Selecting versus Describing:
A Preliminary Analysis of the Efficacy of Categories in Exploring the Web

E.G. Toms[1a], R.W. Kopak[2], J. Bartlett[1], L. Freund[1]

[1]Faculty of Information Studies, University of Toronto,
Toronto, Ontario, Canada
[2]School of Library, Archival and Information Studies, University of British Columbia,
Vancouver, British Columbia, Canada

[a] Person to whom all correspondence should be sent: toms@fis.utoronto.ca

This paper reports the findings of an exploratory study carried out as part of the Interactive Track at the 10th annual Text Retrieval Conference (TREC). Forty-eight, non-expert participants each completed four Web search tasks from among four specified topic areas: shopping, medicine, travel, and research. Participants were given a choice of initiating the search with a search statement (query) or selection of a category from a pre-defined list. Participants were also asked to phrase a selected number of their search statements in the form of a complete statement or question. Results showed that there was little effect of the task domain on the search outcome. Exceptions to this were the problematic nature of the Shopping tasks, and the preference for query over category when the search task was general, i.e. when the semantics of the task did not map directly onto one of the available categories. Participants also evidenced a reluctance/inability to phrase search statements in the form of a complete statement or question. When keywords were used, they were short, averaging around two terms per query statement.

Introduction

We are working toward improved search interfaces and a holistic approach to how interfaces might be designed to facilitate information searching, browsing and encountering. As a first step, we are observing how non-experts seek information on the World Wide Web (the 'Web') noting in particular their mode of interaction with the system.

In this exploratory study we compared how participants used pre-defined categories versus standard search statements. In addition, we examined participant behaviour across three additional factors: the way a search was entered (as question or as keyword), by the source of the task (researcher-specified versus user-personalized), and by task domain (medicine, travel, shopping, and research). We compared the outcomes among these factors using a series of efficiency, effectiveness, and satisfaction metrics. We added verbal protocol data so that we could better understand the reasoning behind participants use of category and string search, and to provide a rich description of strategies used and rationales for observed patterns. In this version of

our analysis, we have included an analysis of quantitative data only.

Method

Participants

Our criteria for selection specified that participants be adult members of the general public (including but not limited to the university community) who have used the Web, who may have searched the Web previously, and who may have had some training, but who had not taken professional search courses. Information science/studies students were eligible only if they were in first term, and had not yet taken a professional search course. The population was one of convenience. Participants were recruited by printed posters posted on bulletin boards on campus, or in libraries and coffee shops in the surrounding area, and via e-mail posted on listservs or e-notice boards at the Universities of Toronto and British Columbia.

The 48 participants (29 women and 19 men) ranged in age from 18-20 to over 65 years; 80% were under 35. Most had university level education, mainly at the bachelor (38%) or masters (30%) level, predominantly from the humanities or social sciences. About half were students; the remainder were from a diverse range of occupations. Most (94%) of the participants had been using the Internet for more than two years, frequently using it for 6 or more hours (50%) per week. Email was the most frequently used application, with all but one person using it daily. All but one participant reported searching the web on a daily or weekly basis. Almost all had no search training of any sort. Overall, they were a relatively young, educated group who were experienced in terms of web use.

Search Interface

We used Google as our Web search engine, and modified the standard Google interface to include both the search box/button, and the Google top level category list (directory). The resulting screen retained Google's simplicity, but added an instruction to either enter a query in the search box or select a category from the directory (See Figure 1). Beyond this initial page, the standard Google interface screens were retained.

Choice of Google as the search engine was based on its current status as the most popular search engine (http://www.searchenginewatch.com/reports/perday.html). Like many search engines, Google accepts natural language queries, joining terms with AND by default. Google uses a stop list, and displays to the user terms which were eliminated from the search. Words such as the questions terms (who, what, where, when, why) and many other common words seem to appear in the stoplist. A query seems to be limited to ten non-stop word terms and to not be stemmed.

Tasks

Sixteen tasks (devised by the TREC 10 Interactive Track participants) were used in the study. The questions came from four domains: Medical, Research, Travel and Shopping. Of the 16 tasks, half were fully specified and half were partially specified so that participants could personalize them.

