TREC11 Web and Interactive Tracks at CSIRO

Nick Craswell[1] *David Hawking*[1] *James Thom*[2]
Trystan Upstill[3] *Ross Wilkinson*[4] *Mingfang Wu*[4]

[1]Enterprise Search
[4]Technologies for Electronic Documents
CSIRO Mathematical and Information Sciences
{Nick.Craswell; David.Hawking; Ross.Wilkinson; Mingfang.Wu}@csiro.au

[2]School of Computer Science and Information Technology, RMIT University
jat@cs.rmit.edu.au

[3]Department of Computer Science, Australian National University
Trystan.Upstill@anu.edu.au

1. Overview

This year, the CSIRO teams participated and completed runs in two tracks: web and interactive.

Our web track participation was a preliminary exploration of forms of evidence which might be useful for named page finding and topic distillation. For this reason, we made heavy use of evidence other than page content in our runs.

In the interactive track, we continue to focus on answer organization issues, aiming to investigate the usefulness of the knowledge about “organizational structure” in organizing and delivering the retrieved documents. For the collection of the US government (.gov domain) web documents, we used their level two domain labels and their corresponding organization names to categorize the retrieved documents. For example, documents from the "nih.gov" domain will be put into the “National Institutes of Health (nih)" category. We compared this delivery method with the traditional ranked list. The preliminary results indicate that subjects achieved a significantly better performance with the category interface at the end of fifteen minutes search, however, there is no significant difference between the two methods during the first five or ten minutes. The experiment result also shows that the category interface assisted subjects answer the more complex topics as time increases.

2. The web track

2.1. Topic distillation

In topic distillation we used the following forms of evidence:

- BM25 on content. Pages returned should be relevant. We indexed the .GOV corpus and applied BM25, sometimes with stemming sometimes without.
- BM25 on content and referring anchor text. An alternative to content-only BM25 is to include referring anchor text words in the BM25 calculation (content and anchors).

- In-link counting and filtering. We expected pages with more in-links to be potentially better answers, and we differentiated between on-host and off-host links. We also eliminated many results on the grounds that they had insufficient in-links.
- URL length. We expected short URLs to be better answers than long URLs
- BM25 score aggregation. We expected sites with many BM25-matching pages to be better than those with few.

Table 1 reports the results for our topic distillation runs. Our (non-submitted) content-only achieved better performance than any of the submitted runs that included "distillation evidence".

Table 1 Runs for topic distillation

Run	*P@10*	*BM25 content only*	*BM25 content and anchors*	*In-link counting and filtering*	*URL length*	*BM25 aggregation*
csiro02td1	0.1000	y		y	y	
csiro02td2	0.0714		y	y		
csiro02td3	0.0184	y		y	y	y
csiro02td4	0.0184		y	y		y
csiro02td5	0.0939	y (stem)		y	y	
csiro02unoff	0.1959	y				

In this year's topic distillation task, the focus on local page content relevance ("BM25 content only") was probably too high for our non-content and aggregation methods to succeed (our "distillation evidence"). We expected most correct answers to be shallow URLs of sites containing much useful content. In fact, correct answers were deeper, and our aggregation method for finding sites rich with relevant information was actually quite harmful (runs 3 and 4). The focus on page content is borne out by the improvement in effectiveness achieved when we apply simple BM25 in an unofficial run (csiro02unoff). To perform better in this year's task, we should have put less (or no) emphasis on distillation evidence and far more emphasis on relevance. However, we also believe that in some Web search situations, the distillation evidence would be more important than it was in this year's task.

2.2. Named Page Finding

In our named page finding experiments we used the following forms of evidence:

- BM25 on content and/or anchor text. We indexed the .GOV corpus and applied BM25 to document content and to surrogate documents that contained all anchor text pointing to a page. Stemming of query terms was also employed.
- Extra Title Weighting. To bias our results towards what we thought would be page naming text we put further emphasis on document titles.
- PageRank. To see whether link recommendation could be used to improve results we incorporated this link popularity measure [3].

