

TREC12 Web and Interactive Tracks at CSIRO

Nick Craswell¹, David Hawking¹ and Trystan Upstill²

nick.craswell@csiro.au, david.hawking@csiro.au and trystan.upstill@cs.anu.edu.au

Alistair McLean¹, Ross Wilkinson¹ and Mingfang Wu¹

alistair.mclean@csiro.au, ross.wilkinson@csiro.au, mingfang.wu@csiro.au

¹CSIRO ICT Centre, Australia

²Department of Computer Science, CSIT Building, ANU

Canberra, ACT 0200, Australia

1. Introduction

This year, CSIRO teams participated in all three tasks of the web track; these being: the automatic topic distillation task, the home/named page finding task and the interactive topic distillation task. This paper describes our approaches, experiments and results. The following section describes our experiments in the two automatic tasks, and Section 3 describes our experiment in the interactive task.

2. The web track

CSIRO submitted a total of 10 runs to the non-interactive portion of the 2003 Web Track - 5 runs for home/named page finding and 5 runs for Topic Distillation. The runs are labeled csiro03[TYPE][RUNID], where TYPE is "ki" for known item runs and "td" for topic distillation runs.

This year we focused on tuning Okapi BM25 and Web evidence parameters. Our home/named page finding submissions use tunings computed for both home and named page finding, and evaluate two run combination methods. Our topic distillation submissions are tuned for home page finding only and test whether the Web evidence evaluated is useful, and whether the use of stemming improves performance.

We did not incorporate PageRank or simple indegree this year because of previously observed poor performance for named page finding and homepage finding. Instead our query-independent Web evidence included URL length and two important sub-types of indegree (off-site and on-site).

Throughout our experiments we tuned Okapi BM25 (through the k_1 and b parameters), anchor-text weighting and other query independent Web evidence. The parameters were tuned using a hill climbing algorithm, with complete exploration of 2 parameters at a time, on a grid computer consisting of cluster of 20 dual processor Intel Xeon machines.

2.1. Home/named page finding

We submitted runs based on three tunings (for a home page task, a named page task and both at the same time), and evaluated two combination methods. We trained using last year's .GOV named page finding query/result set, and using a home page finding training set derived from a first .GOV resource listing.

We submitted runs tuned for both home page and named page finding at the same time (csiro03ki01), tuned for named page finding only (csiro03ki02) and tuned for home page finding only (csiro03ki03). We also submitted two combinations of these runs. The first was an interleaved run (csiro03ki04 -- interleaving csiro03ki02 and csiro03ki03), and the second a run that summed scores achieved in both rankings (csiro03ki05). A summary of our home/named page finding submissions, and their retrieval effectiveness is presented in Table 1.

Table 1: Home/named page submissions summary. To aid our understanding of retrieval performance we computed ARR for home pages only ``ARR (HP)'' and named pages only ``ARR (NP).'' We also computed a further run post-hoc (csiro03kins)

<i>Run</i>	<i>Description</i>	<i>ARR</i>	<i>S@10 (%)</i>	<i>ARR (HP)</i>	<i>ARR (NP)</i>
csiro03ki01	Tuned for HP and NP	0.692	83.7	0.815	0.569
csiro03ki02	Tuned for HP	0.603	77.7	0.774	0.432
csiro03ki03	Tuned for NP	0.702	84	0.755	0.649
csiro03ki04	HP and NP tunings interleaved (HP then NP) w/q.class	0.667	86.3	0.801	0.532
csiro03ki05	HP and NP tunings combined	0.699	81	0.812	0.586
csiro03kins	HP and NP tunings interleaved (NP then HP)	0.717	87	0.781	0.651

Our results show that tuning specifically for the home page finding task significantly harmed our named page retrieval effectiveness (csiro03ki02 vs. csiro03ki03). Our highest ARR was achieved using the NP-only tuning, whilst the best S@10 used interleaved lists from HP and NP tunings. The results report that an overemphasis on home page finding evidence can hinder named page searches.