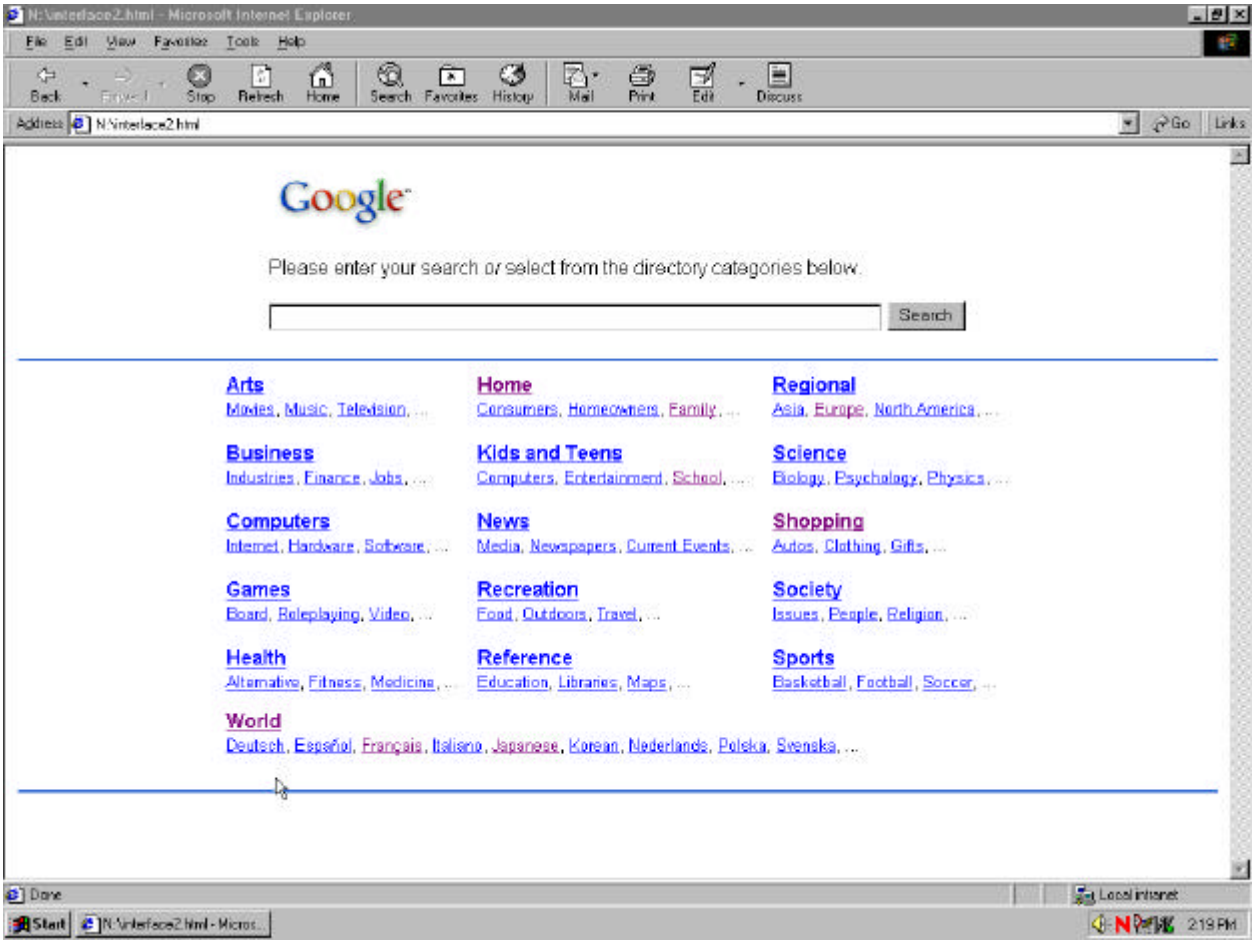

Figure 1. Modified Google interface

Procedure

The participants were recruited in August and September of 2001 in Toronto and Vancouver. Each participant was given four search tasks, one from each of the four domains. Two of the four tasks contained specific questions or imperatives that the participant was asked to respond to by finding relevant Web pages. For the remaining two tasks, participants were asked to provide a topic of personal interest to them, but within the general topic domain pertaining to the task, e.g. Medicine. We used a modified latin squares method to distribute the question variations among the participants.

