

# Towards a predictive model of information seeking: empirical studies of end-user searching

M. Ennis, A.G. Sutcliffe and S.J. Watkinson,  
Centre for HCI Design, City University, UK  
M.Ennis@city.ac.uk

## INTRODUCTION

Previous empirical studies of searcher behaviour have drawn attention to a wide variety of factors that affect performance; for instance, the display of retrieved results can alter search strategies (Allen 1991, 1994), the information need type influences search behaviour, (Elkerton et al 1984, Marchionini 1995); while the task complexity, reflected in the information need can affect user's search behaviour (Large et al 1994). Furthermore, information source selection (Bassilli 1977), and the user's model of the system and domain impact on the search process (Michel 1994); while motivation (Solomon 1993, Jacobsen et al 1992) and the importance of the information need (Wendt 1969) also influence search duration and the effort a user will employ.

Rouse and Rouse (1984) in a review of empirical studies, summarise a wide variety of variables that can effect searching behaviour, including payoff, costs of searching, resource available, amount of information sought, characteristics of the data and conflicts between documents. It appears that user behaviour is inconsistent in the search strategies adopted even for the same search need and system (Davidson 1977, Iivonen 1995).

Theories of searcher behaviour have been proposed that provide explanations of aspects of end-user behaviour, such as the evolution of the user's information need and the problems of articulating a query, [Bates (1979, 1989), Markey and Atherton 1978], effective search strategies in browsing and goal directed searches [Marchionini 1995, Belkin (1987, 1993)], the linguistic problem of matching search terms with indexing terms or content of target documents through an expert intermediary (Ingwersen 1982) or cognitive aspects of IR (Kuhlthau 1984, Ingwersen 1996). These models offer insight into the complexities of retrieval activity but can not predict all user behaviour for the

## TOWARDS A PREDICTIVE MODEL

variety of strategies, tactics, tasks, users and systems designs which occur in the retrieval task. More detailed theories of information searching behaviour are required as a sound basis for improving the functionality and usability of information retrieval systems designs by associating elements of the design with user activity.

The empirical study reported in this paper was motivated by the development of a cognitive model of searcher behaviour (Ennis and Sutcliffe 1996, Sutcliffe and Ennis 1998). The ultimate motivation of our work is to extend our understanding of the search process in an attempt to define the facilities that should be provided by configurable information retrieval systems.

This paper describes experimental studies of end-user performance and search activities for information retrieval sessions using the WinSPIRS interface for MEDLINE<sup>1</sup>. The paper is organised as follows: section two describes the experimental design and analytic method employed. This is followed in section three by performance results. This section is concluded with analysis of query construction, strategies and the systems facilities used. This study examines the cognitive activities performed in the retrieval process. Users' information retrieval sessions are studied through examinations of the verbal reports of activities performed by searchers.

## EXPERIMENTAL METHOD

### Experimental design

Seventeen medical students (thirteen males and four females, aged between 24 and 26) who were three months away from their clinical final examinations took part in the experiments. The subjects' experience in the domain and information retrieval system was assessed by a pre-test questionnaire. The answers given were used to categorise subjects' profile of domain and device knowledge and to allocate individuals into either a group who had some experience in the use of MEDLINE (experts) or novices. Three pilot subjects undertook the experiment before the trials were started, to refine the medical scenarios for the information required and to test the experimental procedure.

All subjects were given the following search tasks which were developed by a independent medical expert.

1. Please use the MEDLINE database to investigate the socio-economic reasons for increased failure rates on the oral contraceptive pill (*Experimental task code OG1*),
2. Please utilise the MEDLINE database to compare caesarean sections being performed in planned and emergency situations from the standpoints of:

<sup>1</sup> MEDLINE is the copyright of the U.S. National Library of Medicine and SilverPlatter

- maternal and foetal safety, infection control, statistical and economic data (*Experimental task code OG2*).
- Using the MEDLINE database please assess the importance of blood sugar levels and lipid profile in the cause of myocardial infarction within the male population (*Experimental task code PH1*).
  - Utilise the database to determine some areas of increased and decreased efficiency within the NHS since the introduction of formal clinical auditing (*Experimental task code PH2*).

The version of the MEDLINE system used for the experiments runs under MS Windows 3.11 running on a personal computer using the WinSPIRS interface, see Figure 1 for the query interface.

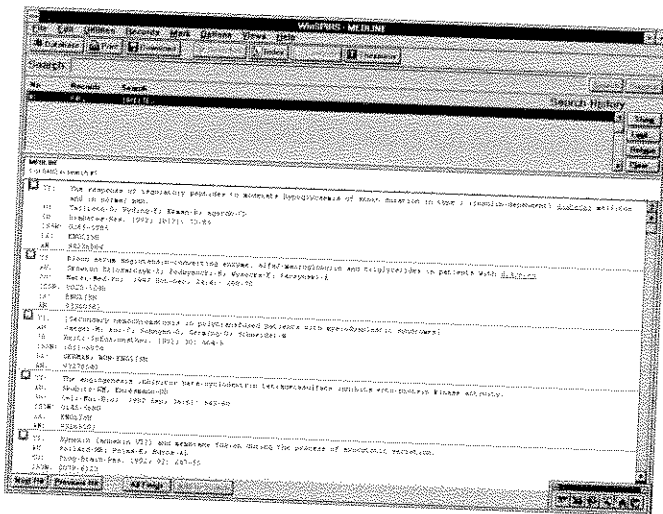


Figure 1. WinSPIRS search interface for MEDLINE showing a summary of retrieved results.