The run with the highest S@10 (csiro03ki04) interleaved the csiro03ki02 and csiro03ki03 runs (i.e. top HP, top NP, second HP, second NP etc.). To improve early precision, if we encountered a keyword that strongly indicated a named page query¹ was occurring we led with the top NP, rather than the top HP result. From further post-hoc evaluations (see csiro03kins) we determined that leading with NP rather than HP would have further improved precision (achieving an ARR improvement of 0.717). In summary, interleaving HP then NP without query classification achieves an ARR of 0.646. Interleaving HP then NP with swapping if query appears to be a named page query achieves an ARR of 0.667. Finally, interleaving NP then HP without query classification achieves 0.717.

We could not find a single tuning that is equally useful for each type of query. This raises interesting query classification, or further combination of evidence questions. A superior classifier may well have taken into account other evidence, such as query length, while a better combination may have taken into account the strength of the home page evidence (and only led with a homepage result if it was sufficiently strong).

There are some limitations inherent in the training sets we used for tuning. The set of home pages was taken from a .GOV portal, which may inadvertently have favored prestigious, or larger and more popular home pages. Further, last year some of the named pages were in fact home pages, whereas this year there was a distinction made between named pages and home pages. Our named page tuning was therefore based on a mixed query set with a smaller ratio of home pages. This may have slightly biased our training towards home page queries.

2.2. Topic Distillation

Our Topic Distillation runs were based on the home page tunings built for the home/named page task. The run results are presented in Table 2.

¹ Query terms were selected from last years query set and included terms such as ``"page", "form" and "2000"``

Table 2: Topic distillation submissions summary. Post-hoc we computed a run based using the name page tunings (csiro03tdns)

<i>Run</i>	<i>Description</i>	<i>Average R-Prec</i>
csiro03td01	Tuned for HP	0.1438
csiro03td02	Tuned for HP without query-ind. hyperlink evidence	0.1162
csiro03td03	Tuned for HP with stemming	0.1636
csiro03td04	Tuned for HP without anchor evidence	0.0988
csiro03td05	Tuned for HP with “normal” bm25 tuning ($k_f=2, b=0.75$)	0.1217
csiro03tdns	Tuned for NP	0.1166

Our best run (csiro03td03) used the home page tuning and incorporated stemming. When removing hyperlink evidence (i.e. csiro03td02 and csiro03td04) we observed a decrease in retrieval performance. Likewise, we observed a 2% decrease in performance when using standard tunings for Okapi BM25 (csiro03td01 vs. csiro03td05). Post-hoc we computed a new run based on the named page finding tunings used in our home/named page finding submission (csiro03tdns), this tuning reduced the Avg R-Prec to 0.1166.

The results from the topic distillation task appear to support the notion that our home page training set favored prominent resources (an advantage for Topic Distillation). Further, our results illustrate the usefulness of web evidence, and stemming, when addressing Topic Distillation.

3. The Interactive Sub-track

In this year's interactive sub-track, searchers were asked to construct a resource list that covers all major aspects of a broad topic through interaction with an information access system. Similar to that in automatic topic distillation task [1], a key resource page is defined as a main page of a website which is:

1. principally devoted to the topic,
2. providing credible information on the topic, and
3. not part of a larger site also devoted to the same topic.

Take the topic “adoption procedures” and the website <http://www.courtinfo.ca.gov/> as an example, the main page that meets the above requirements is <http://www.courtinfo.ca.gov/shelfhelp/family/adlption/>, all the pages referring to this page or referred to by this page would fail one of the above conditions.

To assess whether a web page is a key resource page, a searcher needs to make the following judgments about the page:

- 1) Is it relevant?
- 2) Does it have the right scope? (Is it too broad or too narrow compared with that of other relevant pages from the same site?)

In the interactive track, searchers were also asked to make one more judgment:

- 3) Does it cover a different aspect from the previous saved web pages?

The traditional ranked list provides users with a set of entry points to their corresponding websites, then users have to browse each website to decide whether the entry point is the main page, or if not, whether there is a page within the site could be the main page. The above three tasks (especially the task 2 and 3) are not explicitly supported by this kind of delivery interface.