We also used two different sets of search instructions. For the first two topics, we asked the participant to either enter the query as a list of one or more words or phrases or to select a category from the directory. For the last two topics, we asked the participant to either enter the query as a *complete* question or sentence or to select a category from the directory.

Each participant session lasted approximately two hours. During this time, participants

- completed a demographics and web/search experience questionnaire.

- were assigned four search tasks in sequence. For each task the participant completed a pre-search questionnaire to establish their familiarity with the topic, searched for the topic using the Web interface, and responded to a post-search questionnaire about the search process and their satisfaction with the results.

- described their search while a screen-capture video of the search was replaying. During these retrospective interviews, we tried to elicit the decision-making process used at each stage in the search process.

- responded to a series of questions regarding the search process as a whole. This final interview, which lasted about 10 minutes, was intended to get the participant to comment in a personal way about the challenges that they have in searching the WWW.

This paper reports, primarily, on the results of the data collected in steps 1 and 2 above.

Data was collected using four mechanisms:

1. Questionnaires for demographics, and pre- and post-search evaluations.

2. Audio-tape for all semi-structured interviews.

3. Transaction logs; the *WinWhatWhere* software used captured the titles and URLs of all sites visited, and all keystrokes entered.

4. Screen capture to visually record the user process; Lotus *Screencam* software records in real-time each user session and stores it for playback.

Data Analysis

Data from the pre- and post-search questionnaires and the demographics survey data were combined with data from the transaction logs. Because of some inconsistencies in the WinWhatWhere files, we manually coded the search state, such as query use, category selection, hit list selection, view a URL, and so on, by reviewing the ScreenCam files and the WinWhatWhere files together. The additional coding made it possible to identify the path taken in each search, and to determine the amount of time spent at each state and to identify the rank position on a hit list page of each selected URL. In addition, audio-tapes were transcribed and the content is being analyzed (and is not included in this report).

Results and Discussion

Summary of Results

The 48 participants spent about 7 minutes doing each task. They used the search box for about 66% of the tasks and selected from the directory categories for the remainder. On average, they examined about 5 URLs and about 6 links within each of those URLs. They tended to select about the fourth item on a hitlist and on average examined about two pages of hitlists.

Participants reported little familiarity with the topics for each of the assigned tasks, with few having ever done a search on any of the topics prior to the session. On a five-point scale, they rated the degree of certainty with which they found their answer, the ease of finding the answer, and their satisfaction with the process of finding their answer at around 4.

User-Specified vs. Researcher Specified Task

Half the questions were completely specified and half were fill-in-the-blanks, allowing some user modification toward personalizing the task. There were no significant differences between the two types on any measure. This finding challenges the assumption that information retrieval experimentation with pre-defined queries alters user behaviour in experimental settings. Our participants performed about the same regardless of whether they were assigned a task or allowed to create their own. That said, it is likely that the artificially of the process, e.g., time constraints, lab setting, and so on, may have a greater impact than the nature of the task.

Tactic Used

Participant search paths were analyzed according to the strategy taken in finding information. To start, they could have elected to use a query or a category, and could have changed that tactic to the other technique at any time during the process. Some participants, for example, used a single tactic such as queries only, while some used novel strategies that combined queries and categories as illustrated below:

Strategy	N	Code used in Table 1
used queries only	104	Qry
used categories only	25	Cat
used queries and then selected categories	24	Qry -> Cat
used categories and then selected queries	39	Cat -> Qry

A caveat of this result is the effect of an inherent bias towards the query. While the initial start page contained both categories and a query box, once a query was used, the categories had to be sought out. On the other hand, the query box is an integral part of the second and subsequent category pages, appearing at the top of each, and on each hitlist page.

Twenty efficiency, effectiveness and satisfactory metrics were used to assess the strategies used by participants. Most of this data was derived from the transaction logs or self-reported by the participant in pre- and post task questionnaires. Results from analyses of variance for each of the measures appear in Table 1.