When questioned after the experiment, the subjects confirmed that the tasks were similar in style and complexity to ones they would normally use MEDLINE to solve. Subjects had access to the tasks throughout the trial and were instructed that they could record information as they searched. Each task was given to the subject after they had reported that the preceding task had been completed satisfactorily. The subjects' notes, physical actions, search strategies, views of the problem to be tackled and search history were recorded. The presentation of tasks was randomised to counteract any learning effects. At the end of the experiment the subjects were asked to rate the perceived difficulty of each of the tasks on a five point scale.

To ensure a high and uniform level of device knowledge for the expert group

and to accentuate the differences between the groups, the ten subjects with experience using MEDLINE were given training in the use of the information retrieval system. The system facilities and brief guidance on the production of strategies were explained in a tutorial document and a demonstration. Novices were not trained. Subjects were requested to think aloud during the experiment and their verbal protocols, with associated physical actions, were recorded on video and audio tape. The subjects were advised to take their time when verbalising and not be afraid of verbalising too much, in line with the practices of Ericsson and Simon (1984).

To train the subjects to produce verbal protocols they were required to think aloud while they performed a simple eight puzzle. The puzzle task was used to encourage familiarity with giving verbal protocols whilst operating a computer. The experimenter encouraged the subjects to verbalise at action points, indicating why they had performed an action. The interaction in the puzzle was designed to include time lags so that prompts to verbalise from the experimenter caused minimal interruption to the subjects behaviour.

### Data analysis

*Performance:* performance measures were calculated for each task:

- **Recall:** the proportion of documents marked as relevant by the subject at any time during the session as a percentage of relevant records in the database as contained in a gold standard solution set.
- **Precision:** the proportion of the documents marked as relevant which were also judged to be relevant by an independent expert as a percentage of the records retrieved and marked as relevant by the subject. Note that precision was calculated from the documents the subject marked as relevant, not the overall number retrieved during querying.

Recall performance scores were calculated as a percentage of an ideal 'gold-standard' solution synthesised from all the subjects' searches and queries created as a collaboration between independent domain experts and an information retrieval specialist. Independent medical experts were asked to assess the articles for relevance to the task scenarios. Differences between experts' document classification were discussed and an agreed cluster of relevant documents was produced. This was performed so as to eliminate irrelevant articles from the recall data as some subjects selected all the results of a particular query without assessing the articles for relevance, based on the size of the results set and the query posed.

*Systems facilities used:* It is hypothesised that the use of certain systems facilities may assist users in obtaining better performance. The analysis investigates which were the most popular strategies and which seemed to aid

retrieval success.

*Strategies used:* It is hypothesised that better performers would use more effective search strategies, as reported in the literature (Harter et al 1985, Marchionini 1995). The subjects search histories and timing data were categorised to ascertain the strategy types employed. This analysis tries to identify patterns in the strategies used.

**Mental behaviour categorisation.** Protocol transcripts were analysed by matching mental behaviours to speech segments in accordance with Ericsson and Simon's (1984) method. Thirteen categories of mental behaviour were used in the analysis. The data was entered into an Access database along with user characteristics. This allowed the rapid comparison and identification of patterns of mental behaviour sequences based on the user and task characteristics present using SQL queries. The following categories were used.

Identify Concept	Articulation of new terms/ keywords. ( present in the scenario or the device) e.g., ' so this wants myocardial infarction .....
Identify concept relationship	Reports of concepts/terms with appropriate linking phrases. e.g., '... and within that...' or '... as a subset of that ...'
Formulate strategy	Verbalisations of planning within the search, and the indication of a pre-search map of activities to be performed. This occurs if a subjects' verbal protocols indicate they are trying to find a method of solving the problem. e.g., ' how are we going to .....
Refine strategy	Amend or extend the existing plan of action for their current strategy. This category is dependent on the state of the current strategy changing. e.g., 'lets extend this by narrowing it down...' or 'lets attack this from a different perspective ...'
Explore synonyms for concepts, terms and their inter-relationships	Searchers report looking for lexical variations and synonyms for terms/ concepts already in the search problem space. Reasoning about possible terms to be used and how these should be inter-related. e.g., ' so that's oral contraceptive...or ... OCP.. or possibly contraceptive pill ...' the relationships are bounded by the scope of the concept being explored. Explicitly looking for variations on terms.
Reason about the problem space	Subject verbalises the search option and the reasons why they have chosen one over another. e.g., 'I could always use NHS but that will not help me as it can mean many different things so I'll leave that for now.....'
Select terms	Selection of terms for the current search. This only occurs as an activity if the searcher chooses a term from a previous