We aimed to investigate the effectiveness of a task tailored delivery method to assist searchers in making the above three judgment tasks. Our hypothesis was that searchers would have a better performance on the

assigned tasks by using the interface designed to support the above judgment tasks than a generic interface (such as a ranked list).

3.1. Experimental setting

3.1.1. Delivery interfaces

The Panoptic topic distillation engine was used as the back-end search engine for both interfaces. To concentrate on the comparison of the two delivery interfaces, we decided to fix the query for all topics and for all searchers, i.e. searchers were restricted to explore the same set of retrieved documents. The queries were optimized to return shallow pages from a web site and to make sure the precision at top ten returned documents was acceptable.

The baseline interface (referred to as TDLinear for Topic Distillation with Linear interface), was the delivery interface from the Panoptic topic distillation engine. As shown in Figure 1, this interface provided searchers with a ranked list of top 100 potential relevant key resource pages in five consecutive pages, with each page showing the titles and summaries of 20 documents.

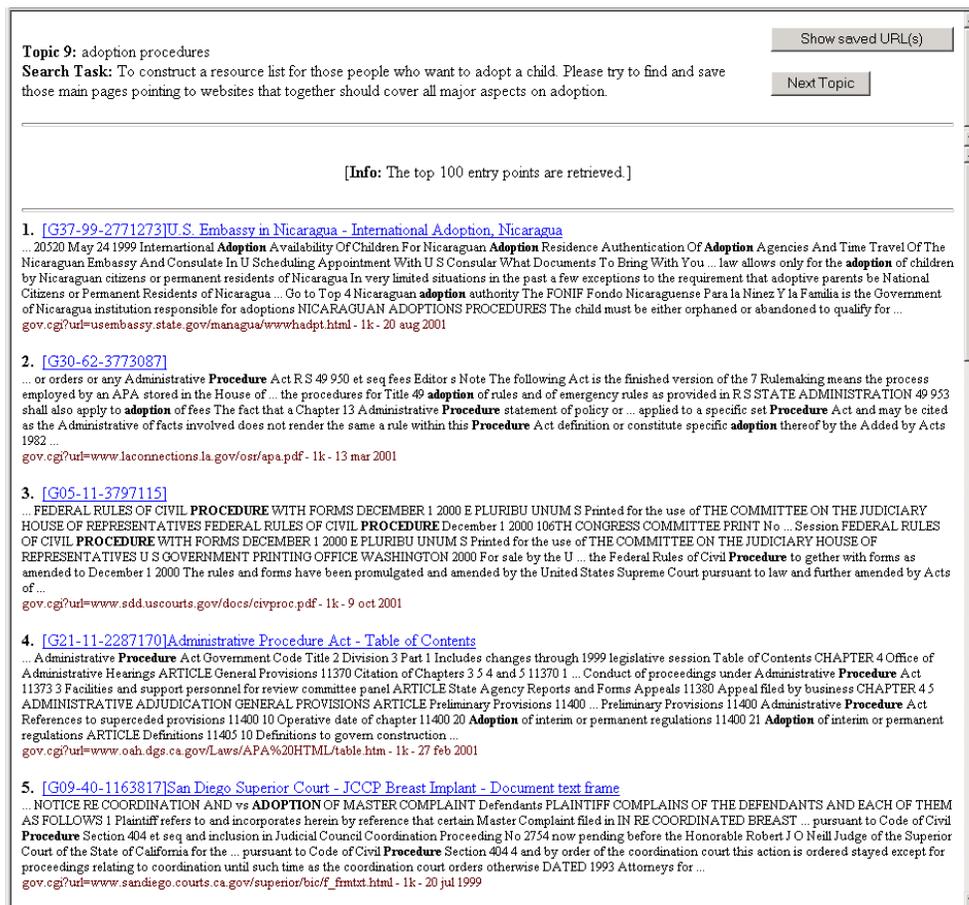


Figure 1. The delivery interface for the ranked list

To validate our hypothesis, we designed the experimental interface (referred to as TDHierarchic for Topic Distillation with Hierarchical delivery interface) that explicitly supports searchers' assessment tasks. The experimental interface consists of two parts: the site summary and the sitemap.