Table 1. Results on all measures by approaches used in the search

Metric	Strategies Used: Qry -> Cat	Cat -> Qry	Qry	Cat	Statistical Significance
Average Number of Instances of Each Search State Per Task					
# of Queries	2.3	.72	1.9	.0	F(3,192)=18.758, p<.001
# of Categories	4.2	5.3	.05	5.9	F(3,192)=35.236, p<.001
# of URLs	6.8	5.4	4.4	3.1	F(3,192)=3.922, p=.010
# of Print	1.8	1.9	2.2	1.5	F(3,192)=2.862, p=.038
# of HitLists	7.9	6.6	6.0	1.0	F(3,192)=9.719, p<.001
# of in Site links	5.8	6.6	6.4	6.2	*ns*
Average Time (seconds) Spent at Each Search State Per Task					
Query Time	29.0	36.2	21.3	.88	**
Hit List Time	120.9	145.9	144.4	41.2	**
URL Time	113.3	114.1	98.5	77.0	*ns*
Category Time	123.8	67.2	1.3	203.5	F(3,188)=5.830, p=.001
Print Time	170.9	84.3	207.8	114.5	**
In Site Time	109.3	138.6	116.8	122.1	*ns*
Position of URLs in HitList					
Avg. Rank	4.7	4.2	4.4	4.8	*ns*
Min Rank	1.7	1.9	2.3	3.4	*ns*
Max Rank	8.6	7.1	6.7	6.9	*ns*
Average User Perception Rating Per Task (scale from 1 to 5)					
Familiarity	2.7	2.4	2.3	2.2	*ns*
Certainty	3.7	3.5	4.0	4.1	F(3,190)=2.901, p=.036
Ease	3.4	3.2	3.9	4.0	F(3,190)=4.543, p=.004
Time allotted	3.6	3.1	3.5	3.8	F(3,190)=2.883, p=.037
Satisfaction	3.4	3.3	3.7	4.0	F(3,190)=2.593, p=.054

There were two key differences in the strategy data. There was a key distinction between the two single-tactic strategies and as well, between the single and mixed strategies. Participants who used only Categories looked at significantly fewer hit lists and spent less time looking at those lists than those who used only Queries. The use of categories seems to have led participants to sites that were more specifically related to the topic, while those who used Queries seemed to examine more pages of hitlists and spent more time doing so. But there were no user perception differences between those who used the single tactic strategies. Both query only and category only participants were equally satisfied with the task and with the ease with which the task was completed.

Participants who chose either Queries only or Categories only tended to find it easier and more satisfying, and were more certain about their results that the mixed approaches. In addition, participants who used a single tactic, i.e., categories or queries, seem to be more successful. Those who used mixed tactics felt the task was more difficult and were less satisfied than the single tactic strategies. These also, selectedly, scored lower on some of the count and time

efficiency measures.

Search as Question vs. Search as Keyword

Participants were required to answer half the tasks using a question or statement and half using keywords or phrase. Because of the interface, they could, however, choose to use the search box or select from the categories. In general participants used the searchbox 66% of the time, but the selection from categories versus use of the searchbox was clearly related to the way that participants were required to enter the query. When asked to use search with a question, the use of categories increased significantly ($?^2$=6.0, p=.014).

Table 2. Participants first tactic by the type of search entry

		Search Entered		
		in question form	**as keyword(s)**	Total
Start With:	**Directory**	40	24	64
	Searchbox	56	72	128
	Total	96	96	192

Participants examined fewer hitlists when using categories (4.3) than searchbox (6.3) (F(1,192)=161.461, p=.017). In addition, when asked to provide a question, they tended to provide keywords or phrases for a significant number of question-based queries (F(2,189)=3.844, p=.023. Participants also tended to rate the task as a question as more difficult that the task as a keyword (F(2,189)=5.986, p=.015. We believe that participants were challenged by this task as it did not represent the way that they normally conceptualize the search process. Thus to avoid asking a question, they opted for categories.