Formulate Query	verbalisation of possible search terms. The subject translates their information need into a query. This activity is observed when the subject produces a new search which doesn't incorporate aspects of the previous search, but does refer explicitly to composition of query syntax. e.g., 1 inceas* Form Query 2 decreas* Form Query
Revise Query	The subject produces a search which is an amendment to a previous query. This phase is the composition of previous queries or editing the existing query. e.g., 1 inceas* Form Query 2 decreas* Form Query 3 #1 AND #2 Revise Query
Execute actions to implement	Interaction with the computer to implement the search.
Evaluation of initial systems results	Evaluation of search results presented by the system is inferred. This activity has occurred if a subjects verbal protocol indicate that they are evaluating their search based on the context in which the verbal report is found. e.g., ' so lets have a look at what this has found ...' or ' so that's 222 articles ...' This activity occurs also when a searcher removes dialogue boxes associated with a null search.
Evaluate the content of the results	Inferred from users observed actions and verbal protocols that indicate articles have been read to ascertain their importance to the information need. This may be based on the time spent reviewing an article; the display the subject has selected for the articles retrieved; moving through the results set using scrolling operators and by the subject paraphrasing the article's content. e.g., ' ...so this ones concerning ....'
Evaluate and diagnose search success/ failure	Inferred from verbalisations as subjects review the results/ articles retrieved. These reports may indicate that the results expected didn't occur and the search should progress accordingly. e.g., ' ... just checking to see if I've got enough information to ...' or ' ...this isn't quite what I want ... it doesn't ...' The verbalisations tend to precede a decision to reformulate or give up.

**Protocol categorisation rules.** During the search process a searcher may perform many retrievals before inspecting the results set; furthermore, they may return to previous results sets if they don't find the information they require. If

this happened it was categorised as 'execute actions' to alter the results set followed by subsequent evaluation.

## RESULTS

### Search performance.

Recall and precision results of the subjects for the four tasks was poor (average recall 13.94%) when compared to the gold standard. The gold standard queries represented a single search without any evaluation and gave a average recall of 76.66% of all the relevant articles possible, demonstrating that a high percentage of the relevant articles could be accessed if the need was articulated correctly. Expert searchers had, on average, better recall than novices, but some novices had good recall scores. Significant differences were found between the groups for recall in only one of the four tasks in which recall success was also correlated to evaluation time (PH1  $p < 0.05$ , Mann Whitney U) when novices outperformed experts. There is no immediately apparent reason for the performance difference on this task; but on this task alone novices spent longer on a retrieval session than experts and time spent evaluating articles was much higher in novices (74%) than experts (57%). In PH1 recall was positively correlated (Spearman's rank) to the users perception of task difficulty ( $p < 0.005$ ), total time spent ( $p < 0.05$ ) and time spent evaluating ( $p < 0.05$ ). In PH1 it appears that searchers who were more diligent in the amount of time spent in the retrieval process, and evaluating the results reaped rewards in terms of improved recall. Correlations for the other tasks were not significant. The experts had significantly similar ranking of the tasks for recall ( $p < 0.01$  Friedman test) in the order of PH2, PH1, OG2, OG1 ( $p < 0.001$  L Page test) and novices didn't.

### Query construction

On average expert searchers used more term alternatives (9.1 terms) than novice users (6.6 terms) even though they have equivalent domain knowledge. There was a significant difference in the number of different terms used by each of the groups in the two more complex tasks (experts > novices for OG2  $p < 0.05$ ; PH2  $p < 0.05$  Mann-Whitney U). However, the total number of terms used by each group was not significantly different for all tasks, although individuals did show a common ranking in the number of terms used per task (PH2, PH1, OG1, OG2,  $P < 0.001$  for both groups). This indicates individual searcher's reactions to each of the tasks were similar and they adjusted the number of terms needed to articulate the search based on the task components.

The proportion of the terms used from the gold standard query was low (expert average 22.81%, novices 15.85%) and significant inter-group differences

## TOWARDS A PREDICTIVE MODEL

were found on the public health tasks (experts > novices PH1  $p < 0.05$ , PH2  $p < 0.05$ , Mann Whitney U). However, the terms used by most subjects that were shared with the gold standard constituted a high proportion of the total terms they used (> 70%) apart from H7, H10 (67%) and L6 (60%). Experts expanded more of the components of the original need statement in queries than novices by adding terms from the systems thesaurus or from their knowledge of the domain (experts 60% of task elements explored, novices 37% of task elements expanded with synonyms or refined with sub-terms). On average experts produced 43% more queries (average 12.73 per task) than novices (average of 8.93 per task). The number of query iterations performed and the terms used that were shared with the gold standard were positively correlated (OG1  $p < 0.05$ , OG2  $p < 0.01$ , PH1  $p < 0.01$ , PH2  $p < 0.01$  Spearman rank order correlation). Significant correlations were also found between the number of terms searchers added to queries and the number of query iterations (OG1  $p < 0.001$ , OG2  $p < 0.05$ , PH1  $p < 0.001$ , PH2  $p < 0.05$ ) so prolonged querying also appears to lead to richer and more complete queries.

