MODELLING INFORMATION SEEKING BEHAVIOUR OF AEC PROFESSIONALS ON ONLINE TECHNICAL INFORMATION RESOURCES

RECEIVED: March 2003
REVISED: September 2003
PUBLISHED: October 2003 at http://www.itcon.org/2003/20/
EDITORS: Robert Amor and Ricardo Jardim- Gonçalves

Sameh Shaaban
University of Newcastle upon Tyne, UK
email : s.shaaban@ncl.ac.uk

John McKechnie
RIBA Enterprises ltd, UK
email: john.mckechnie@ncl.ac.uk

Stephen Lockley, Prof.
RIBA Enterprises ltd. UK
email: Stephen.lockley@theNBS.com

SUMMARY*: With the increasing popularity of architectural, engineering and construction (AEC) online information resources, studies have emphasized the need for domain specific systems that acknowledge both the user's information tasks and skills. This study concentrates on analysing the users' information behaviour when involved in an online information seeking session. This analysis aims to find out whether there are patterns of information seeking behaviour among the AEC professionals. The study is based on a live web-based information access system, which contains a large collection of technical AEC documents. Web transaction logs, of around 200,000 user sessions, were gathered and statistically examined. Cluster analysis methods have been performed in order to find the optimal natural groupings of information seeking behaviours among the system users. Results shows the popularity of the 'exploring results' and 'simple searching' activities among all users. Common usages of short queries have also been noted. Four clusters of user seeking behaviour have been found. The statistical characteristics of each cluster as well as the authors' interpretations of their common usage patterns have been discussed.*

1. AEC ONLINE INFORMATION

Online Information access systems are becoming increasingly popular sources of information in the AEC industry. Surveys show the high increase of the use of online resources among the industry professionals as a source of information, (RIBA Enterprises, 2002; Barbour, 2000). They share a common advantage, as reported by user surveys, which is they provide a medium for rapid and up to date industry information. Developments in AEC online information systems have taken three basic forms, mostly based on the type of information they provide. They are scientific literature, product catalogues and technical documents.

Rich and structured published information is, inevitably, the foundation for successful information delivery. It facilitates the design of efficient online information access systems. Research efforts have identified several factors that influence the information delivery as well as the development of online AEC information sources (Amor et al, 1996; Kelly et al, 1997; Augenbroe, 1998; Lockley et al, 2002; Björk and Turk, 2000). The collated information from the literature, together with our experience from developing online fielded AEC information systems, has led to identify the following four groups of factors:

Information authoring: A key determinant to the design of information resources often referred to as data models or underlying data structures. The AEC industry research is caught in the middle between two directions of data modelling. Core models, e.g. product and process models, model deep knowledge areas but are difficult to apply to large data sets (Froese, 1996; Tolman, 1999; Eastman, 1999). Classification models that model shallow knowledge areas and are mostly applied to wide data sets, e.g. UNICLASS in UK and BSAB in Sweden (Crawford, 1997; Ekholm et al, 2000). The diversity of building standards and regulations and the variety of international constraints has complicated the aim of developing unified models that could represent the AEC

knowledge base. Consequently, none of these models has yet been implemented successfully for current information resources (McKechnie et al, 2001).

Market pressure: A factor that could be considered out of the scope of the academic research arena. However, it directly influences the development of information resources in terms of providing funding and incentive for improvement. For example, it is acknowledged that users are not willing to pay for building products information and trade literature. This led manufacturers to publish their own biased information on the internet. The opposite scenario is emerging regarding the technical document information resources. The rising customer demand for such systems in the UK has resulted in competition to provide better services, backed by scientific research, between information systems, e.g. construction information services (McKechnie et al, 2001) and Construction Expert (Barbour, 1999).

Infrastructure: A major concern when developing on-line information resources. The Speed of connection as well as allocated band width affects the performance of any information system. The emerging online technologies for improving information access systems, e.g. information visualisation or raster image detection, are striving to comply with minimum connection configuration.

User involvement: The swift development of online environments has been accompanied by a rather sceptical and slow change of information seeking culture among industry users. Where traditional information resources were deemed to be easy to use and navigate through their paper form, users now have to acquire new searching skills. The domain independent developments in search engines have contributed to the users' lack of enthusiasm. It has resulted in overloading the users with information that could be irrelevant to the construction domain. It also ignores the users' information needs and domain implicit information searching tasks. It expects users to learn advanced query languages in order to find the items of information they are looking for.

2. FACETS OF INFORMATION SEEKING BEHAVIOUR

A person engaged in an information seeking session performs two distinctive tasks: information seeking and information retrieval (Marchionini, 1995; Hearst, 1999). While seeking is characterised as a more human oriented and open ended process, retrieval implies that the object must have been known at some point; most often had been previously organised for later use. Seeking connotes the process of acquiring knowledge. It is more problem oriented as the solution may or may not be found. It is closer to answering questions or learning. According to Marchionini, (1995) information seeking behaviour could be organised into four levels of granularity, (Figure 1):

1. At the coarsest level, people exhibit information **seeking patterns**. Patterns are mostly unconscious sequences of behaviours that can be discerned over time and across different information problems and searches. They are influenced by user disciplines, domain and systems.
2. **Strategies** are the approaches that information seekers take to a problem. Two classes of strategies are formalised as analytical searching and browsing strategies. They are the extremes of a range of flexible combinations of strategies (Belkin et al, 1993). Strategies mostly are consciously selected and mainly search specific.
3. **Tactics** are discrete intellectual choices during an information seeking session. Tactics are more focused than strategies, for example narrowing the search space by selecting a date range. Tactical skills clearly distinguish between expert and novice users of on-line systems, are often mentioned as searching skills.
4. **Moves** are finely grained actions manifested as discrete behavioural actions, e.g. doing search, going to advanced search, downloading a document, or even clicking a mouse. Moves are evidences of tactics. They offer observable clues for interface usage and mapping the intellectual activity at higher levels of action. This study concentrates collecting user moves data and aims to utilise it in order to build models of their seeking patterns.

