword search pattern unit. Each unit includes a preceding keyword, a succeeding keyword, the type of link that connects two keywords and other log information, including the number of times that the pattern has been learned and the time since the last recording of the pattern.

Search patterns with the same pair of keywords may have a link recorded multiple times, multiple links of different types or bidirectional links of the same type. Appropriate ranking of suggested keywords, similar to the webpage ranking of Google [8], reduces the time the buyer needs to select one from them. The learned outbound links from a keyword point to keywords that the buyer might need in a subsequent search. The proposed guide system ranks suggested subsequent keywords by considering the number of links from the preceding keyword; the number of links to the succeeding keyword, the experience of the buyer who is annotating the link, and the number of times that the buyer has actually adopted the keyword suggested by the link. Suppose that a user adds a link between two keywords, K_1 and K_2. When the user inputs K_1 next time, the proposed guide uses Eq. (1) to calculate the rank of K_2.

$$\begin{aligned} & R\ (K_1, link, K_2, user, \Delta t) \\ & = w_1\left(S_{pattern}(K_1, link, K_2, user, \Delta t)\right) + w_2\left(S_{keyword}(link, K_2)\right) \\ & = w_1\left(p \sum_{(K_1, link, K_2)}(user) + q\left(1/\min \Delta t\right)\right) + w_2\left(\frac{Q_{link}(link, K_2)}{Q_{link}(link)}\right) \end{aligned} \quad (1)$$

$R(\ )$: rank of K_2 given K_1 as the initial input;
link: type of link that connects K_1 and K_2;
user: level of experience of the user (the default setting is 1 for expert user, 0.7 for experienced user, 0.4 for inexperienced user and 0.1 for novice user);
Δt: (the current date) – (the date when the link was recorded)
w_1, w_2: weights determined by the system user, $w_1 + w_2 = 1$, $0 \le w_1, w_2 \le 1$;
$S_{pattern}()$: strength of the pattern (K_1, *link*, K_2);
$S_{keyword}()$:strength of K_2 from the perspective of *link*;
p, q: weights determined by the system user, $p + q = 1$, $0 \le p, q \le 1$;
$Q_{link}(link, K_2)$: number of links of the *link* type to K_2;
$Q_{link}(link)$: total number of links of the *link* type in the keyword net;

Equation (1) comprises two parts - $S_{pattern}$ and $S_{keyword}$. $S_{pattern}$ concerns the importance of the pattern and the time since the pattern was learnt.

Table 1 shows the calculation of $S_{pattern}$ for pattern ("premixed concrete", *location*, "#6 deformed rebar"). The pattern comprises three *location* links, indicated by Fig. 3 and detailed by Links 1, 2, and 3 in the table. The most recent link was recorded by an expert user two days ago, and the oldest link was recorded by an expert user six days ago. Assuming $p = 0.7$ and $q = 0.3$, $S_{pattern}$ is 2.040 for "#6 deformed rebar". Table 1 also shows that $S_{pattern}$ for pattern ("premixed concrete", *location*, "form") is 2.246.

$S_{keyword}$ measures the importance of the succeeding keyword K_2 from the perspective of the link of interest. More links of the type to a keyword indicate a higher probability that the keyword is the next one to be needed by the user. The net in Fig. 3 comprises 18 (= 3+5+3+5+2) *location* links, of which three enter "#6 deformed annealing rebar" and five enter "form". Thus, "form" ($S_{keyword}$ = 5/18) is more likely than "#6 deformed annealing rebar" ($S_{keyword}$ = 3/18) to be the next *location*-related keyword required by the user.

Table 1. Spattern calculation for patterns

	"premixed concrete" location "#6 deformed rebar"			"premixed concrete" location "form"				
Link #	1	2	3	1	2	3	4	5
user	1	0.7	1	1	0.7	0.7	1	1
Δt	2	5	6	1	3	4	6	6
p×∑user	1.89			3.08				
q×(1/minΔt)	0.15			0.3				
Spattern	2.04			2.246				

Assuming w_1= 0.7 and w_2 = 0.3, R is 2.04×0.7+(3/18)×0.3=1.478 for "#6 deformed rebar" and 2.246×0.7+(5/18)×0.3=1.656 for "form". Hence, "form" is ranked higher than "#6 deformed rebar" in the list of suggested location-related keywords with respect to "premixed concrete".

5. SYSTEM IMPLEMENTATION AND EVALUATION

The system is implemented on the Microsoft Windows 2000 Server platform with Internet Information Server v5.0, using PHP [9] and MySQL [10]. The system database consists of three categories of procurement-related data, including suppliers, supply catalogs and specifications.

The evaluation described herein focuses on the potential reduction in the number of input keywords. The experiment conducted involved monitoring 12 procurement engineers as they performed procurement tasks for a US$

9,000,000 project to construct a school building, which involved 14 procurement packages (such as soil preparation and excavation) and a total of 64 procurement items (such as an excavator, unskilled laborers, Type IV Portland cement). Descriptions of the procurement items given to the participants were taken directly from the actual contract. The participants' procurement experience ranged from one to 11 years, and web search experience ranged from under one to seven years.