Table 2 shows the results for the named page finding runs. The BM25 content-only submission performed the best. We tried combining content evidence with anchor-text and PageRank but both combinations harmed retrieval effectiveness.

Table 2 Runs for named page finding

Run	*ARR*	*S@10*	*BM25*	*Stemming*	*Extra Title Weighting*	*Small Crawl PageRank*
csiro02np01	0.573	0.77	Content			
csiro02np02	0.241	0.34	Anchor text			
csiro02np03	0.416	0.59	Content		y	
csiro02np04	0.318	0.51	Content and anchor text	y	y	
csiro02np16	0.307	0.49	Content and anchor text	y	y	y

Prior to submission we generated 20 training queries and found content with extra title weighting performed best. We expected page titles to be important evidence in named page finding, however this appeared not to be the case – in fact extra title weighting for the TREC queries appeared to reduce effectiveness (run 1 vs run 3). While there was some anchor text evidence present for the query set (run 2) when we combined this evidence with content (runs 4 and 16) results were noticeably worse than for the content-only run (run 1). PageRank harmed retrieval effectiveness (run 16 vs run 4).

3. The interactive track

On the Internet, the information source and its information provider indicate not only the quality and credibility of the information, but also the type and content of the information. When people try to access information from an organization's website, they very often try to match their mental model about that organization with their information needs. They can usually identify a few related departments in that organization, and search the information within these departments.

We can consider the whole worldwide web as the web site of a global organization with a hierarchical structure. Documents in this space could be categorized by their "functional departments" corresponding to their domain names. For example, the level one domain labels can categorize the documents into government (.gov), university (.edu), military (.mil), and commercial (.com) etc. (In fact, they should be the level two domain labels, with the level one label of .us) ; the level two domain label can be used to further categorize the documents within the first level domain.

In this year's interactive track, all documents in the collection are gathered from the US government domain (.gov). The test topics also cover various areas, such as government policy, medicine/health and travel. To this collection and the topic set, our intuition was to organize and delivery the retrieved documents according to the US government functional (or departmental) structure. We intent to use this dynamically generated organizational structure to organize the distributed documents retrieved from the web, and guide users to focus their attentions on the information sources and/or information providers. We hypothesized that this structure (called categorization structure) would serve as a better guide for a user to locate relevant and authoritative information than the traditional ranked list, thus improving the user's performance with the search tasks.

3.1. Experimental setting

3.1.1. Delivery interfaces

The Panoptic [1,2] is used as the back-end search engine in both delivery methods. In the categorization delivery method, the categoriser classifies the retrieved documents according to the

level two domain labels. Each category label is obtained by expanding the domain label into its owner's organizational name through the "whois" server (http://www.whois.nic.gov). For example, all documents from the "nih.gov" domain will be put into the "National Institutes of Health" category. The documents in a category are ranked according to their original rank in the returned ranked list, and the categories are ranked according to the original rank of the first document of each category. The category interface shows the first category by default.

The interfaces for the two different delivery methods are shown in Figure 1 and Figure 2. We have been trying to keep the two interfaces as consistent as possible, differing only in their presentation of the alternate structures. Both interfaces are divided into three areas: the top area shows the current search topic and provides three buttons for the subjects to save answer and move on to the next topic. The middle area is the query area that has a query box and information on query word matching. The bottom area is the main area that shows either the ranked list or the categorized result.