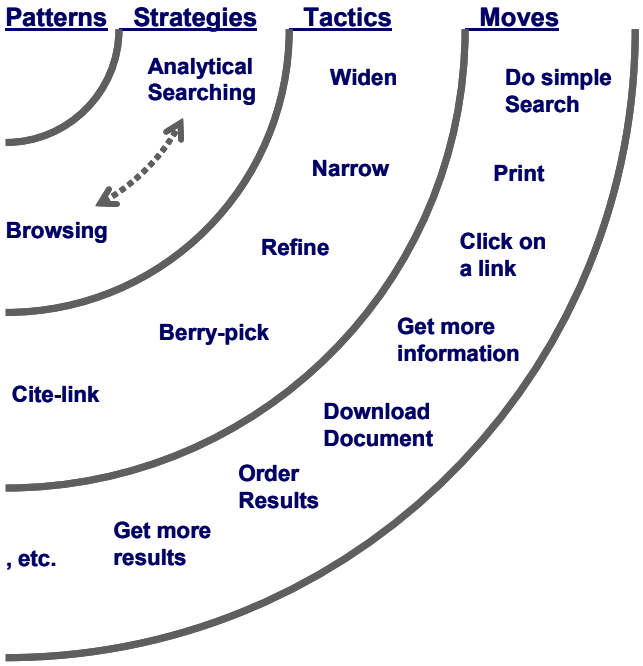

Figure 1: Information behaviour; grouped by level of granularity

3. RESEARCH DESIGN AND METHODOLOGY

3.1 Research aim

The aim of this investigation is to detect whether or not there are patterns of information seeking behaviour among the architectural, engineering and construction (AEC) industry professionals. The study concentrates on the users' interaction with online industry information resources.

3.2 Background

In the information seeking field, researchers typically focus on the information seeking process, resources individuals or groups of individuals use when seeking information to resolve a problem or when seeking information serendipitously, and/or outcomes of the information seeking process (Sonnenwald & Wildemuth, 2001; Wilson, 2000; Jin & Fine, 1996). It is proven that information seeking could involve indeterminate sequences of events. These events are influenced by the user's behaviour in each of its levels, i.e. strategies, tactics, and moves.

Methods are typically based on creating a controlled test environment where user samples are involved with some test procedures. Data collection methods vary from psychometric measures, surveys, interviews; think-aloud protocols, to direct observations (Sonnenwald & Iivonen, 1999; Sonnenwald & Wildemuth, 2001; Fabritius, 1998; Chen & Cooper, 2001; Downs et al, 1988; and Höök, 2000). These methods are most effective in stimulating evidences of the cognitive aspects of the information seeking process, which could include pre-search and post-search activities. However, a number of problems have been stated in these studies. They are (1) sampling mechanisms include demographic considerations formed in subjective manners; (2) Users behave differently in controlled environments; (3) User feedback is hindered by the user's perception of their own seeking behaviours and existing technologies; (4) Results are not reliable due to small samples.

On the other hand, direct observation methods based on collecting data about the user interactions while engaged in an information seeking session try to correct these problems (e.g. Cothey, 2002; Chen & Cooper, 2001). The main problems that face these types of research methods are; (1) it is often difficult to gain access to user log data; (2) it can only be interpreted subjectively since it does not provide insight into participants' perspectives or cognitive reasoning (Sonnenwald & Wildemuth, 2001).

3.3 Research design

The research design is a large-scale cohort based case study. It involves an empirical enquiry that investigates a contemporary phenomenon within its real-life context, which conforms to the definition of Yin (1994, p. 13). It uses Web access transaction log records of a live web based information access system, Construction Information Services (CIS). Anonymous recoding of user identification ensured their privacy. However, server identification of unique user authentication allows individual information seeking sessions to be compiled. Two datasets have been collected over two different periods with the sum of 104,224 sessions over the first half of the year 2001 and 86,678 sessions over January and February of 2002. Clustering analysis is performed in order to find whether there are natural groupings of usage patterns on the online system.

4. DATA COLLECTION AND PREPARATION

4.1 Transaction log records

The data used for this study are web transaction log records of the server activities on the Construction Information Services (CIS) online system. Log data typically consists of electronic records of the user's requests and the system's response. Figure 2 shows an example transaction log record for a user searching the system with the phrase "fire alarm".

```
Date: 2001-01-26
Time: 00:57:13
Method: GET
Uri Query (State):
/scripts/Web.cis?MfcISAPICommand=SearchAnywhere&AuthCode=1454689&anywhere
=fire+alarm&SortBy=rank&lLastRec=10&PreviousNewOrRevised=
Client system: Mozilla/4.7+[en]+(Win98;+I)
Referer:
http://www.tionestop.com/scripts/CIS.CIS?Welcome&AuthCode=1454689
```

Figure 2: An example web transaction log record on CIS online system: example search for "Fire Alarm"

Each log record contains a number of parameters describing the user's state, i.e. which page he/she has accessed and the page he/she is coming from including all the parameters describing the server request. Hence each line of the transaction logs represents a particular user move, e.g. searching, moving from a results' page to another or to download a document.

For example, from the record illustrated in Figure 2, besides the date, time and user's machine configuration one can learn the following information about this particular user move: this move could be the second move on a search session as the user has come from the welcome page (notice the "Referer" parameter). It also indicates that the user has used the "Search Anywhere" feature in the welcome page using the phrase "fire alarm" as a search term. Hence the user arrived to this state by submitting that server call, i.e. "Uri Query". Other information can also be obtained, such as number of results returned (10) and "relevance rank" as the default results sort order. Users IP addresses and Usernames were removed from the logs as the first pass on preparing the data for analysis. This is due to privacy policy considerations. One last parameter that is important to emphasise here is the "AuthCode", (bold faced in Figure 2). This denotes the unique authentication code that identify the user's session. The server automatically generates this code once the user logs on and gets authenticated.

4.2 Identifying the system's states

In order to investigate the user moves on the CIS online information system, all possible user actions have been mapped into 21 states. Due to the trade-off between resolution of states and computational cost (Chen and Cooper, 2002), these states are grouped into eight higher level ones, which are used as the user states in this study. Each transaction record contains the full url of the server request which includes commands such as 'SearchAnywhere'(see Figure 2). These commands were used as the manifestations of these user states. Figure 3 shows these states, on level 2, and their relationship with the gathering user state groups, on level 1.

Table 1 lists these eight user states and the transaction record manifestations used to identify them. Each user state has been letter coded. Thus the user move in the previous example (Figure 2) could simply be represented as: "A|B" which means: a user move from "Start Session" to "Simple Search".