A pretest was performed to divide the project into two "equivalent" subprojects, which both required approximately the same number of input keywords, to reduce the learning effect. Fifteen graduate students participated in the pretest; each was required, for each procurement item, to complete the procurement task that includes finishing item specifications and RFQs, as well as finding contact information of three prospective suppliers. The quantity takeoff for each procurement item was given. Each participant could find information from the system only by inputting search keywords and browsing the resulting URL references. The items were then divided into two groups, namely Group-1 and Group-2, according to the average number of input keywords used for each item. The difference between the average numbers of input keywords for the two groups was less than 1%. Therefore, the two groups of items were assumed to require the same amount of search work.

In the following experiment, the 12 participating engineers used the system to complete the procurement tasks for the items in Group-1 without a guide, and for those in Group-2 with the guide. Only 12 engineers could participate in this experiment because each procurement task took a long time. Thus, the result may lack conclusive statistical meaning due to the limited number of samples. Nevertheless, the following statistics help to explicate the extent of the efficiency improvement of the guided search.

Table 2. Average number of search keywords used with and without guide

Group-1 (without guide)									
Procurement package	Site preparation	Survey	Excavation	2500psi concrete	Brick laying	Rebar assembly			Total
procurement items	$\bar{}$ 4	3	2	9	4	4			26
keywords	$\bar{x}$ 18.58	15.67	10	33.17	17.08	27.5			122
	σ 2.84	4.39	3.38	7.2	2.75	5.71			
Group-2 (with guide)									
Procurement package	Pebble pavement	Backfill	3200psi concrete	Form-work	Floor tiling	painting	Water-proofing	Wall tiling	Total
procurement items	$\bar{}$ 3	1	8	5	6	5	5	5	38
keywords	$\bar{x}$ 6.83	1.75	12	19.58	12	13	10.42	14.25	89.83
	σ 1.95	0.97	3.15	2.16	1.1	1.09	1.56	1.45	

Table 2 lists the average numbers of search keywords with and without the guide for each work package of Group-1 and Group-2. The number of procurement items in a work package ranges from one to nine. On average, the total number of keywords used in the procurement task is 122 for Group-1 (without the guide), and 89.83 for Group-2 (with the guide); these values represent a reduction in the number of keywords of 32.17 and 26%.

Table 3 shows cross analyses of data based on the participants' experiences. For instance, the average number of keywords used to complete the procurement task without the guide is 101.67 for the half of the participants with the most professional procurement experience, measured in years, and 142.33 for those who have less experience. With the guide, the average number of keywords used by more experienced participants is 75.67 and the number used by less experienced participants is 104.00. Thus, the guide reduces the number of keywords used by more experienced participants by 26 (26%), and by less experienced participants by 38.33 (27%).

Table 3. Comparison of reductions in numbers of keywords used by two groups of participants

		Group			
		More procurement-experienced	Less procurement-experienced	More Web-search-experienced	Less Web-search-experienced
Number of keywords	No guide	101.67	142.33	121.67	122.23
	With guide	75.67	104	88	91.67
Saving	# of keywords	26	38.33	33.67	30.67
	%	26%	27%	28%	25%

The findings in Tables 2 and 3 can be summarized as follows. First, procurement experience considerably reduced the number of keywords used in the search, but web-search experience did not. Engineers with more procurement experience used fewer keywords to complete the task without the guide than did less experienced engineers (101.67 vs. 142.33); but the number of keywords used by engineers with more experience of web-searches was approximately that used by those with less experience of web-searches (121.67 vs. 122.23).

Second, the guide reduced the number of input keywords by about 26%. The reduction was greater for engineers with less experience of procurement than those with more experience (38.33 vs. 26.00) because less experienced engineers used more input keywords originally; however, the difference between the percentage reductions was insignificant.

6. CONCLUSIONS

This work developed a guide system that suggests possible subsequent keywords based on a keyword initially input by a user. A subsequent keyword is predicted according to a learning net that connects the recorded keywords using links, which are annotated by expert users during their search. The system offers seven guides - *correction*, *specification-by-equivalence*, *specification-by-detail*, *extension-by-time*, *extension-by-location*, *extension-by-team* and *extension-by-component*. The order of the suggested keywords is calculated by considering the number of links from the input keyword; the number of links to the subsequent keyword that is being considered, the experience level of the user who annotates the link, and the number of times that the user has actually adopts the keyword that was suggested based on the link.

The evaluation experiment demonstrated that users with more procurement experience required fewer input keywords than those with less experience. The number of input keywords is independent of the users' experience of web-searches. The guide system reduced the number of keywords input without the guide by 26%.

ACKNOWLEDGEMENT

The authors would like to thank the National Science Council of Republic of China, for financially supporting this work under Contract No. NSC-92-2211-E009-064.

REFERENCES

1. B.C. Björk, "Electronic document management in construction- research issues and results," Electronic Journal of IT in Construction, vol. 8, pp. 105-117, 2003.
2. AEC Info (2004). Available: www.visualibrary.com.
3. BuildPoint (2003). Available: www.buildpoint.com.
4. bLiquid.com (2003). Available: bliquid.com.
5. Citadon, Inc. (2004). Available: www.citadon.com.
6. AltaVista (2004). Available: www.altavista.com.
7. Google (2004). Available: www.google.com.
8. P. Craven (2004). Google's PageRank explained and how to make the most of it. Web Workshop: http://webworkshop.net/pagerank.html.
9. PHP (2003) PHP: Hypertext Preprocessor. [Online]. Available: http://www.php.net
10. MySQL (2003). Available: http://www.mysql.com.