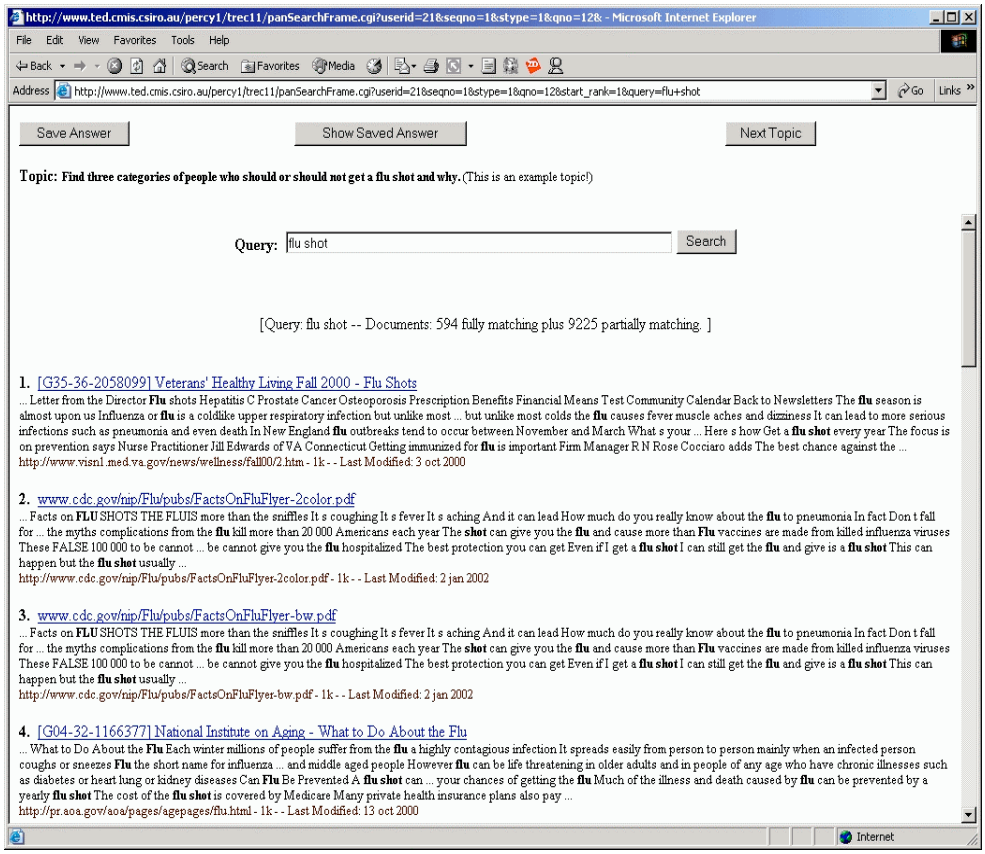

Figure 1 The delivery interface for the ranked list

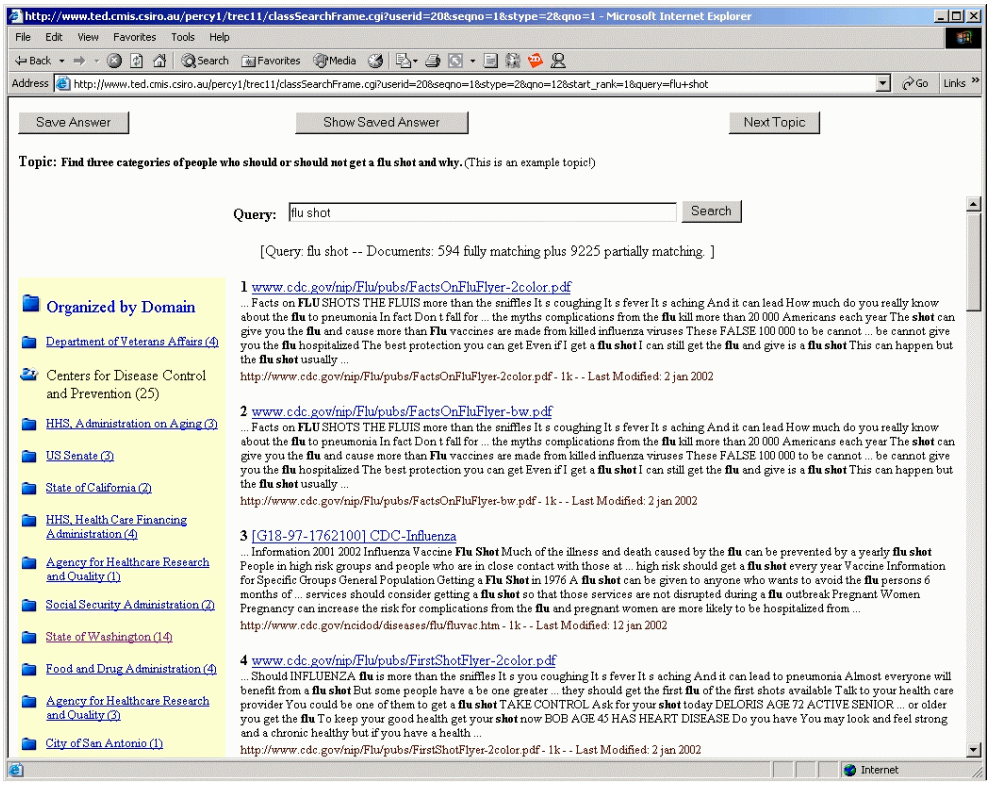

Figure 2 The delivery interface for domain categorization

3.1.2. Experimental procedures

During the experiment, all subjects are asked to follow the following procedures.

- Subjects filled in the pre-search questionnaire about their demographic information and their search experience.
- Subjects were then shown the two experimental interfaces, and were free to ask any question related to the use of the two interfaces.
- Subjects were assigned to the experimental design that was used by all participant groups in the interactive track. In this experimental design, subjects searched four topics on each interface, the sequence of interface and topics varied among subjects. A complete such a design requires a group of 16 subjects.
- Prior to each interface, subjects had hands-on practice with an example topic, and got familiar with each interface.
- Prior to each interface, the query "information retrieval" was issued by the corresponding system automatically to calibrate the difference between two systems' response time. Subjects were asked to click the "Next Topic" button when they saw the search result appeared. The average response times are 6.8 seconds for the ranked list interface and 8.3 seconds for the category interface.
- Prior to the search of each topic, subjects were required to fill in a pre-search questionnaire about their familiarity with the topic. After the search of the topic, subjects filled in a post-search questionnaire about their experience of that particular search topic.
- Subjects filled in a post-system questionnaire after each interface.

- Subjects filled in an exit questionnaire in the end of the experiment.