1. The site summary (Figure 2): The top 100 retrieved pages (from Panoptic topic distillation engine) were firstly grouped according to their corresponding departments (organizational structure), and then further sub-divided into their secondary business units (websites). Each of the websites was summarized and represented by using the titles of the top three most relevant pages. The summary not only described the content of the site, but also provided three candidate entry points to the site. We decided not to show the summary of a document directly for two reasons: the summary of the document may not be suitable for the topic distillation task and showing the summary of a document would make the interface cluttered. Instead, we placed a “Summary” icon next to each title. If a searcher wanted to read the summary of a document, he/she could hover the mouse over the “Summary” icon, a pop-up window would appear next to the icon to show the summary. The content of this summary is the same as that for the same document in TDLiner interface.

We expected that the grouping mechanism in this interface would help searchers select a relevant website and a web page from the website as a starting point to browse from, and also support searchers with the third judgment task – the websites of different departments (or different sectors of the same department) would provide different perspectives on the searched topic.

2. The sitemap (Figure 3): After a searcher entered a web site, a hierarchical sitemap was provided to support the second judgment. The same query was used to retrieve the top 100 documents from just that site. The sitemap provided an outline view of the distribution of these retrieved pages according to the directory structure of the website. By using this sitemap, the searcher would be able to see the distribution of retrieved pages above or under the current directory, and to have an overview of the location of current page in the whole site.

Therefore, our hypothesis could be rephrased more specifically as that a searcher may perform the topic distillation task better with TDHierarchic interface than with TDLiner interface.