Experts used the term exploration facilities sparsely, whereas only one novice used the facilities at all. Eighty percent of experts used the thesaurus or term suggestion facilities, but use of these facilities was inconsistent. None of the subjects used the index facilities and only one subject consulted the help system. Both groups of subjects (14 out of the 17 subjects) re-used queries at some point in the experimental session; however, expert searchers were more consistent in this (9 out of 10 subjects on all tasks). Experts used term extensions to increase the scope of queries, but novices did not. The expert searchers who used term extension facilities did so consistently irrespective of the task.

Experts used Boolean and query structuring operators to sub-divide complex queries into functional components, whereas the novices constructed only simple queries; and this pattern was consistent across tasks. Experts using ordering mechanisms (use of () operators to group query components) do so consistently across tasks (see Table 1) while novices consistently only used Boolean 'AND' relationships, whereas experts use more complex Boolean syntax. The query relationship use of the novices confirms the findings of Sewell *et al.* (1986) and Marchionini (1989) in which the 'AND' relationship was used most commonly and users submitted multiple parallel searches instead of using 'OR' relationships. This may provide some of the reasons for the consistent use of successive term substitution strategies by novice subjects (Table 2).

	Task support facilities				
	Thesaurus	Term Suggestions	Query Re-use	Term exploration using operators **?	Order of execution
H1		μ			
H2	μ		μ μ μ μ	μ μ μ μ	
H3	μ μ μ	μ μ μ	μ μ μ μ	μ	μ
H4	μ μ		μ μ μ μ		μ μ μ μ μ
H5			μ μ μ μ	μ	
H6			μ μ μ μ	μ μ	
H7		μ μ μ μ		μ μ μ	
H8	μ μ μ μ		μ μ μ μ	μ μ μ μ	μ μ μ μ
H9	μ	μ	μ μ μ μ	μ μ μ	
H10		μ	μ μ μ μ	μ μ μ μ	μ μ μ μ
Total	13	10	36	26	16

	Task support facility				
	Thesaurus	Term Suggestions	Query Re-use	Term exploration using operators **?	Order of execution nesting operators ()
L1					
L2			μ μ μ		
L3			μ		
L4			μ μ μ μ	μ	
L5	μ μ				
L6			μ μ μ μ		
L7			μ μ μ μ		
Weighted Total	2.9	0	22.9	1.42	0

Table 1: Analysis of systems facilities used by searchers. The spatial distribution reflects the task order OG1, left hand side, OG2, PH1 middle locations, PH2, right hand side. The totals for novices are weighted to account for differences in the number of subjects in the novice group

**Information searching strategies**

Query logs were analysed to highlight patterns in the need articulation process. Analysis concentrated on cycles of narrowing, broadening and successive term substitution as illustrated in table 2. A subject was determined to be using successive term substitution if the queries used showed a pattern whereby a term relating to a particular concept is successively amended as subsequent queries are submitted. A subject was determined to be using narrowing cycles if in the query history a pattern of progressive search narrowing by using more specific terms, constraints were tightened, new concepts were added using Boolean

'AND' clauses or Boolean relationships between query elements were altered indicating narrowing (e.g., substituting 'AND' for 'NEAR'). A subject was determined to be using broadening cycles if in the query history a pattern of pro-

	Strategy types				
	Narrowing cycle	Broadening cycle	Successive term substitutions	Evaluation strategy	Trial and error
H1	μ μ μ μ	μ μ μ μ	μ μ μ	μ	μ μ μ μ
H2	μ μ μ μ		μ μ μ	μ	μ μ μ
H3	μ μ μ		μ μ	μ μ μ	μ μ μ μ
H4	μ μ		μ μ μ μ	μ μ μ μ	μ μ μ μ
H5	μ μ		μ μ μ μ	μ μ μ μ	μ
H6	μ μ μ	μ μ		μ μ μ	
H7	μ μ μ μ	μ μ μ μ	μ μ μ		μ
H8	μ μ μ μ	μ μ μ μ	μ μ μ	μ	μ
H9	μ μ μ μ	μ μ	μ	μ μ μ	
H10	μ μ	μ μ	μ		μ μ μ
Total	32	18	23	19	20

	Strategy types				
	Narrowing cycle	Broadening cycle	Successive term substitutions	Evaluation strategy	Trial and error
L1	μ		μ μ μ	μ μ μ	μ μ μ
L2			μ μ μ	μ μ μ	μ μ μ
L3	μ		μ μ μ	μ μ μ	μ μ μ
L4			μ μ μ μ	μ μ μ μ	μ μ μ
L5	μ		μ μ μ μ	μ μ μ μ	μ μ μ
L6	μ μ μ	μ	μ μ μ	μ μ μ	μ μ μ
L7	μ μ	μ μ	μ μ μ	μ μ μ	μ μ μ μ
Weighted Total	11.42	4.3	32.4	31.4	30

Table 2: Strategy types used by searchers. The spatial distribution reflects the task order OG1, left hand side, OG2, PH1 middle locations, PH2, right hand side. The totals for novices are weighted to account for differences in the number of subjects in the novice group

gressive query expansion occurred if more general terms were used, constraints were relaxed, new concepts were added using Boolean OR clauses, or Boolean relationships between query elements were altered indicating broadening (e.g., substituting 'NEAR' for 'AND'). At a higher level assessments were made as to whether a searcher followed patterns of trial and error or as favouring evaluation. A subject was determined to be using trial and error if the queries showed little systematic development. A subject was determined to be favouring evaluation strategies if they spent 25% or greater of their total retrieval time on one evaluation session.. For example, if a subject was to submit myocardial

infarction and males as a query and in the next query submit blood glucose and lipids then they would be determined to be using trial and error (or sub-goaling) as there is no direct link between the queries.