At any time during a topic search, subjects could move on to the next topic whenever they found the required answer and were satisfied with what they have found. We encouraged our subjects to find answers to a topic within ten minutes, however they could have an extra five minutes in case they could not find the required answer in the first ten minutes and want to continue their search.

Transaction logging, questionnaire, and screen recording are the main methods to collect data. During each search session, every significant event - such as document read, the instance saved and the supporting source document and the query sent - was automatically captured. Questionnaires are those common to all participant groups in the interactive track. Screen recording was used to capture the search process for further detailed analysis.

3.1.3. Subjects

All our sixteen subjects were university students. These subjects came from various backgrounds, such as computer science, media study, law and mechanical engineering. Of the sixteen subjects, fourteen are male and two are female. Fifteen of them are in the age group 18-27 years, only one is in the age group 38-47 years. Table 3 lists subjects' responses to the selected questions from the pre-search question (all are on 7-point Likert scale). From the table, we can see that our subjects search the web very often (Q1, mean=5.81), can usually find what they are looking for (Q5, mean=5.38), and generally regard themselves as experienced searcher (Q10, mean=4.73). Comparatively, subjects use the search box (Q6, mean=5.19) more often than browsing mechanism (Q7, mean=4.06). These subjects very often search for information related to assignments (Q8-1, mean=5.38) and entertainment (Q8-6, mean=5.19), while search less on shopping (Q8-2, mean=3.19), government policy (Q8-5, mean=3.06), and traveling (Q8-3, mean=2.94), and least on medical/health (Q8-4, mean=1.94). (While our test topics cover the government policy, traveling, and medical/health.)

Table 3 The selected questions[1] from the pre-search questionnaire.