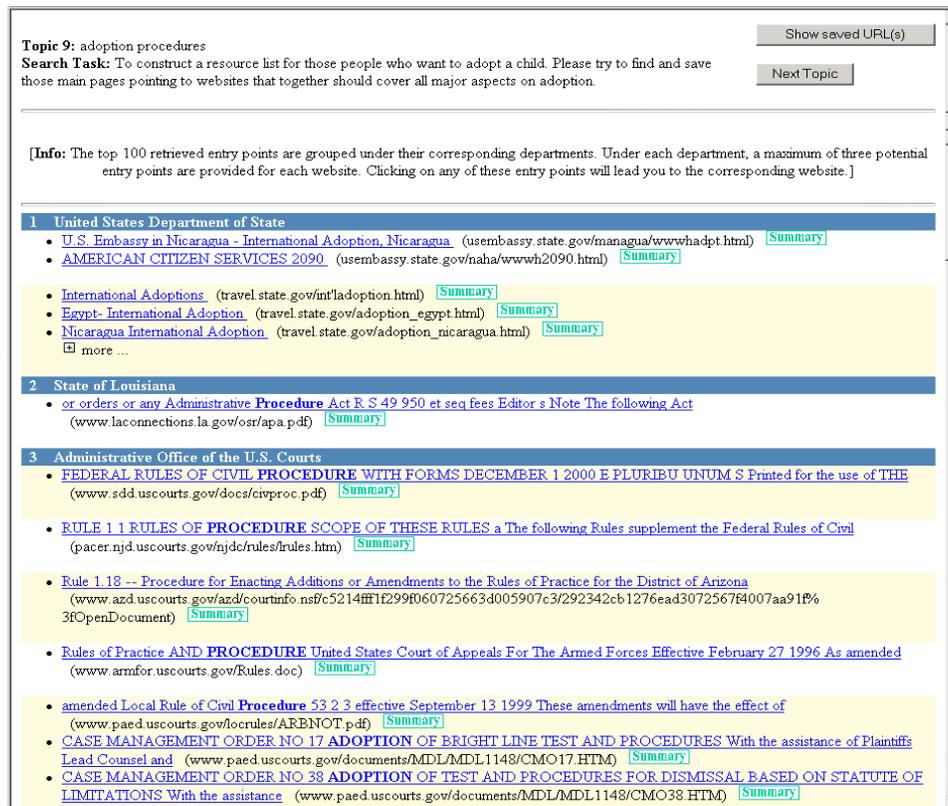


Figure 2. The site summary interface

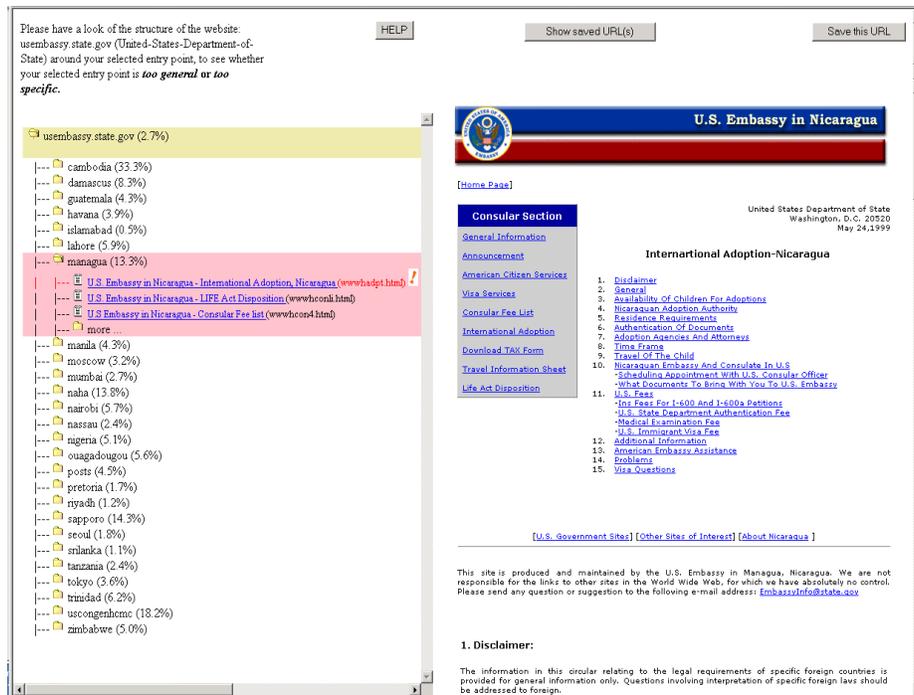


Figure 3. The sitemap interface

3.1.2. Experimental procedure

We adopted the same experimental design as used by all participating groups in the interactive track. In this experimental design, subjects searching four topics on each interface, the sequence of interface and topics varied among subjects. A complete design requires a group of 16 subjects.

During the experiment, all subjects were asked to follow the following procedure:

- Subjects filled in the pre-search questionnaire about their demographic information and their search experience.
- We explained the search task to the subjects and gave subjects an example as recommended by the track guidelines.
- After acknowledged their understanding of the search task, subjects were then presented with the two experimental interfaces, and were free to ask any question.
- Subjects were randomly given a search number. The sequence of each topic and its associated interface for each search number was pre-programmed according to the experimental design. Subjects had 15 minutes for each topic, and were prompted to move to the next topic when their times run out.
- Prior to each interface, subjects had a chance of hands-on practice with an example topic. This helps them to get familiar with the 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 were also required to fill in a post-search questionnaire about their experience of that particular search topic.
- Subjects filled in a post-system questionnaire after each interface (with four search topics).
- Subjects filled in an exit questionnaire at the end of the experiment.

3.1.3. Subjects

Sixteen students were recruited from local universities. They are all from computer science background. Among them, one is a PhD student; four of them are undergraduate students; and the rest eleven are all

Master students. They have an average age of 23.8. On average, they have 4.4 years of online searching experience, they regarded themselves as experienced searcher (Mean=5.44, Std=0.73); fifteen of them search the web daily. Comparatively, they have more experience with web search engines (Mean=6.19, Std=0.66) than the web site directory (Mean=5.56, Std=1.71).