In task OG1 eight expert subjects adopted a query narrowing strategy; however, these subjects were not the best performers in this group. In OG2 and the PH tasks most experts used a narrowing strategy. The pattern is less consistent in task PH2, with only five subjects (H1, 2, 7, 8 & 9) following the same strategy, again without much reward in terms of recall. Overall, about half of the expert subjects adopted a consistent search strategy across tasks, the more consistent subjects being H1, 7 and 8. In spite of adopting this apparently effective strategy these subjects did not achieve better recall, so it appears that although there may be an 'expert' behaviour pattern, unfortunately, it is not always successful.

Most subjects followed more than one strategy. Expert searchers concentrated on cycles of narrowing and broadening whereas novices favoured a trial and error approach by substituting the individual query terms. However, there were individual differences, for example, experts H2-5 did not use query broadening. Novices made use of evaluation strategies more than experts. It is noticeable all the novices used evaluation strategies on the public health tasks (PH1, PH2) and all but one of the subjects used this approach for task OG1. However this was not observed in task OG2, possibly because this task promoted a richer query representation.

The novice subjects followed a more consistent pattern, however, they showed few facets of what may be considered to be expert behaviour. In task OG1, complex queries are noticeably absent although iterative querying was adopted by three subjects, two of whom followed a narrowing strategy and scored well (L6, 7). Both groups substituted terms in queries, although novices favoured this strategy slightly more than experts. Sub-goaling to break a complex need into sub parts was practised by both groups.

#### IMPACT OF USER DEVICE EXPERIENCE AND TASK ON THE MENTAL BEHAVIOURS PRODUCED

A frequency distribution for the different behaviours was created from the category analysis of verbal protocols. The effects of task and users device knowledge on the behaviour observed are investigated. The analysis considers three measures of activity: the average frequency and the average rate of activity and the activities performed in a specific activity category as a of the proportion of total activities performed.

#### Results

Table 3 highlights some interesting differences in the searchers activity patterns. Expert searchers perform on average a third more activity transition than novices. The differences between experts and novices occur in the following categories: strategy revision, problem space reasoning, exploration of synonyms, new query formulation, query revision and implementation actions for all of which experts perform more activity than novices. It is also interesting that the evaluation of the content of articles and search diagnosis activity do not vary with device knowledge. Experts perform different levels of activity on the four tasks. The majority of differences between the four tasks are restricted to the execution of the device and the results interpretation (i.e., identify concept, form query, revise query, evaluate initial results and evaluate the content of the results). The activities produced by novices show less variation for querying based activities than experts, with the exception of task PH2 where subjects appeared to give up.

Category	OG1		OG2		PH1		PH2	
	E	N	E	N	E	N	E	N
Identify concept	3.22	2.57	7.67	4.14	4.22	3.43	3.33	2.71
Identify relation	1.00	1.00	2.56	2.14	1.56	1.57	0.89	1.00
Form strategy	1.00	0.86	1.00	1.00	1.00	1.00	1.00	1.00
Revise strategy	4.22	1.86	4.89	2.43	4.00	2.43	3.56	1.71
Reason about problem space	3.11	0.86	3.89	0.71	3.44	0.86	3.22	0.86
Explore synonyms	5.44	2.57	6.11	1.43	4.67	2.14	3.67	0.86
Select terms	2.33	2.57	3.00	1.43	2.67	2.28	2.00	1.00
Form query	5.11	2.86	6.67	4.43	4.00	3.29	4.00	2.43
Revise query	8.00	6.00	10.78	7.70	6.89	6.00	5.67	2.86
Execute implementation actions	15.55	9.71	19.78	13.00	12.78	10.00	13.00	5.71
Evaluate initial results	12.56	8.86	17.22	11.85	11.44	9.14	9.44	5.43
Evaluate results content	5.11	4.71	7.44	7.14	3.67	3.86	4.67	2.71
Evaluate and diagnose search success/failure	2.11	2.71	3.89	2.85	2.78	2.71	3.33	1.57
Total activities	68.76	47.14	94.90	60.25	63.12	48.71	57.78	29.85

Table 3: Average occurrence of activity types for device novices (N) and experts (E)

The tentative observations on the raw data were investigated for statistical significance (see table 4). The analysis enables the testing for significant differences between novices and experts on each of the experimental tasks using unrelated t tests enabling the data to be divided based on experimental task, user knowledge (novice and expert) and behaviour category.