	Q1	**Q5**	**Q6**	**Q7**	**Q8-1**	**Q8-2**	**Q8-3**	**Q8-4**	**Q8-5**	**Q8-6**	**Q10**	**Q11**
Mean	5.81	5.38	5.19	4.06	5.38	3.19	2.94	1.94	3.06	5.19	4.73	4.53
Std	1.05	0.96	1.83	1.98	1.09	1.64	1.06	1.18	1.65	1.42	1.34	1.61

3.2. Results

3.2.1. Performance with two interfaces

The effectiveness of the two interfaces is measured by the success rate: the ratio of the correctly saved instances. There are two types of topics in this year's interactive track. Type I topic is "find X instance of …". Type II topic is "find a website that is a good resource on Y". For the topics of type I (topic 1, 2, 4, 5 and 6), each instance correctly identified and supported by a document will be given a score 1/n, where n is the number of required instances for the search topic. Type II topic can be regarded as a special case of the type I where the required instance is 1. So for the topics of type II (topic 3, 7 and 8), the score is binary: 1 - for the correctly identified website, and 0 – for a website that does not give information on the topic. (A 5- or 7- point Likert scale could be used to judge the degree of "goodness" of the saved website. However, this kind of judgement might be too subjective to reach consistence. So we adopted the binary score.)

Table 4 shows the subjects' search performance at three period cut-off - after five minutes, ten minutes, and fifteen minutes. On the average, the performance with the category interface is

[1] Questions are listed in the Appendix I. All responses are on a 7-point Likert scale.

Table 4 Subjects' search performance per topic at three period cut-off

Topics		1	2	3	4	5	6	7	8	Mean	Std	p <
5Min	**List**	0.38	0.08	0.75	0.44	0.42	0.21	0.13	0.63	0.38	0.24	0.26
	Cate	0.38	0.04	0.63	0.22	0.25	0.42	0.00	0.63	0.32	0.24	
10Min	**List**	0.38	0.25	1.00	0.75	0.79	0.42	0.38	0.88	0.61	0.28	0.48
	Cate	0.58	0.33	1.00	0.53	1.00	0.63	0.25	0.88	0.65	0.29	
15Min	**List**	0.38	0.54	1.00	0.84	0.92	0.50	0.50	0.88	0.69	0.24	0.05
	Cate	0.75	0.63	1.00	0.75	1.00	0.88	0.63	1.00	0.83	0.16	

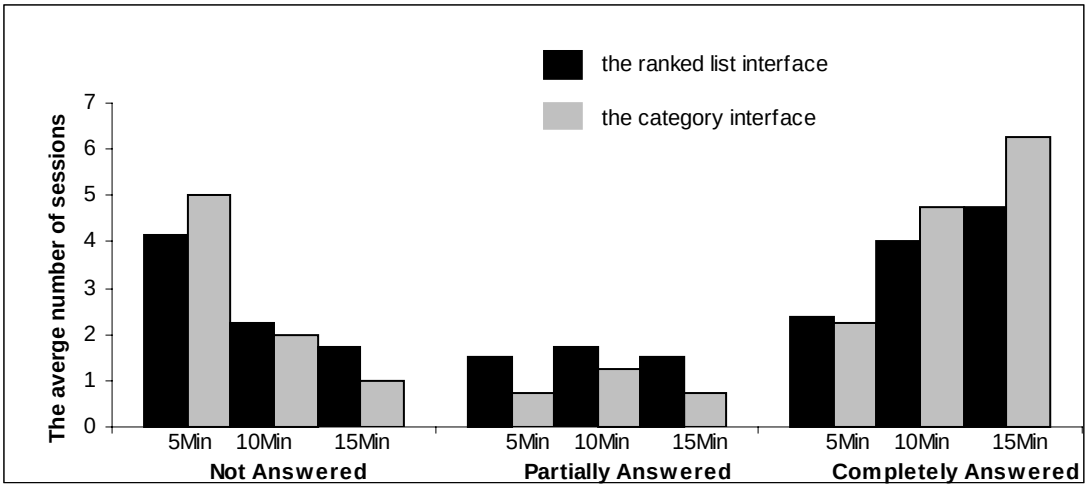

Figure 3 The completeness of the saved answers

lower than that with the ranked list interface at the end of the five minutes, higher than the ranked list interface at the end of the ten minutes, and significantly outperform the ranked list interface at the end of fifteen minutes. (two tailed, paired t-text)

Before the search of each topic, subjects were asked about their level of familiarity with the topic on a 7-point Likert scale. On the average, our subjects have low familiarity with all topics (Ranked list: Mean = 2.14, Std = 0.70; Category: Mean = 2.19, Std = 0.72). Although the correlation between the success rate with the ranked list and the familiarity ($r = 0.51$) is higher than the correlation between the success rate with category and the familiarity ($r = -0.0004$), nevertheless, neither of the correlations is significant.