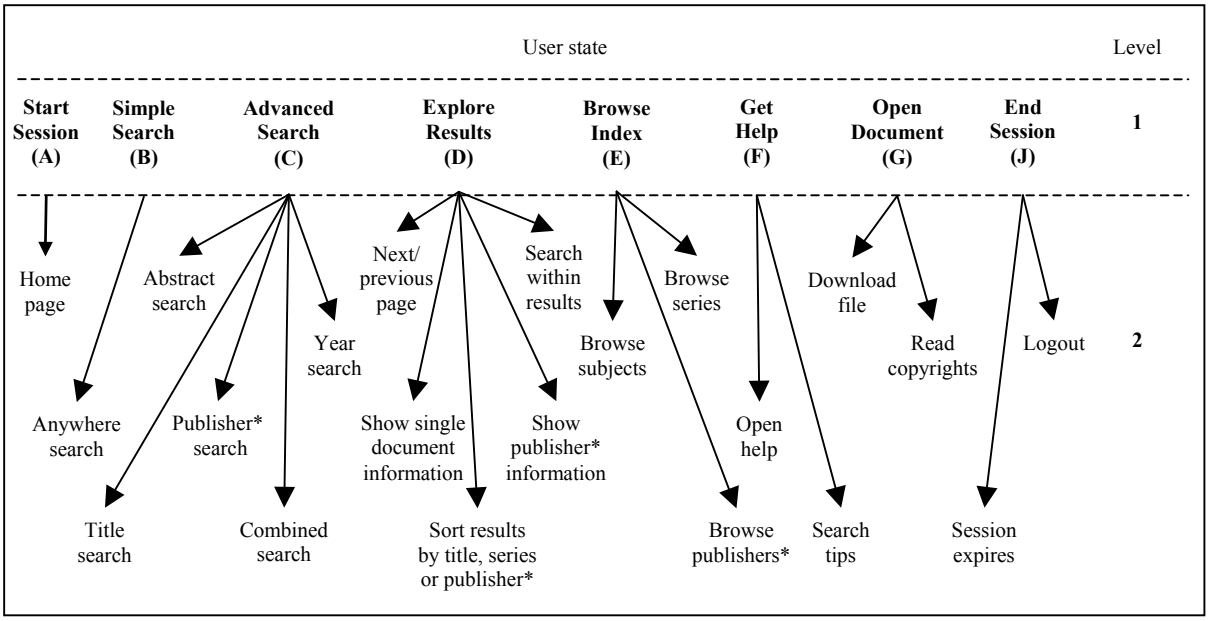

Figure 3: The two levels of abstractions of user moves on the Construction Information Service (CIS) online catalogue system.

** Publishers can be treated as authors on systems that provide technical literature.*

Table 1 user states during a session

User moves	Code	Manifestations in Web transaction logs
Start session	A	Being a subscription based service; users have to logon to their organisation's account in order to have access to the system. However, the login pages could not be used to mark the start of the sessions due to the users privacy statements. Instead, the system's home page is used.
Simple search	B	Searching indiscriminately across all bibliographic data fields. Search anywhere pages are used.
Advanced search	C	A user looking for specific documents. In that case, the user employs information retrieval (IR) techniques. This includes doing title search, abstract search, publisher search, combined search and year search. In the context of technical literature information resources, publishers act as authors of documents, e.g. British Standards. Hence publisher searching can be compared with author searching in normal electronic information system.
Explore results	D	Sifting through search results to locate the required documents or examine the search tactic. Includes: show single document information, show a publisher's information, sort results, search the results and forward to next or previous page of results.
Browse index	E	Navigate the pre-classified indexes. Includes: browse the whole document collection by subject categories, publisher names, or document series, e.g. legislations or code of practice.

User moves	Code	Manifestations in Web transaction logs
Get help	F	Using the system's help feature. This includes: accessing the main help or help tips in the search pages.
Open document	G	This user move assumes that the user has found a document of interest. It includes the user accessing the download pages and/or reading the publishers' copyright pages. Copyright pages are required by some publishers, e.g. Chartered Institution of Building Services Engineers.
End session	J	End of user's session. Two methods are adopted to determine the end of session. First, monitor users who manually logout from the system by accessing the logout page. Secondly, measure the time a user stays idle and not requesting any page from the server. Automatically the system logs users after 20 minutes. In that case End session is calculated on the last user activity.

4.3 Identifying user sessions as unit of analysis

In order to identify unique session characteristics of the study's data samples, a three stage transformation method has been implemented.

The first stage was to identify sessions and extract them from the logs. This was done using the unique authentication codes. Together with the time stamp on each record, the analysis program compiled complete episodes of user's seeking moves during a single session. Hence, raw log records have been transformed into sequences of user moves. Figure 4 shows an example session episode for the same "Authentication code" used in

```
1454689,2001-02-14 12:59:11,A|B
1454689,2001-02-14 12:59:20,B|D
1454689,2001-02-14 12:59:41,B|G
1454689,2001-02-14 13:08:50,B|G
1454689,2001-02-14 13:12:34,B|D
1454689,2001-02-14 13:12:38,D|G
1454689,2001-02-14 13:16:10,A|B
1454689,2001-02-14 13:16:17,A|B
1454689,2001-02-14 13:16:39,A|E
1454689,2001-02-14 13:17:25,A|B
1454689,2001-02-14 13:17:38,A|B
```

Figure 4: An example session episode

The second stage of session Data preparation was to transform these session episodes into vectors by considering each move as a vector attribute. The count of each move became the value of there attributes. Total session lengths in seconds were also calculated and included as a vector attribute.

The third and last stage of data preparation was to remove duplicate sessions and disqualify unusable sessions across the whole data sample. The resulted sessions are referred to in the paper as 'unique sessions'. The criteria for selecting these unique sessions were:

1. Identical values of all session vectors including user moves and session lengths in seconds.
2. Sessions that lasted less than 60 seconds were excluded from the sample. This is based on measuring several cases of using the system to do one search and one download, which lasted on average more than one minute.
3. The frequencies of unique sessions (CS) were included in the session vectors for future frequency measurements.
4. The following user moves have been excluded from the session attributes because they had constant value of zero across all sessions: A|D, A|G, B|C, B|E, E|J, F|B, F|C, F|D, F|E, F|G, F|J, G|A, G|C, G|D, G|E and G|G.

In order to obtain comparable values for user moves, all moves values were converted into ratio of each individual move to the total count of moves during a session. For example, taking the example in Figure 4, the value of A|B was converted from 5 to 45%. Figure 5 illustrates an example of the session vectors in their final format.

ID	CS	TS	TM	A|A	A|B	A|C	A|E	A|F	A|H	A|J	B|A	etc.		
1454689	1	1107	11	0	45	0	9	0	0	0	0	…		
1454940	1	94	5	0	20	0	0	0	0	0	0	…		
1454A72	1	150	6	33	0	33	0	0	0	0	0	…		
1454F72	4	52	3	0	33	0	0	0	0	0	0	…		

ID: Unique session identification
CS: Count of covered identical sessions
TS: Session length in seconds
TM: Total count of user moves in the session.
A|A, A|B, etc.: Percentages of user moves within the session

Figure 5: Parts of example session vectors.

4.4 Characteristics of the data set

The aim of this investigation is to ascertain whether there are patterns of usage while seeking information among the specified user-discipline, i.e. AEC professionals. For that purpose, two sets of transaction log records have been collected. The first dataset was intended for the main analysis while the second for validation. The first set consisted of 104,224 sessions collected from 15th January to 29th June 2001. The second set was collected from 7th January to 28th February 2002. It includes 86,678 sessions. For the purpose of ease of discussion in this paper, the first dataset is called 2001 data while the second is called 2002 data (Table ***2***).