Category	Tasks							
	OG1		OG2		PH1		PH2	
	t	SD	t	SD	t	SD	t	SD
Identify concept	2.10	p < 0.05	3.60	0.005	2.57	0.025	1.25	
Identify relation	0		0.52		0.03		0.37	
Form strategy								
Revise strategy	2.12	p < 0.05	2.39	0.025	1.46		2.02	0.05
Reason about problem space	2.06	p < 0.05	4.02	0.005	3.00	0.005	2.91	0.01
Explore synonyms	1.13		2.87	0.01	2.68	0.01	2.19	0.025
Select terms	0.22		1.65		0.49		1.66	
Form query	1.18		0.84		0.39		1.31	
Revise query	0.80		1.20		0.43		1.60	
Execute implementation actions	1.06		1.32		0.80		1.72	
Evaluate initial results	0.49		1.16		0.72		1.69	
Evaluate results content	0.20		0.24		0.16		1.71	
Evaluate and diagnose search success/ failure	0.87		1.10		0.08		2.69	0.01

Table 4: Significant differences between the categories of activities for device novice/expert df = 14

The results in table 4 show that significant differences occur with experts reasoning more than novices (revise strategy, reason about the problem space). The data indicates significant differences between expert and novices in the concepts identified and the exploration of synonyms.

The percentage of the total activities by category is used in an attempt to identify strategy differences between experts and novices. Any difference in the configuration of activities observed would indicate a strategic difference existed either across tasks or user group. Unrelated t tests are used to investigate any strategic differences between the activity performed by searchers with differing device knowledge. This analysis is performed for each of the tasks.

Category	Tasks							
	OG1		OG2		PH1		PH2	
	t	SD	t	SD	t	SD	t	SD
Identify concept	0.40		1.05		0.65		1.88	0.05
Identify relation	0.22		0.55		1.07		1.42	
Form strategy	0.75		3.22	0.005	1.38		1.98	0.05
Revise strategy	2.83	0.01	0.97		1.48		0.29	
Reason about problem space	3.06	0.005	3.00	0.005	3.33	0.005	2.44	0.025
Explore synonyms	2.49	0.025	2.73	0.01	3.20	0.005	1.18	
Select terms	0.54		1.00		0.42		0.72	
Form query	0.81		0.47		0.46		0.05	
Revise query	1.46		1.45		0.49		0.29	
Execute implementation actions	0.80		0.93		0.10		0.80	
Evaluate initial results	0.55		1.56		0.17		1.40	
Evaluate results content	2.86	0.01	2.20	0.025	1.26		0.45	
Evaluate and diagnose search success/ failure	2.47	0.025	0.34		0.96		0.58	

Table 5: Significant differences between the different tasks performed on each of the categories of activities for high and low device knowledge populations (2 d.p) df = 14

Table 5 shows that the differences between the configuration of searchers activities which exist for novice and experts are significant. PH2 also shows differences with novices spending a greater percentage of activities spent identifying concepts than experts. There are expert-novice differences in the % of the total activities performed as a single activity category on the following categories: reason about the problem space, exploration of synonyms and the evaluation of the results contents.

Unrelated t tests are used to investigate activity rate differences between the activity performed by searchers with differing device knowledge. This analysis is performed for each of the tasks.

The significant differences in the rates of activity between novices and experts occur only for revision of the strategy, reasoning about the problem space and exploration of synonyms (table 6). Significant differences in the rate

of activity occurrence are expected between novices and experts because of the proceduralisation of knowledge as skill and thus increased speed of activity completion.

Category	Tasks							
	OG1		OG2		PH1		PH2	
	t	SD	t	SD	t	SD	t	SD
Identify concept	0.91		1.52		1.44		1.01	
Identify relation	0.01		0.05		0.20		0.74	
Form strategy	0.87		1.95	0.05	0.15		1.33	
Revise strategy	3.07	0.005	1.97	0.05	1.77	0.05	0.34	
Reason about problem space	2.30	0.05	3.18	0.005	4.10	0.005	2.58	0.025
Explore synonyms	0.67		2.92	0.01	3.50	0.005	2.07	0.05
Select terms	0.01		1.22		1.07		1.11	
Form query	1.48		0.25		0.62		0.37	
Revise query	0.58		0.03		0.98		0.47	
Execute implementation actions	1.20		0.48		1.22		0.14	
Evaluate initial results	0.90		0.14		1.27		0.53	
Evaluate results content	0.62		1.84	0.05	0.18		0.53	
Evaluate and diagnose search success/ failure	0.05		0.23		0.54		0.07	

Table 6: Shows the significant differences between the different tasks performed on each of the categories of activities for high and low device knowledge populations (2 d.p)  $df = 14$

### SUMMARY

- Overall, searcher performance was poor. This may be attributed to the system's user interface rather than the underlying retrieval mechanism given the performance achieved by the gold standard queries.
- Experts exhibited significantly similar ranking of recall on the four tasks whereas novices did not, so it appears that novice searcher performance was more due to chance. This and the idiosyncratic use of system facilities by the subjects indicates that IR systems should provide specific assistance, or targeted task support, for users.
- Multiple strategies exist and searchers are not consistent in their strategy across task. The strategies used by the searcher change during a retrieval session. Experts generally spend longer completing a retrieval session and consider more

### TOWARDS A PREDICTIVE MODEL

alternative term expressions than novices but this is still only a small proportion of the possible search concepts (expert average 22.75%; novices 15.97% of the gold standard queries).