In addition, those who asked questions tended to create longer queries – from 2.6 to 5.8 words (F(2,165)=421.469, p<.001). But the increase in size was accounted for primarily by stopwords (F(1, 166)=65.663, p<.001). Queries as questions had approximately 3.7 stopwords while those entered as keywords had on average about 2.5 stopwords.

Type of Task

Four different task domains were used in this study: Medicine, Research, Shopping and Travel. Results for various measures across each task domain appear in Table 3. There are few significant differences among the four domains.

There were however differences in post hoc Bonferroni adjusted tests. Participants did more printing in Research than Shopping (p=.035) and spent less time in Categories while responding to a Research task than a Travel task (p=.010). Research was perhaps the most complex and cognitively challenging task. Categories were rarely used for Research tasks, and participants spent little time examining categories while doing a Research task. However, Categories were used almost identically across the other three tasks. We can speculate that the presence of top level categories that were semantically related to the assigned task made it easier to use in Shopping, Medicine and Travel, and, additionally, that those tasks were more specific than the Research task.

The type of task had an effect on user perception. In general Participants found the

Shopping task more difficult and less satisfying than the other tasks, rating these an average of 3.3 and 3.1 on a five-point scale.

Table 3. Various Metrics across Type of Task

		Type of Task				
Metric	**Mean**	**Medical**	**Shopping**	**Travel**	**Research**	**Statistical Significance**
Average Number of Instances of Each Search State Per Task						
# of Queries	1.7	1.3	1.7	1.5	2.0	*ns*
# of Categories	2.4	1.1	2.4	2.8	2.2	*ns*
# of URLs	4.7	4.2	4.7	5.0	5.0	*ns*
# of Print	2.0	2.0	1.6	2.2	2.2	**
# of HitLists	5.7	5.4	5.3	5.2	6.8	*ns*
# of In-Site links	6.4	3.8	8.0	8.2	5.5	*ns*
Average Time Spent at Each Search State Per Task						
Query Time	22.6	15.5	24.1	29.4	21.6	*ns*
Hit List Time	128.3	144.3	107.2	100.8	161.0	*ns*
URL Time	100.7	117.2	90.9	87.7	107.0	*ns*
Category Time	56.3	15.8	51.2	99.5	58.9	*ns*
Print Time	166.0	174.1	129.2	205.0	155.6	*ns*
In site Time	121.0	70.3	156.0	159.5	98.1	*F(3,192)=5.072, p=.002*
Position of URLs in HitList						
Avg. Rank	4.4	4.0	4.9	4.4	4.5	*ns*
Min Rank	2.3	2.0	2.7	2.3	2.1	*ns*
Max Rank	7.0	6.5	7.7	7.0	7.0	*ns*
Average User Perception Rating Per Task						
Familiarity	2.4	2.5	2.2	2.1	2.6	**
Certainty	3.9	4.2	3.7	4.0	3.8	**
Ease	3.7	3.9	3.3	3.8	3.7	**
Time allotted	3.5	3.7	3.2	3.5	3.5	**
Rating	3.6	3.9	3.1	3.7	3.8	*F(3,189)=5.191, p=.002*

** selected results were significant in Bonferroni adjusted post hoc tests

In addition, we mapped strategies by domain as illustrated in Figure 2. The mixed strategies are evenly used across the four domains. But when the directory categories were used as a tactic, it tended to be for travel topics.

Discussion and Conclusions

We discovered that researcher-specified versus participant-personalized queries had no effect on results. The domain of the task, too, appears to have had little effect, although the Shopping tasks tended to be more difficult to complete, and were generally the least satisfying. The use of categories seems to have influenced the search process itself, where more time was spent contemplating the nature of the search task at the beginning of the process, resulting in