Table 2: Summary of study's datasets

	2001	**2002**
Sum of Moves	889,848	713,740
Total sessions	104,224	86,678
Unique sessions	65,611	60,460

The time difference between the two sets and the change of the covered period were intentionally selected in order to validate the resulted analysis. Because, the study's users are mostly professionals, their behaviour is not expected to change according to certain time of the year, unlike studies that investigate mostly users from academia (Chen and Cooper, 2001).

5. DESCRIPTIVE ANALYSIS

The means of the session variables in both datasets, 2001 and 2002, have been collected in order to investigate the diversity of behaviour among all users. High dispersion values have been noted, evident in the high values of Standard Deviation and Variance parameters (Nachmias & Nachmias, 1996) (Table 3). These results indicate the lack of a general usage behaviour among AEC users. However, it did not disprove that there could be several clusters of usage patterns, where each would have its central tendency. The next section describes the search for clusters of behaviour types within the data.

Table 3: Frequency table of session variables in 2001 and 2002

	2001			2002		
Variable	Mean	Std. Deviation	Variance	Mean	Std. Deviation	Variance
TS	424.68	574.87	330473.35	453.49	640.66	410448.20
A|A	3.22	8.05	64.83	2.88	8.10	65.55
A|B	16.45	16.03	256.92	19.36	17.76	315.40

A|C	4.43	9.74	94.86	3.84	9.54	90.94
A|E	2.68	7.54	56.85	2.12	7.12	50.70
A|F	0.14	1.79	3.20	0.09	1.53	2.35
A|J	0.03	0.20	0.04	0.02	0.24	0.06
B|A	1.78	5.46	29.79	1.60	5.52	30.49
B|B	6.60	13.53	183.13	7.43	14.81	219.31
B|D	9.48	12.23	149.61	9.70	12.50	156.37
B|F	0.15	1.70	2.90	0.11	1.46	2.14
B|G	3.35	10.36	107.28	4.55	12.33	152.05
B|H	5.72	10.93	119.49	6.53	11.97	143.34
B|J	0.07	0.28	0.08	0.05	0.23	0.05
C|A	0.46	2.45	6.00	0.28	1.90	3.61
C|B	0.00	0.03	0.00	0.00	0.00	0.00
C|C	2.63	8.59	73.83	2.00	7.54	56.87
C|D	2.17	6.00	36.04	1.62	5.31	28.18
C|E	0.01	0.18	0.03	0.00	0.12	0.01
C|F	0.07	0.94	0.88	0.04	0.67	0.45
C|G	0.54	3.68	13.51	0.56	3.82	14.56
C|H	1.19	4.77	22.77	1.06	4.56	20.83
C|J	0.03	0.20	0.04	0.02	0.14	0.02
D|A	0.85	3.14	9.86	0.67	2.85	8.14
D|B	0.72	2.93	8.57	0.71	2.96	8.77
D|C	0.66	3.16	9.98	0.56	3.11	9.69
D|D	11.67	19.33	373.67	11.17	19.32	373.30
D|E	0.41	2.41	5.79	0.33	2.23	4.99
D|F	0.04	0.61	0.38	0.02	0.54	0.29
D|G	2.67	7.31	53.40	2.84	7.73	59.69
D|H	3.31	6.75	45.59	3.05	6.78	45.92
D|J	0.12	0.59	0.35	0.09	0.44	0.20
E|A	0.25	1.87	3.51	0.15	1.48	2.19
E|B	0.33	2.53	6.38	0.30	2.53	6.40
E|C	0.06	0.86	0.74	0.04	0.76	0.57
E|D	1.32	4.57	20.93	0.98	4.00	16.02
E|E	2.25	6.73	45.34	1.79	6.35	40.33
E|F	0.01	0.30	0.09	0.00	0.22	0.05
E|G	0.33	2.99	8.92	0.29	2.80	7.86
E|H	0.22	2.10	4.41	0.15	1.68	2.81
F|A	0.01	0.49	0.24	0.01	0.33	0.11
F|F	0.13	1.95	3.79	0.06	1.23	1.51
F|J	0.00	0.07	0.01	0.00	0.05	0.00
G|B	0.00	0.07	0.00	0.00	0.00	0.00
G|F	0.04	0.82	0.67	0.02	0.67	0.45
G|H	0.00	0.00	0.00	0.00	0.08	0.01
G|J	0.00	0.08	0.01	0.00	0.08	0.01

The characteristics of the two datasets have been compared in order to verify their data contents. Figure 6 shows the distribution of the percentage of user moves over the sum of moves across all sessions. A matching pattern is found between the two datasets. While values themselves could have less significant meaning for the purpose of exploring usage patterns, this pattern match suggests the validity and comparability of the two sets (Yin, 1994). This figure also provides the following information:

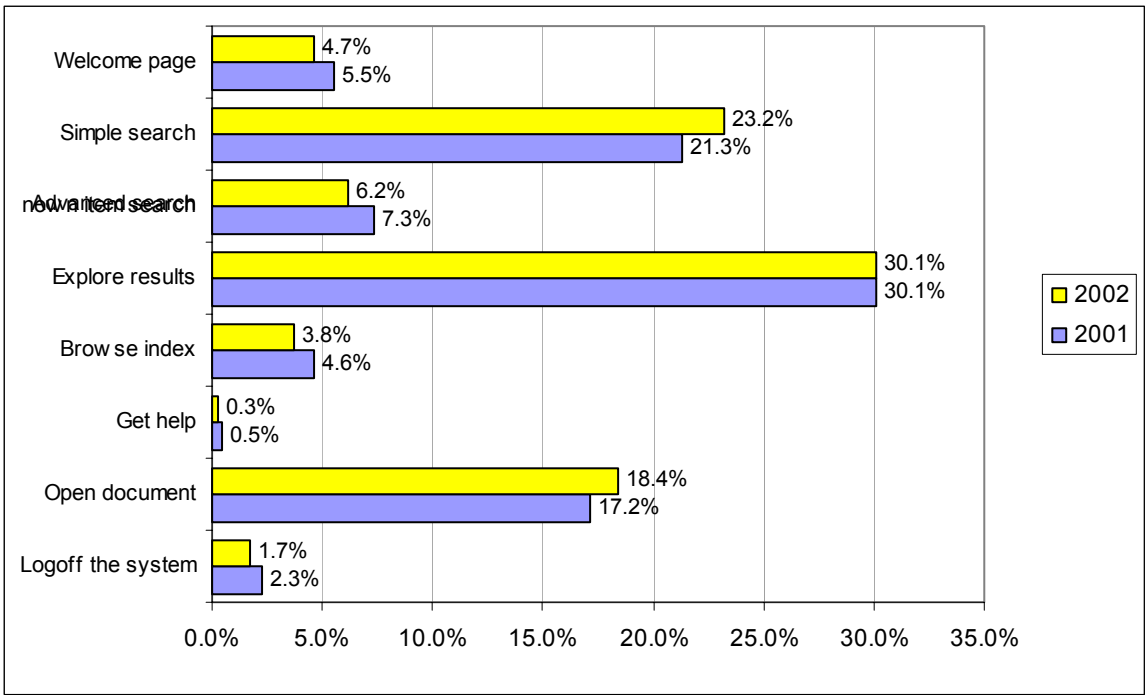

Figure 6: Percentages of user moves over the sum of moves across the whole dataset