- Experts expanded more components from the original need statement than novices by adding terms from the systems thesaurus or from their knowledge of the domain (experts 60% explored, novices 37% terms expanded with synonyms or refined with sub terms).
- Experts consider more term alternatives than novices but this is only significant on more complex tasks. However, individuals did show a common ranking in the number of terms used per task. The number of terms used is related to the characteristics of the task for both groups.
- Experts re-used queries more, and used substantially more queries to articulate their need to the system than novices (experts average 12.73 queries; novices average 8.93 queries). Experts concentrate on cycles of narrowing and broadening whereas novices favour trial and error by substitution. Novices favour evaluation while experts use systems facilities to explore alternative terms more than novices. Novices only use Boolean 'ANDs' to express relationships between keywords whereas the Boolean relationships used by experts are more diverse. The number of query iterations performed are positively correlated to the coverage of the gold standard.

### DISCUSSION

Although we did find evidence for behavioural differences between novice and expert searchers, no simple correlations between behaviour and performance were immediately apparent. Instead, our analysis revealed a complex picture. There are many contributing factors to effective performance, such as use of complex queries, many search cycles, narrowing strategies and careful evaluation of retrieved results. However, these factors were not evenly distributed across our subjects. Experts relied more on complex query formulation and iterative cycles of searching whereas novices relied on careful evaluation. However, success was neither guaranteed by any one strategy nor by a combination of all of them. Some of our subjects exhibited expert behaviour with poor results, and conversely some followed poor strategies with good results. Although these exceptions were a minority, we can only explain them in terms of choice of appropriate search terms.

Several explanations for the searchers' difficulties are possible. First, choice of appropriate search terms that matched the document contents in free text search. It is possible the searchers were unaware of alternative terms held by the system thesaurus, considering that only a minority of our subjects consulted it. It seems that education of users about these facilities only goes part of the way to solving the problem as shown by the differences in facility use by the two

subject groups. Secondly, they may have had difficulty using these facilities as demonstrated by three of the searchers who explicitly complained about the usability of the thesaurus (out of the eight who used the thesaurus). Finally, the system thesaurus did not always contain appropriate keywords and spellings, hence leading to failed searches. MEDLINE's thesaurus has been set up with US spellings hence articles with English spellings and keywords are not found by free text retrieval.

The lack of correlation between effective strategies and search success has also been noted in ecological studies of on line searches with intermediaries (Smithson 1994), who also found that poor evaluation of results early in the search cycle caused poor performance. The inter-task performance differences we encountered may be interpreted in terms of how tasks may determine search strategies. Bystrom and Jarvelin (1995) propose a model of search strategy selection for tasks according to the user domain and the knowness of the need. In the complementary theoretical studies (Sutcliffe and Ennis 1998) we have elaborated Bystrom and Jarvelin's approach to propose rules for strategy selection according to properties of the information need and task. However, motivation and perceived difficulty of the search clearly has an important effect that we shall have to account for.

Our subjects were not only inconsistent in their search strategies but also in their choice of search terms. There was little overlap with the gold-standard solutions and the inter-subject consistency was also low. This agrees with Iivonen's (1995) findings, although we have not analysed consistency in detailed terms of lexical agreement.

If computerised retrieval systems aim to take over the expert intermediary role of the librarian then their designers will have to pay more attention to the users task needs when designing the user system dialogue. This requirement is becoming more evident due to the shift to end-user searching. The study has shown that the systems interface, task support facilities and user guidance are failing to assist the searcher in need articulation and their retrieval activity, as even domain experts are unable to retrieve a high proportion of the relevant information in the database. The performance figures highlight the requirement for improved IR interfaces and intelligent systems guidance to assist searchers. The differences between the query articulation activity and search strategies of the subject groups shows a need exists for targeted assistance based on the device knowledge held and the pattern of a user's query history. Differences in the query patterns and terms used further support this requirement for user specific task support. Experts generally considered more term alternatives than novice searchers even though they have equivalent domain knowledge. This may be due to differences in their mental models of system functionality, index structures and document representations. This has design implications for the articulation assistance provided by the system and the advice offered to

searchers. This study demonstrates a need to understand the cognitive activities of searchers as a first step towards supporting and enhancing their retrieval behaviour as current IR designs aren't fully supporting user activity.

#### FUTURE WORK

We will adapt our cognitive model of searching behaviour (Ennis and Sutcliffe 1996, Sutcliffe and Ennis 1998) in the light of these results. A complete and fully scenario tested computational implementation of the model in the COGENT modelling environment (Cooper et al 1996) is currently under development. This will lead to the development of guidelines for the provision of task support facilities and the design of IR interfaces.

#### ACKNOWLEDGEMENTS

We wish to thank Dr K.A. Ennis and the Medical students of UCL for their assistance with this work. Mark Ennis is supported by an EPSRC post graduate studentship under grant number: 95306422.