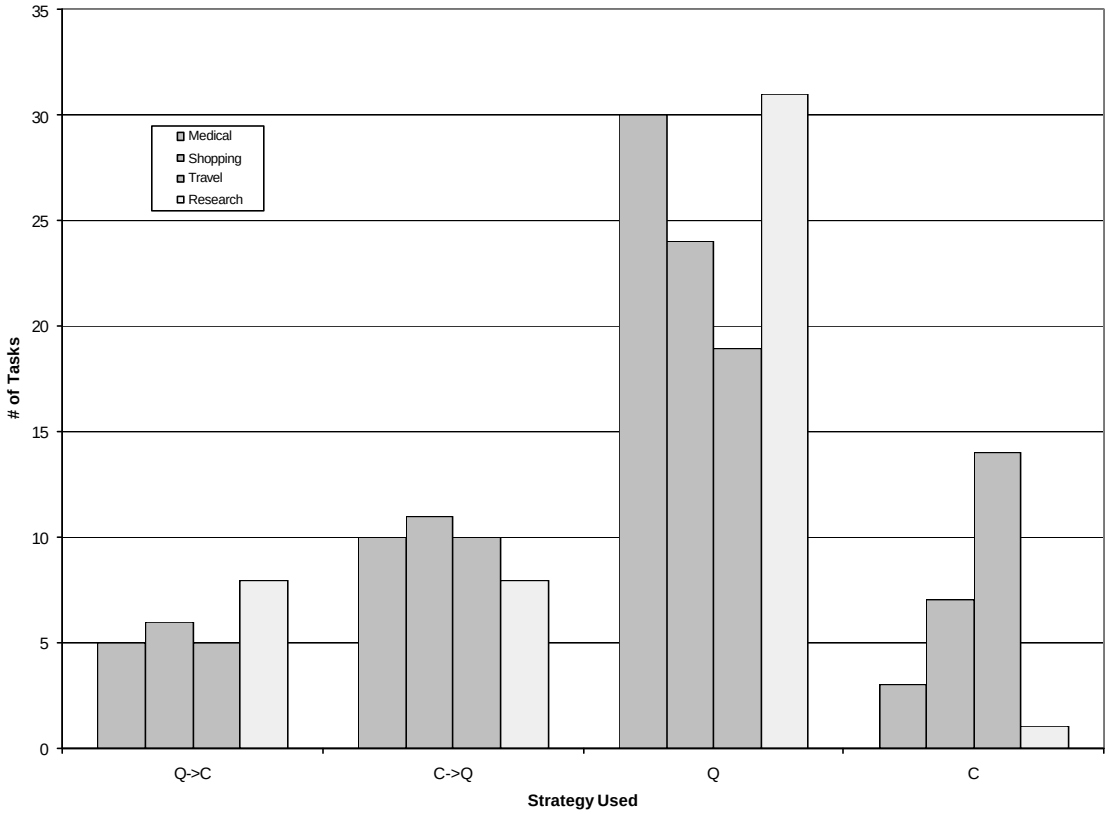

Figure 2. Domain by Strategy Used

fewer items being selected from the hit list, and marginally less navigation within a site once there. Anecdotally, participants indicated a need to develop a broad perspective before focusing on specific results.

When participants are asked to express a search statement in the form of a question or statement that they had only modest success in even achieving that task. The choice between initiating a search with a search box or with a selection from the categories seems dependent at least partially on the manner in which the query is entered. Participants were more likely to search using the categories when they were requested to create the query as question. It seems the prospect of using a question posed difficulties for participants. When entering keyword queries, the number of keywords used, on average, was quite small. Likely participants have learned one way of conceptualizing the search process and have developed a fixed mental model of that process which constrained their ability to provide richer search statements.

Future analysis will use the verbal protocol data collected to enhance our interpretation of the current findings. From the examination of this protocol data we hope to gain a richer explanation of not only what was observed in the findings reported here, but of why participants chose the courses of action they did for the various tasks performed. For example, why was the query constructed in that particular manner? What did the participant think it would achieve? Why did they choose to search using categories or queries? How did they select from the results list? And, how did they decide if a site was useful? In addition, we hope to pinpoint the problems within

the search process. When participants appeared to be off course, what might have been useful to help get them back on track?

Future Research

Based on this current work we also hope to carry out two additional studies that will focus on: i) developing a more refined experimental approach to the category and query integrated search, and ii) manipulating how people conceptualize the query process.

Acknowledgments

This work was partially funded by an Natural Sciences and Engineering Research Council of Canada grant to the first author. The authors wish to thank Research Assistants, J. Heaton and A. Olsen in Vancouver and A. Lebowitz in Toronto.