3.1.4. Measurements

The saved lists from each search session (per topic, per interface) were gathered and sent to NIST for assessment. The assessment was based on four criteria: relevance, depth, coverage, and repetition. The assessors were asked to answer each of the following questions/statements on a five-point Likert scale.

Relevance: The page is relevant for the topic.

1 = Agree strongly, 2 = Agree slightly, 3 = Neutral, 4 = Disagree slightly, 5 = Disagree strongly

Depth: Is the page too broad, too narrow or at the right level of detail for the topic?

1 = Too broad, 2 = Bit broad, 3 = Right level, 4 = Bit narrow, 5 = Too narrow

Coverage: The set of saved entry points covers all the different aspects of the topic.

1 = Agree strongly, 2 = Agree slightly, 3 = Neutral, 4 = Disagree slightly, 5 = Disagree strongly

Repetition: How much repetition/overlap is there within the set of saved entry points?

1 = None, 2 = Minimal, 3 = Some, 4 = A lot of, 5 = Way too much

From the questions, we can see that the relevance and the depth judgment are document based, while the coverage and repetition are list based.

Transaction logging, questionnaire, and screen recording are the main methods used to collect data. During each search session, every significant event - such as reading a document and saving the URL - was automatically captured. Questionnaires common to all participating groups in the interactive track were adapted to our testing hypothesis. Screen recording was used to capture the search process for further detailed analysis.

3.2. Results

3.2.1. Performance with two interfaces

Tables 3 to 6 show each objective measure over all search sessions for each interface, averaged over topics and subjects. Overall, there is no significant difference between two interfaces (TDLinear vs TDHierarchic) by any measure, although there are topic variations.

As we discussed earlier, one motivation for this year's interactive track was to compare the results from the interactive topic distillation with that from automatic topic distillation. Thus, for each topic, we take a list of top N documents generated by Panoptic topic distillation engine, where N is the number nearest to the average size of all saved lists for that topic. For each of these lists, we can measure its relevance and depth, given that assessors had provided with corresponding assessments for each document. In the rare occasions when one of the top N documents was not picked up by any searcher as relevant, we would then assign it to the "highly irrelevant" category. For the lists automatically generated by Panoptic, their relevance and depth are shown in Tables 3 and 4 denoted as TDAuto (for Topic Distillation from Automatic system). From Table 3 and Table 4, we can find that, in six out of eight topics, the lists saved by searchers (TDLinear) are more relevant and closer to the right level than the lists from the automatic approach (TDAuto). Overall, these differences are significant ($p < 0.0003^2$ and $p < 0.0001$ for the relevance and depth respectively). The difference between TDHierarchic and TDAuto is not found significant in terms of relevance, but significant ($p < 0.005$) in terms of depth.

² All significant tests in the interactive part used paired t-test.

In the automatic topic distillation track, systems are judged according to the number of good answers they found in the top ten results. Here the “good” answers are those of high relevance and right depth. To compare the interactive system with the automatic tool using an equivalent measure, we also use the relevance and depth as the indicator of a “good” answer: if the relevance score of a saved page is 1 or 2, and the depth score of the page is between 2 and 4 inclusively, we would assume the page is a good answer. By applying this rule, the Tables 3 and 4 can be converted into the Table 7³. The difference between TDAuto and TDLinear is significant at 0.02 (paired t-test).

Table 3: Relevance of the saved/retrieved documents (The closer a score is to 1, the better)

	T1	T2	T3	T4	T5	T6	T7	T8	Mean
TDAuto	3.17	1.14	2.50	2.86	1.67	2.75	3.43	2.83	2.54
TDLinear	2.81	1.37	2.7	2.41	1.13	2.38	2.43	2.49	2.22
TDHierarchic	3.56	1.85	2.52	2.74	1.22	2.96	2.81	2.03	2.46

Table 4: Depth of the saved/retrieved documents (The closer a score is to 3, the better)

	T1	T2	T3	T4	T5	T6	T7	T8	Mean
TDAuto	4.00	3.71	2.50	4.14	3.33	3.13	4.14	3.67	3.58
TDLinear	3.77	3.19	2.26	3.83	2.99	3.40	3.62	3.46	3.32
TDHierarchic	4.30	2.88	2.47	3.83	2.87	3.37	3.71	3.01	3.31