#### REFERENCES

- ALLEN, B.L. (1991) Cognitive research in information science: implications for design. *Annual Review of Information Science and Technology*, 26, 3-37.
- ALLEN, B.L. (1994) Perceptual speed, learning and information retrieval performance. In: *Proceedings of 17th SIGIR Conference, Dublin*. ACM Press, pp. 71-80.
- BORGMAN, C.L. (1985) The user's mental model of an information retrieval system: an experiment on a prototype on-line catalogue. *International Journal of Man-Machine Studies*, 24, 47-64.
- BATES, M.J. (1979) Information search tactics. *Journal of the American Society for Information Science*, 30, 205-214.
- BATES, M.J. (1989) The design of browsing and berrypicking techniques for the on-line interface. *On-line Review*, 13, 407-424.
- BASSILI, J.N. & REGAN, D.T. (1977) Attributional focus as a determinant of information selection. *Journal of Social Psychology*, 101, 113-121
- BELKIN, N., COOL, C., STEIN, A. & THIEL, U. (1993) Scripts for information seeking strategies. In: *Case-based reasoning and information retrieval - exploring the opportunities for technology sharing: Papers from the AAAI Spring Symposium Series*, Technical Report SS-93-07, 8-17.
- BELKIN, N.J. (1987) Information concepts for information science. *Journal of*

- Documentation*, 34, 55-85.
- BEAULIEU, M. (1997) Experiments on interfaces to support query expansion. *Journal of documentation*, 53, 8-19
- COOPER, R., FOX J., FARRINGDON, J. & SHALLICE, T. (1996) A systematic methodology for cognitive modelling. *Artificial Intelligence*, 85, 3-44.
- DANILOWICZ, C. (1994) Modelling user preferences and needs in Boolean retrieval systems. *Information processing and management*, 30, 363-378.
- DAVIDSON, D. (1977) The effect of individual difference of cognitive style on judgements of document relevance. *Journal of the American Society for Information Science*, 28, 273-284.
- DERVIN, B. (1977) Useful theory for librarianship: communication, not information. *Drexel Library Quarterly*, 13, 16-32
- ELKERTON, J. & WILLIGES, R.C. (1984) Information retrieval strategies in a file search environment. *Human factors*, 26, 171-184
- ENNIS, M & SUTCLIFFE, A.G. (1996) A cognitive model of search behaviour: a first step towards improving interface design for IR. In: *Proceedings of the IR and HCI workshop, British Human Computer Interaction Group*, BCS, September 96.
- ERICSSON, K & SIMON, H. (1984) *Protocol analysis: verbal reports as data*. Cambridge, MA: MIT press
- HARTER, S.P. & ROGERS-PETERS, A. (1985) Heuristics for on-line information retrieval: a typology and preliminary listing. *On-line Review*, 9, 407-424.
- HIOVONEN, M. (1995) Searchers and searchers: differences between the most and least consistent searchers. In: E.A. Fox, P. Ingwersen, & R. Fidel, eds. *SIGIR '95. Proceeding of the 18th annual international ACM-SIGIR conference on research and development in information retrieval*. New York: Association for Computing Machinery, 1995, 149-157.
- INGWERSEN, P. (1982) Search procedures in the library analysed from the cognitive point of view. *Journal of Documentation*, 38, 165-191
- INGWERSEN, P. (1996) Cognitive perspectives of information retrieval. Interaction elements of a cognitive IR theory. *Journal of Documentation*, 52, 3-50.
- JACOBSON, T. & FUSANI, D. (1992) Computer, system and subject knowledge in novice searching of a full text multifile database. *Library and Information Science Research*, 14, 97-106.
- LARGE, A. BESHESHTI, J., BREULEUX, A. & RENAUD, A. (1994) A comparison of information retrieval from print and CD-ROM versions of an encyclopaedia by elementary school students. *Information Processing and Management*, 30, 499-513.
- MARCHIONINI, G. (1989) Information seeking strategies of novices using a

- full text electronic encyclopedia. *Journal of the American Society for Information Science*, 29, 165-176
- MARCHIONINI, G. & LIEBSCHER, P. (1991) Performance in electronic encyclopaedias: implications for adaptive systems. *Proceedings of the 54th Annual meeting of the American Society for Information Science*, 28, 39-48.
- MARCHIONINI, G. (1995) *Information seeking in electronic environments*. Cambridge University Press.
- MARKEY, K. & ATHERTON, P. (1978) ONTAP: on-line training and practice manual for ERIC database. Syracuse, NY: Syracuse University, ERIC Clearing House on Information Research. 1978 (Report no. ED106109).
- MICHEL, D.A. (1994) What is used during cognitive processing in information retrieval and library searching? Eleven sources of search information. *Journal of the American Society for Information Science*, 45, 498-514.
- RASMUSSEN, J. (1983) Skills, rules and knowledge: signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man and Cybernetics*, 13, 257-266.
- ROUSE, W. & ROUSE, S. (1984) Human information seeking and design of informationsystems. *Information Processing and Management*, 20, 129-138.
- SEWELL, W. & TEITELBAUM, S. (1986) Observations of end-user on-line searching behaviour over eleven years. *Journal of the American Society for Information Science*, 37, 234-245.
- SOLOMON, P. (1993) Children's information retrieval behaviour: a case study of an OPAC. *Journal of the American Society for Information Science*, 44, 245-264
- SUTCLIFFE, A.G. & ENNIS, M. (1998) Towards a cognitive theory of information retrieval. *Interacting with Computers*, 10, 321-351.
- WENDT D. (1969) Value of information for decisions. *Journal of Mathematical Psychology*, 6, 430-443