- It clearly shows the popularity of the 'exploring results' activities among all users, almost 30% of all moves in both 2001 and 2002.
- It also shows a strong bias toward 'simple searching', 23.2% in 2002 and 21.3% in 2001, compared to 'advanced searching', which are 6.2% and 7.3% in 2001 and 2002 respectively.
- The popularity of the 'simple searching' activity is increasing from 2001 (21.3%) to 2002 (23.2%). This increase correlates with a decrease of the number of moves in both 'advanced searching', from 7.3% in 2001 to 6.2% in 2002, and 'browsing the library indexes', from 4.6% in 2001 to 3.8% in 2002.

Simple searching activity has been monitored to measure query lengths. Longer queries would indicate more complex searching and high searching skill levels. Results in Figure *7* show the frequency of the count of terms used in simple search queries in 2001 and 2002. It clearly shows that there is a broad range in the number of terms that get used. However, 77% and 76% of the queries in both datasets are one or two terms long.

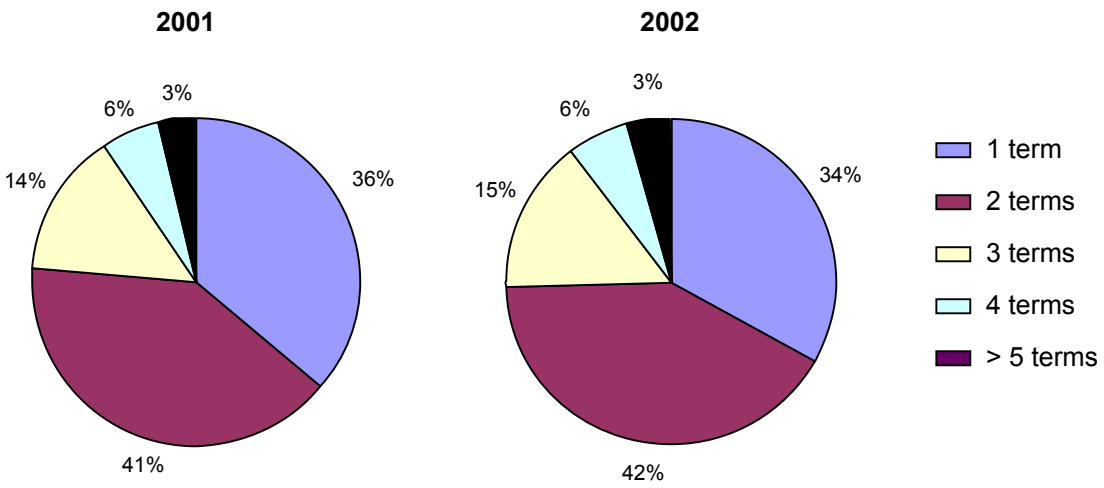

Figure 7: Frequency of query lengths in simple searching activities

6. CLUSTER ANALYSIS

The study utilises cluster analysis as a statistical method for finding the natural grouping of session vectors. The aim of this exercise is to find out whether there are clusters of common information seeking patterns among the study user sample and how many they are. The study employs a three phases clustering process, which concurs with a similar framework introduced by Chen and Cooper (2001). In the first phase, the 2001 session vectors are transformed into clusters using a 2 stages methodology. In the second phase, the 2002 session vectors are transformed into clusters using the same methodology. In the third phase, the 2002 session vectors are classified into the clusters derived from the 2001 clustering in phase 1. The purpose from phase 2 and 3 is to validate that the resultant clusters from both 2001 and 2002 datasets are not formed by chance.

6.1 Phase 1 – clustering 2001 sessions

The main goal of this phase is to locate the natural grouping of sessions. This can be achieved by doing a hierarchical cluster analysis procedure. However, it works best with datasets that contain a small number (less than a few hundred) of cases to be clustered, (Aldenderfer & Blashfield, 1984). Hence a 2 stages methodology has been implemented.

In the first stage, a non-hierarchical method adopting K-means clustering analysis procedure has been implemented. This method is useful to cluster large number of cases where the number of clusters is pre-defined. It has been used to reduce the number of cases from 65,611 sessions to 100 clusters. Fifteen clusters with five or fewer cases each have been treated as outliers and removed from further analysis. The remaining 85 clusters are used as the input for the second stage.

In the second stage, a hierarchical clustering procedure is implemented. The clustering procedure adopted a "Squared Euclidean Distance" measurement between pairs of furthest objects when calculating the distance between the clusters. This approach provides more statistically distinct clusters, (Aldenderfer & Blashfield, 1984). SPSS 11.5 has been used to run this cluster analysis.

Table 4 shows the agglomeration schedule of the last 20 stages in the final hierarchical solution. It provides a numerical summary of the clustering stages. For example, in stage 71, the cluster formed in stage 1 was joined by the one formed in stage 39 as the distance they had the minimum distances from each other. They both joined the cluster that was formed in stage 66. In turn the formed cluster in this stage, 71, has been used later in stage 77 where it was joined by the ones from stages 1 and 24.

Table 4: The agglomeration schedule of the 2001 hierarchical solution

Possible number of clusters	Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
		Cluster 1	Cluster 2		Cluster 1	Cluster 2	

20	65	1	5	99.09	63	64	66
19	66	1	47	108.15	65	0	71
18	67	51	77	112.65	0	0	81
17	68	2	25	119.24	59	0	74
16	69	65	79	124.21	55	0	73
15	70	3	84	139.24	56	0	74
14	71	1	39	151.89	66	0	77
13	72	9	66	170.51	0	0	80
12	73	12	65	170.59	32	69	83
11	74	2	3	172.51	68	70	78
10	75	6	8	173.78	0	62	76
9	76	6	59	197.30	75	0	79
8	77	1	24	201.98	71	0	80
7	78	2	41	204.28	74	54	79
6	79	2	6	228.32	78	76	81
5	80	1	9	273.03	77	72	82
4	**81**	**2**	**51**	**279.95**	**79**	**67**	**82**
3	82	1	2	364.00	80	81	83
2	83	1	12	421.50	82	73	84
1	84	1	49	482.39	83	43	0

A good cluster solution sees a sudden jump (gap) in the distance coefficient. The solution before the gap indicates the good solution, i.e. the optimum number of clusters. In this solution (Table 4) there are two possible jumps in the distances coefficient, i.e. between stages 79 and 80 and between stages 81 and 82 with jumps of 44.71 and 84.05 respectively. The maximum gap indicates the optimal number of clusters. Therefore the 4 cluster solution has been selected (shown in bold face in Table 4).