Table 5: Coverage of the saved list (The closer a score is to 1, the better)

	T1	T2	T3	T4	T5	T6	T7	T8	Mean
TDLinear	1.63	2.13	3.25	4.5	1.25	1.25	1.00	4.88	2.48
TDHierarchic	2.38	2.63	2.50	4.63	2.25	1.25	1.00	3.63	2.53

Table 6: Repetition of the saved list (The closer a score is to 1, the better)

	T1	T2	T3	T4	T5	T6	T7	T8	Mean
TDLinear	2.57	2.25	1.75	2.38	1.50	3.0	3.0	2.63	2.38
TDHierarchic	2.50	1.63	2.13	3.38	2.00	3.25	3.25	1.25	2.42

Table 7: Precision of the saved/retrieved list

	T1	T2	T3	T4	T5	T6	T7	T8	Mean
TDAuto	0.33	1.00	0.67	0.29	0.83	0.50	0.43	0.50	0.57
TDLinear	0.48	0.88	0.53	0.52	0.97	0.62	0.54	0.48	0.63
TDHierarchic	0.18	0.78	0.61	0.48	0.95	0.51	0.49	0.60	0.58

³ Compared to the automatic topic distillation task our assessment is fairly lenient and the queries have been manually adjusted to the task. Although the absolute values of the precision are high, it is the relative differences that are noteworthy.

3.2.2. Searcher effort

The numbers of unique documents (after removing duplicated occurrences and un-accessible pages) read and saved are shown in Table 8. The second row shows that subjects read significantly more documents from TDLiner interface (Mean=24.17) than that from TDHierarchic interface (Mean=17.73) ($p < 0.0002$). However, the third row did not show much difference in the number of saved documents from each interface.

Table 8: The number of read and saved documents

	TDLiner Mean (Std)	TDHierarchic Mean (Std)
Read-unique	24.17 (8.71)	17.73 (5.00)
Saved-unique	6.64 (3.00)	6.65 (3.79)

To understand how and where subjects put their effort, we had a closer look on how subjects divided their effort on each interface.

The TDLiner interface has two parts: the window for the ranked list (TDLiner-R) and the window to show the content of a selected document (TDLiner-C). The TDHierarchic interface has three parts: the window for the grouped ranked list (TDHierarchic-R), the frame for the tree structure of a selected web site (TDHierarchic-T), and the frame to show the content of a selected document (TDHierarchic-C).

There is not much difference between TDLiner-C and TDHierarchic-C, except their window sizes. The difference is that TDHierarchic has an extra interface panel (TDHierarchic-T), and TDHierarchic-R is probably more complex than TDLiner-R.

Table 9 shows the division of effort from the first four searchers. By examining the recorded screen actions from the these four searchers, we observed that these searchers spent an average 36% of their total search time and on average opened 15 (unique) documents to read from TDLiner-R. While in TDHierarchic-G window, searchers spent similar amount of time (37% of their total search time), but opened only 9.3 (unique) documents. We observed that searchers picked up documents to open sequentially and spent less time to read document summaries in TDLiner-R, while they spent more time to read document summaries (by hovering the mouse over the “Summary” icon) and even read summaries from a few documents before they opened a document in TDHierarchic-G.

While these four searchers spent on average 64% of their total search time and opened 7.9 documents to read from TDLiner-C, they divided their effort in two frames in TDHierarchic. These four searchers spent on average 19% of their search time on TDHierarchic-T, 44% on TDHierarchic-C, but opened a similar number of (unique) documents. This implies that the searchers used the tree structure more often to help them to browse the selected web site.