Figure 3 shows the breakdown of the success rate according to the sessions in which a question is either "not answered", "partially answered", or "completely answered" at three cut-off periods.

At the end of the first five minutes, the number of "not answered" sessions with the category interface is more than the number of those with the ranked list interface. However, with the increase in time spent, the number of "not answered" sessions with the category interface decreases, although the difference is not significant at each cut-off period.

At the end of each cut-off period, the number of "partially answered" session with the category interface is always less than the number of those with the ranked list interface, although the difference is not significant either.

At the first cut-off period, subjects have less "completely answered" sessions with the category interface than that with the ranked list interface (not significant). However, at the second and third cut-off period, subjects have significantly more "completely answered" sessions using the category interface ($p < 0.05$ at tenth minute, and $p < 0.01$ at the fifteenth minute). Looking at

topic by topic at the fifteenth minute, the category interface is performing better for 7 out of 8 topics. In the only exceptional topic – the topic 3, the two interfaces performed the same with the same number of "completely answered" sessions). For the topics 1, 2, and 6, the number of "completely answered" sessions with the category interface is twice that with ranked list interface. Here the topic 1, 2 and 6 are all of the type I.

We had assumed that the type I topics might need to gather instances from multiple documents, but this is not always the case – sometimes a document may contain enough information to cover all required instances. Table 5 shows the distribution of the "completely answered" sessions from either the multiple documents or one document only. In four out of five such type I topics, there are more sessions with the category interface in which the saved answers come from multiple documents. This may suggest that the category interface is more helpful for the more complicated tasks.

3.2.2. Subject's effort

The subject's effort for getting an answer is measured by the time, the number of documents read, and the number of queries sent in order to get a complete answer or reach the end of each session.

Table 6 shows the average time spent in order to get a complete answer by the two quickest subjects using each interfaces. On the average, the quickest two subjects using the category interface took less time than the quickest two using the list interface, but the difference is not significant here. If we look topic by topic, the two subjects are quicker using the category interface only for three topics – the topic 1, 2 and 4; there three topics are of type I.

Table 7 shows the interaction between the subject and the interface. On the average, subjects read more documents with the category interface (Mean=4.81) than with the ranked list interface (Mean=4.74), but sent less queries with the category interface (Mean=3.0) than with the ranked list interface (Mean=3.54). This indicate that the ranked list interface may encourage subjects to rephrase queries, while the category interface may encourage subjects to browse the answer structure, thus read more documents.

Usually subjects read and saved documents high in rank from the ranked list interface - the average rank of the read and saved documents is 4.97 and 4.87 respectively. While the average rank of the read and saved documents from the category interface is 19.10 and 16.5 respectively. This may indicate that the category interface may be able to bring relevant (or related) documents in a category; these documents may scatter in the ranked list, while a subject may not go that far to get that relevant document with the ranked list interface.

Table 5 The source of the complete answers

(M: from multiple documents; S: from one document only)

Topic	**1**		**2**		**4**		**5**		**6**	
	M	**S**	**M**	**S**	**M**	**S**	**M**	**S**	**M**	**S**
List	0	3	0	2	2	2	1	6	2	1
Cate	1	5	2	2	2	3	3	5	5	1

Table 6 The average time (in minute) spent to get a complete answer by the two quickest subjects.