6.2 Phase 2 – clustering 2002 sessions

The second phase in the clustering process was to implement the same procedures of phase 1 on the 2002 dataset. The same two stage methodology was performed. In the first stage, the 60,460 sessions were transformed into 100 non-hierarchical clusters. 22 clusters, with five or less cases each, were removed from further analysis. The remaining 78 clusters were used as the input for the second stage where the hierarchical cluster analysis has been performed.

Table 5: The agglomeration schedule of the 2002 hierarchical solution

Possible number of clusters	Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
		Cluster 1	Cluster 2		Cluster 1	Cluster 2	
20	58	1	6	82.61	53	56	60
19	59	8	68	88.74	50	0	61
18	60	1	11	93.09	58	46	63
17	61	8	34	109.78	59	57	69
16	62	23	69	124.07	36	0	76
15	63	1	19	124.16	60	0	67
14	64	5	22	124.70	55	0	74
13	65	7	76	128.69	0	0	72
12	66	4	29	130.38	54	0	70
11	67	1	55	142.84	63	0	68
10	68	1	10	161.42	67	0	69
9	69	1	8	182.17	68	61	73
8	70	4	42	188.70	66	0	71
7	71	4	25	204.93	70	0	72

6	72	4	7	214.46	71	65	75
5	73	1	67	215.90	69	0	74
4	**74**	**1**	**5**	**227.13**	**73**	**64**	**75**
3	75	1	4	296.63	74	72	76
2	76	1	23	465.06	75	62	77
1	77	1	3	695.80	76	0	0

Table 5 shows the last 20 stages in the solution. There are three possible cluster numbers: two with a jump of 230.74 between stages 76 and 77, three with a jump of 168.43 and four with relevant jump of 69.50 between stages 74 and 75. Opting for, too few groups could jeopardize sufficient description of the differences in usage behaviour; four clusters have been considered as the optimal groupings.

6.3 Phase 3 – validating the clustering results

The main goal of the third and last phase of the clustering analysis is to validate the four clusters of the 2001 dataset. The objective is to prove that these clusters did not occur as a random phenomenon and are replicable from the 2002 dataset.

A non-hierarchical clustering analysis adopting the K-means method was performed to cluster the 2002 dataset into 4 clusters. This time, the 2001 cluster centres resulted from phase 1 were used as the initial cluster centres for this phase. In other words, the 2002 sessions have been categorised into the four clusters of the 2001 dataset. By the end of phase 3 clustering, each session in the 2002 dataset has been assigned to two different clusters, i.e. one from the phase 2 clustering and the other from phase 3 which is a 2001 cluster.

In order to validate the two clustering results, crosstabulation analysis has been performed. The crosstabulation analysis is a basic technique for examining the relationship between two categorical (nominal or ordinal) variables. It offers tests of independence and measures of association and agreement for nominal and ordinal data, nominal in our case.

Table 6: The symmetric measures between phase 2 and phase 3 clusters

		Value	Approx. Sig.
Nominal by Nominal	Phi	0.841	0.00
	Cramer's V	0.488	0.00
	Contingency Coefficient	0.642	0.00
N of Valid Cases		78	

Table 6 shows the symmetric measures between the two clustering systems. In the case of nominal variables, i.e. the four cluster names, the three most indicative measures are: Phi, Cramer's V and Contingency Coefficient (Aldenderfer & Blashfield, 1984). The three measures were proved significant (Sig. value < 0.01). This indicates that the relationship between the two clustering systems is statistically significant. Moreover, the positive values of the three measures with higher than 0.30 indicates strong relationship.

These results indicate that the clustering is unlikely to have been discovered by chance and demonstrate the validity of the clustering analysis of user sessions on the Construction Information Services (CIS) online system. 1 user sessions.

Table **7** shows the validated four clusters of the 2001 user sessions.

Table 7: The four clusters of AEC user seeking patterns, based on 2001 dataset

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4
TS	2548.84	959.68	324.69	130.93
A|A	0.87	1.61	11.30	2.07
A|B	6.30	16.59	5.22	4.40
A|C	1.70	2.98	17.08	1.60
A|E	1.46	1.67	3.26	22.90

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4
A|F	0.03	0.10	0.29	0.14
A|H	0.00	0.00	0.00	0.00
A|J	0.03	0.04	0.07	0.03
B|A	0.62	1.48	2.69	1.20
B|B	4.81	8.62	1.74	1.31
B|D	6.70	9.47	2.00	1.86
B|F	0.05	0.13	0.09	0.11
B|G	1.76	4.99	0.45	0.63
B|H	3.31	7.63	0.44	1.89
B|J	0.04	0.09	0.02	0.03
C|A	0.23	0.46	1.35	0.19
C|B	0.00	0.00	0.00	0.00
C|C	2.89	2.33	9.97	0.39
C|D	1.64	1.77	7.71	0.63
C|E	0.00	0.02	0.00	0.01
C|F	0.06	0.05	0.25	0.02
C|G	0.35	0.40	1.75	0.22
C|H	0.80	0.95	4.52	0.46
C|J	0.03	0.03	0.11	0.01
D|A	0.84	0.78	0.90	0.82
D|B	1.26	0.80	0.54	0.55
D|C	0.56	0.62	0.93	0.62
D|D	35.94	10.56	5.00	3.84
D|E	0.25	0.31	0.26	0.34
D|F	0.04	0.03	0.04	0.01
D|G	5.80	2.93	1.74	2.60
D|H	5.71	3.75	2.53	2.15
D|J	0.49	0.12	0.10	0.11
E|A	0.13	0.20	0.44	2.00
E|B	0.15	0.23	0.42	3.49
E|C	0.06	0.08	0.11	0.20
E|D	1.65	0.99	1.56	9.43
E|E	2.10	1.62	2.89	20.05
E|F	0.01	0.00	0.02	0.08
E|G	0.49	0.32	0.41	2.10
E|H	0.21	0.17	0.25	1.78
F|A	0.00	0.01	0.02	0.02
F|F	0.05	0.13	0.20	0.09
G|B	0.00	0.00	0.00	0.00
G|F	0.01	0.04	0.02	0.02
G|H	0.00	0.00	0.00	0.00
G|J	0.01	0.01	0.00	0.00

7. GROUPS OF INFORMATION SEEKING PATTERNS AMONG AEC USERS

Four clusters of information seeking behaviour have resulted from the clustering exercise. As discussed earlier, user moves were used to represent the user sessions as the most discrete form of the system usage manifestations. Table 8 shows description of the resulted clusters using variables derived from the above mentioned user moves. Using these variables enables the reader to build his/her own interpretations of the differences between clusters. However, the authors will introduce an interpretation later in this section.