Table 9: The split of efforts in each interface

	% of total time		
TDLiner	Ranked list: 36%	Page content window: 64%	
TDHierarchic	Grouping: 37%	Tree: 19%	Page content window: 44%
	Average number of documents opened		
TDLiner	Ranked list: 15	Page content window: 7.9	
TDHierarchic	Grouping: 9.3	Tree: 4.4	Page content window: 4.2

3.2.3. Subjective measures

After each topic, subjects were required to fill in a post-search questionnaire about their experience of the search topic and their sense of the task completeness. All questions are on 7-point Likert scale with 1=strongly disagree, 4=neutral, and 7=strongly agree. Table 10 shows that subjects gave higher score to TDHierarchic interface on all seven questions.

Table 10: Post-search questionnaire

	Q1	Q2	Q3	Q4	Q5	Q6	Q7
TDLinear	5.02	5.04	4.55	5.03	5.16	4.50	4.82
TDHierarchic	5.25	5.13	4.82	5.02	5.23	5.05	4.91

Q1: The search process is easy.

Q2: The pages I just saved focuses on the topic well.

Q3: The pages I just saved are the main pages of their corresponding websites.

Q4: The pages I just saved together provide a good coverage of the topic.

Q5: The pages I just saved will be helpful for the targeted audiences.

Q6: I have enough time to do an effective search.

Q7: I believe that I have succeeded in my performance of the task.

After each system, subjects were asked to fill in a post-system questionnaire to get their opinion on the usability of each system. Table 11 shows the average score for each interface for seven questions. There are significant difference between two interfaces for question 3 and question 4, that is: searchers strongly agreed that the organization of the search results of TDHierarchic interface was clearer and more useful for them to select an entry point to start with.

Table 11: Post-system questionnaire

	Q1	Q2	Q3	Q4	Q5	Q6	Q7
TDLinear	5.50	5.38	4.25	3.81	3.94	3.81	3.81
TDHierarchic	5.63	5.88	5.81	5.50	4.94	4.50	4.88

Q1: It was easy to learn to use this system.

Q2: It was easy to use this system.

Q3: The organization of the search results is clear to me.

Q4: the organization of the search results is useful for me to select an entry point to search.

Q5: The summary of each search results helped me to decide the relevance of that website.

Q6: The summary of each search result is useful for me to select an entry point to search.

Q7: The web structure of my selected entry point is useful for me to judge whether the entry point is the main page.

Table 12 shows searchers' answer to the three questions in the exit search questionnaire. Overall, most of the searchers perceived that TDHierarchic interface is easier to use and supporting their task better, and they liked TDHierarchic interface the best overall.

Table 12. Exit questionnaire

	Q1	Q2	Q3
TDLinear	6	2	3
TDHierarchic	10	14	11
No Difference	0	0	2

Q1: Which of the two systems did you find easier to use?

Q2: Which of the two systems did you think supporting your task better?

Q3: Which of the two systems did you like the best overall?

3.3. Discussion

In this experiment, we found that our searchers preferred the experimental interface (TDHierarchic) and perceived that they fulfilled their task better by using the experimental interface than the ranked list interface (TDLinear). However, we didn't find any significant difference between the two interfaces on searchers' performance in terms of relevance, depth, coverage and repetition.

One of our hypotheses was that we could increase performance by encouraging searchers to compare items rather than make individual judgments. This was implemented on the site summary interface. By further examining searchers' behavior, we found that the interface for grouping documents into sites changed search behavior: searchers spent time selecting amongst the results from a specific site by looking at and comparing the summaries. Searchers selected fewer pages to examine and the overall results were similar to the ranked list interface indicating that users had compared and made good selection decisions. Also in the post-system questionnaire, searchers stated strongly that the grouping interface was useful for them to select an entry point to search. However, confounded by many other factors, it is not clear whether this behavior would be beneficial to the overall task.

Comparing the results from our interactive system with that of the corresponding automatic system, we found a significant improvement in terms of relevance, depth and precision. That indicates that engagement of a searcher's effort has a positive effect on the system performance, and that there is room for improvement for systems to reduce searcher effort.

4. References

- [1] Web track guidelines: http://es.cmis.csiro.au/TRECWeb/guidelines_2003.html
- [2] Interactive track guidelines: <http://www.ted.cmis.csiro.au/TRECInt/guidelines.html>