Topic	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Mean**
List	4.03	12.85	0.75	6.59	2.01	6.08	4.55	1.32	4.77
Cate	3.54	9.03	2.68	4.36	3.93	6.24	5.82	1.44	4.61

Table 7 Subject-interface interaction

		Mean	Std	P < (2 tail t-test)
Number of documents read	**List**	4.74	2.52	NS
	Category	4.81	1.54	
Number of queries	**List**	3.54	1.93	NS
	Category	3.00	1.39	
The ranking of the read documents	**List**	4.97	2.05	0.0007
	Category	19.10	7.94	
The ranking of the saved documents	**List**	4.87	2.62	0.04
	Category	16.5	14.32	

Table 8 Subjects' response to the post-search questionnaire

	PS1 (easy to start)	**PS2** (east to search)	**PS3** (satisfaction)	**PS4** (timeliness)	**PS5** (knowledge helped?)	**PS6** (learn something new)
List	4.59	4.24	4.49	5.22	2.22	4.13
Cate	4.11	3.97	4.25	4.19	2.19	3.69

3.2.3. Subject's satisfaction

After the search of each topic, subjects filled in a post-search questionnaire that was to get the subject's satisfaction of that particular search topic. Table 8 shows the subjects' response to each question. For all questions, the average response from the subjects using category interface is lower than that from the subjects using the ranked list interface, although no significant difference is found between the two interfaces for any questions.

We checked the correlation between the each question and the success rate, significant positive correlation is found only between the PS3 (satisfaction) and the success rate (in both interfaces, $r = 0.69$, significance at 0.05). That may be truism: if subjects saved more answers, they are getting more satisfied.

In the exit questionnaire, when the subjects were asked about which of the systems they like the best overall, 11 subjects chose the category interface, 3 subjects chose ranked list interface, while the remaining 2 thinking there is no difference between the two interfaces.

3.3. Discussion

Our experimental results indicate that the users may be able to find the answer quicker with the ranked list interface for those easy search tasks where the search engine is able to bring the relevant documents on the top of the ranked list. However, for more complicated tasks where an answer is to be synthesized from multiple documents, and those documents are scattered along the ranked list, the user may perform better with the category interface. This performance is achieved by spending longer reading or browsing time. One possible reason might be related to the categorization structure itself: the current one-level flat structure may not be very clear to the subjects. It could be enhanced by having a multi-level hierarchical structure closely reflecting the US governmental structure.

4. References

[1] CSIRO, http://www.panopticsearch.com

[2] David Hawking, Peter Bailey and Nick Craswell. Efficient and Flexible Search Using Text and Metadata. CSIRO Mathematical and Information Sciences, TR2000-83, http://www.ted.cmis.csiro.au/~dave/TR2000-83.ps.gz

[3] Lawrence Page, Sergey Brin, Rajeev Motwani and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. *Standford University Database Group*. http://dbpubs.stanford.edu:8090/pub/1999-66

5. Appendix

5.1. Selected responses from the pre-search questionnaire

Q1: How much experience have you had searching with WWW search engine?

Q5: When I search the WWW, I can usually find what I am looking for.

Q6: I always use the query box – keeping rephrase my queries until I find the right information.

Q7: I always browse web directory (e.g. Yahoo subject directory, etc) to get the information.

Q8-1: How often do you conduct searching for information about assignment/work related project?

Q8-2: How often do you conduct searching for information about shopping?

Q8-3: How often do you conduct searching for information about traveling?

Q8-4: How often do you conduct searching for information about medical/health?

Q8-5: How often do you conduct searching for information about government policy?

Q8-6: How often do you conduct searching for information about entertainment?

Q10: Please indicate your level of expertise with searching. (Novice(1)…..Expert(7)).

Q11: Overall, for how many years have you been doing online searching? _____ years.

5.2. Post-search questionnaire

PS1: Was it easy to get started on this search?

PS2: Was it easy to do the search on this topic?

PS3: Are you satisfied with your search results?

PS4: Did you have enough time to do an effective search?

PS5: Did your previous knowledge help you with your search?

PS6: Have you learned anything new about the topic during your search?

(All above questions on a seven-point Likert scale.)