The variables are grouped into four groups according to their nature. (1) Session characteristics – includes the total number of sessions and their mean of. (2) User activities – includes sums of user moves accessing each part

of the CIS online system. (3) Download decision – includes variables that indicate what triggered the user to make the download decision. (4) Measures – includes three session performance measures, i.e. the ratio of successful information seeking activities, ratio of system difficulty and ratio of seeking activities per download.

Table 8: Description of the four behaviour groups

		Clusters				
Var.	**Description**	**1**	**2**	**3**	**4**	**Definition**
	(1) Session characteristics					
Ns	Number of sessions	9277	37822	6326	1728	
Np	Percentage of number of sessions	16.82	37.44	10.01	3.04	
Sm	Mean of Session length in minutes	43	16	5	2	
	(2) User activities					
B	Simple search	12.52	26.23	7.92	9.75	A|B+B|B+C|B+D|B+E|B+G|B
C	advanced search	5.21	6.01	28.09	2.80	A|C+C|C+D|C+E|C
D	Explore results	45.93	22.79	16.28	15.75	B|D+C|D+D|D+E|D
E	Browse index	3.81	3.63	6.42	43.30	A|E+C|E+D|E+E|E
F	Help	0.23	0.49	0.91	0.48	A|F+B|F+C|F+D|F+E|F+F|F+G|F
Gn	Open non-copyrighted document	8.40	8.63	4.36	5.55	B|G+C|G+D|G+E|G
H	Read document copyright	10.02	12.50	7.74	6.28	A|H+B|H+C|H+D|H+E|H+G|H
Gc	Open copyrighted documents	9.59	12.02	7.28	5.87	H|G
G	Open document	17.99	20.65	11.64	11.42	Gn+Gc
Ah	Users hitting home buttons	2.76	4.73	16.85	6.38	A|A+B|A+C|A+D|A+E|A+F|A
	(3) Download decision					
Gs	Open document after searching	6.21	13.96	7.16	3.21	B|H+B|G+C|G+C|H
Gb	Open document after browsing	0.70	0.49	0.66	3.88	E|G+E|H
Ge	Open document after exploring results	11.52	6.68	4.27	4.75	D|G+D|H
	(4) Measures					
Rs	Ratio of successful seeking (100x)	83.52	57.57	27.43	20.45	G/(B+C+E)x100
Rf	Ratio of system difficulty (100x)	0.35	0.83	1.54	0.67	F/(B+C+D+E)*100
Rd	Ratio of seeking activities per downloads	2.36	0.77	0.46	0.19	Sm/G

Cluster 1, (Table 8), contains the longest session lengths, with mean of 43 minutes (Sm). Users spend most of that time performing explore results behaviour, consuming around 45% of session lengths (D). Compared to advanced searching and browsing indexes, users most prefer accessing the system by doing simple searches with a percentage of 12.52% (B) as opposed to 5.21% and 3.81% for the other activities respectively (C & E). Users of that group appear to be less confused about their information seeking strategies as they rarely go back to the home page with the least percentage of 2.76% among other clusters (Ah). Users tend to explore the search results extensively before making a download decision (Ge). Sessions falling in that group have the highest successful seeking rates (Rs) and the lowest system difficulty measures (Rf). However, users tend to take more time to find the required information, which is evident from the ratio of seeking activities to download (Rd). In summary, this group of users constitute around 17% of the total number of session (Np). They undertake the lengthiest sessions. Based on their behaviour, one can infer a common trend of exploratory information seeking with highly interactive system usage.

Cluster 2 has the largest number of sessions of 37.44% (Np). The average session length is around 16 minutes (Sm), of which 26% in simple searching activity. Lesser emphasis on exploring results activity compared to Cluster 1 users (D). This could be due to the short session lengths. Users of this group tend to make the download decision on documents found on the first page of the search results (Gs). This could simply mean that they find the right document on the first results page for reasons like: high system performance or trust of the search engine (McKechnie, 1999). However, taking into consideration that both the document collection and the

system are consistent between the four clusters and the increase of the help request (F), the lack of exploring results before downloading a document is better explained by the users' lack of information skills in terms of manipulating the system interface. This assumption is smaller measure of successful searching (Rs). In summary, Cluster 2 represents the majority of user sessions (Ns & Np). Users in that cluster tend to utilise the minimum system features. They mostly perform simple searching with number of downloads almost equal to the number of searches (B & G), which indicates successful searching in most of the times. Based on these finding one could infer that this group are mostly knowledgeable users, who tend to use subject searches derived from their domain experience and they are less skilled in terms of information system usage.

Cluster 3 is largely different from the previous two in that its users mostly know what documents they are looking for. High percentage of advanced searching, e.g. title or publisher searching (C). Session lengths are much shorter, of around 5 minute (Sm). The number of downloaded documents is less than half the searches (G & C), which is explained by either unsuccessful searching or the documents do not exist in the collection. However, the high percentage of resetting the system to start a new search by going back to the home page (Ah) supports the former suggestion. This group of users performs the most help-intensive sessions (F & Rf). In summary, they could be described as fast-paced users who perform known-item searching accompanied by help-intensive behaviour.

Finally, cluster 4 has the shortest session lengths of average 2 minutes (Sm). It is also distinctive in terms the highest percentage of browsing index activity (E), the lowest successful seeking rate (Rs) and the fewest number of sessions with 3% (Np). This behaviour suggests two types of users. The first is a user who is coming back to the system to download a previously downloaded document which he/she knows its location within the indexes. The second type is user that is quickly navigating the system in order to check whether a document, or most probably a group of documents, is included in the system. In summary, the results suggest that users of this group are largely passive users who tend to use browsing as the main information seeking strategy; they mostly experience unsuccessful seeking episodes, which they hardly spend time on.

8. CONCLUSIONS

The goal of this investigation was to find out whether there are common patterns of information seeking among the AEC industry professionals. Transaction log records of users' interaction of the construction information service (CIS) online information system have been collected. Two sets of data were gathered over the first six months of 2001 and the first two months of 2002. Both datasets were prepared for analysis by transforming them into a list of unique session vectors. Each session vector comprised of 47 variables that represent the user moves on the seeking session.

Studying the characteristics of the two datasets showed the popularity of the 'exploring results' and 'simple searching' activities among all users. It also showed a discrepancy between simple searching and advanced searching in favour of the former activity. Results also show that around 76% of search queries used over the whole period of study are of a maximum of two terms long.

A three phase clustering analysis methodology was performed in order to outline natural groupings of seeking patterns. In the first phase, a two step procedure of non-hierarchical and hierarchical clustering resulted in identifying four optimal clusters of user behaviours. In phase 2 and phase 3 these four clusters have been validated and proved statistically significant.

Together with the statistical characteristics of each usage group, i.e. cluster, a subjective interpretation was discussed. The four groups could be described as 1) exploratory information seeking behaviour with highly interactive system usage. 2) Knowledgeable users who tend to employ their domain experience in doing subject searches with inexperienced system usage. 3) Fast-paced users who perform known-item searching accompanied by help-intensive behaviour. 4) Passive users who experience unsuccessful short seeking episodes.

There is clear evidence that the industry users lack information searching skills, which is clear from their short query searches and rare advanced searching. This finding highlights the importance of devising new mechanisms that provide the industry professional with the means to recognize and improve their information skills. It also emphasises the need for enhancing systems with better user interfaces.

Although there are time constraints imposed on construction industry users, especially when asked to perform an information seeking task, they tend to seek information in an exploratory manner. The more time they spend on

the system the more exploratory behaviour they perform. It is clearly evident that exploring results returned from the system is the dominant information seeking activity.

Enhancing the results exploration facilities within a system, e.g. information visualisation techniques, relevance feedback and finding similar documents features should reduce the time needed to find the required documents and in finding relevant information more easily.

This study focused on identifying information seeking behaviour of the AEC industry professionals on specialised online information resources. Demographics within the user sample have been excluded for the purpose of finding the natural grouping of users' usage patterns. However, future research is needed to examine the distribution of the usage groups, introduced in this study, on the diverse user cultures within the AEC industry.

REFERENCES

Aldenderfer, M. S., and Blashfield, R. K. (1984). Cluster Analysis, Newbury Park: Sage Publications

Amor, R., Langham, M., Fortmann, J. and Bloomfield, D. (1996). "The UK industry knowledge base feasibility study", CIB W78 and TG10 workshop: Construction on the Information Highway.

Augenbroe, G. (1998). Building product Information Technology, Executive White Paper, Construction Research Centre, Georgia Institute of Technology.

Barbour (1999). The Sourcing and Exchange of Information: Across the Building Team, Barbour Index plc.

Barbour (2000). Influencing Product Decisions: Specification and Beyond, Barbour Index plc.

Belkin, N., P. Marchetti and Cool, C. (1993). "BRAQUE: Design of an interface to support user interaction in information retrieval", Information Processing and Management 29(3): 325-344.

Björk, C. and Z. Turk (2000). "A Survey of the Impact of the Internet on Scientific Publishing in Construction IT and Construction Management", Electronic Journal of Information Technology in Construction (ITcon) 5: 73-86.

Chen, H.-M., and Cooper, M.D. (2001). "Using Clustering Techniques to Detect Usage patterns in a Web-Based Information System", Journal of the American Society for Information Science and Technology, 52(11): 888-904.

Chen, H.-M., and Cooper, M.D. (2002). "Stochastic Modeling of Usage Patterns in a Web-Based Information System", Journal of the American Society for Information Science and Technology, 53(7): 536-548.

Cothey, V. (2002). "A Longitudinal Study of World Wide Web Users' Information Searching Behaviour", Journal of the American Society for Information Science and Technology 53(2): 67-78.

Crawford, M., C. Cann, J. and O'Leary, R. (1997). Uniclass: Unified Classification for the Construction Industry, Royal Institute of British Architects.

Downs, E., Clare, P. and Coe, I. (1988). Structured Systems Analysis and Design Method: Application and Context, London, Prentice-Hall.

Eastmen, C. (1999). Building Product Models: Computer Environments Supporting Design and Construction, CRC Press LLC.

Ekholm, A., Häggström, L. and Karlsson, H. (2000). BSAB 96, the Swedish Construction Industry Classification System, Swedish Building Centre.

Fabritius, H. (1999). "Triangulation as a Multiperspective Strategy in a Qualitative Study of Information Seeking Behaviour", In T. Wilson and D. Allen, Exploring the Contexts of Information Behaviour. London, Taylor Graham: 406-419.

Froese, T. (1996). "Models of Construction Process Information", Journal of Computing in Civil Engineering, American Society of Civil Engineers 10(3): 183-193.

Hearst, M. A. (1999). "User Interfaces and Visualization", In Modern Information Retrieval. Ricardo Baeza - Yates and Berthier Ribeiro - Neto. New York, ACM press: 257-323.

Höök, K. (2000). “Steps To Take Before Intelligent User Interfaces Become Real”, Journal of Interaction with Computers 12(4): 409-426.

Jin, Z. and Fine, S. (1996). “The Effect of Human Behaviour on the Design of an Information Retrieval System Interface”, The International Information and Library Review 28(3): 249-260.

Kelly, J., Aouad, G., Rezqui, Y. and Crofts, J. (1997). “Information Systems Developments in the UK Construction Industry”, Automation in Construction 6: 17-22.

Lockley, L., Watson, R., and Shaaban, S. (2002). “Managing E-Commerce in Construction - Revolution or E-Business as Usual?”, Engineering Construction and Architectural Management, 9(3): 232-240.

Marchionini, G. (1995). Information Seeking in Electronic Environments, Cambridge university press.

McKechnie, J. (1999). “Towards Intelligent Information Retrieval for the Construction Industry”, 5th International Conference on the Application of Artificial Intelligence to Civil and Structural Engineering. Oxford.

McKechnie, J., Shaaban, S. and Lockley, L. (2001). “Computer Assisted Processing of Large Unstructured Document Sets: A Case Study in the Construction Industry”, 1st ACM Symposium on Document Engineering, Atlanta.

Nachmias, C., and Nachmias, D. (1996). Research Methods in Social Science. (5th edition), Arnold

RIBA-Enterprises (2002). The NOP Information Sources Survey 2002, RIBA Enterprises.

Sonnenwald, D. and Iivonen, M. (1999). “An Integrated Human Information Behaviour Research Framework for Information Studies”, Library and Information Science Research 21(4): 429-457.

Sonnenwald, D. and Wildemuth, B. (2001). “Investigating Information Seeking Behavior Using the Concept of Information Horizons”, SILS Technical Report 2001-01. 22 Manuscript pages. Winner of the ALISE 2001 Methodology Paper award.

Tolman, F. (1999). “Product Modelling Standards for the Building and Construction Industry: Past, Present and Future”, Automation in Construction 8: 227-235.

Wilson, T. (2000). “Human Information Behaviour”, Informing Science 3(2): 45-55.

Yin, R. (1994). Case Study Research, Design and Methods, (2nd edition), Newbury Park, Sage Publications.