Melbourne TREC-9 Experiments

Daryl D'Souza Michael Fuller
James Thom Phil Vines Justin Zobel
Department of Computer Science, RMIT University
GPO Box 2476V, Melbourne Victoria Australia 3001
{*djds, msf, jat, phil, jz*}*@cs.rmit.edu.au*

Owen de Kretser
Department of Computer Science and Software Engineering
The University of Melbourne, Victoria 3010, Australia
oldk@cs.mu.oz.au

Ross Wilkinson Mingfang Wu
CSIRO, Division of Mathematics and Information Science
723 Swanston St., Carlton VIC 3053
{*Ross.Wilkinson, MingFang.Wu*}*@cs.mu.oz.au*

February 1, 2001

1 Introduction

We report results for experiments conducted in Melbourne—at CSIRO, RMIT, and The University of Melbourne—for TREC-9. We present results for the interactive track, cross-lingual track, main web track, and the query track.

2 Interactive Track

2.1 Introduction

We have been continuously investigating technologies for delivering retrieved documents to support interactive question answering. In this year's interactive track, we focused on the role of a document surrogate in the interactive fact finding task. In this experiment, we compared two types of document surrogates

Submitted to the 9th Text Retrieval Conference (TREC-9), Gaithersburg, MD, USA, November 13–16, 2000.

in the two experimental systems. One system uses the document title and the first 20 words of a document as the document's surrogate, while the other system uses the document title and the best three Answer Indicative Sentences extracted from the document as the document's surrogate. The results show that subjects can find significantly more facts from the system using 3 sentences than from the other system.

2.2 Hypothesis

This year's track reflects two of the major characteristics of interactive information searching: the questions are concentrated on the fact finding, and the time for answering each question is very short (5 minutes). As an average reader can only scan a limited number of words within 5 minutes, the challenge is how to help the user to locate the facts or find the documents that may contain the facts while reading the limited number of words.

For most web search engine (e.g. Altavista, Excite), this has been achieved by displaying the surrogate of a document, which mainly includes the title and the first *N* words from the document. The purpose of the surrogate is to indicate the main theme of the document. This kind of surrogate may be more suitable for the learning and exploration types of information needs, but less suitable for fact finding type of information need. Based on our pilot investigation into the interactive fact finding task, we observed that:

- The relevant facts may exist within a small chunk of documents, and this small chunk may be not necessarily related to the main theme of the document.
- This small chunk usually contains the keywords, and is in the form of a complete sentence. We call this sentence the answer indicative sentence (AIS).
- When a user is scanning through a document to search for facts, s/he usually tries to locate an answer indicative sentence by looking around the query keywords, and therefore either find the facts, or decide whether to read the document further or discard it.

Our hypothesis is that the above-mentioned answer indicative sentences should provide a better surrogate of the document than the first *N* words, for the purpose of interactive fact finding. Therefore our experiment focused on the comparison and evaluation of two systems using different surrogates. The control system *First20* uses the title and the first twenty words as the surrogate of a document, and the test system *AIS3* uses the title and best three answer indicative sentences as the surrogate of a document. The performance was measured by the effectiveness of each system in helping to locate answer facts, users' subjective perception of the systems, and the effort required by users to locate answers.

2.3 System Description

Both systems in this experiment use the mg [4] search engine for indexing and retrieval. The two systems provide natural language querying [2] only. For each query, both systems present a user with the surrogates of the top 100 retrieved documents in 5 consecutive pages, with each page containing 20 document surrogates. Each system has a main window for showing these surrogate pages. A document reading window is popped up when a document surrogate is clicked. If a user finds a fact from the document reading window, s/he can click the "Save Answer" button in this window and a save window will be popped up for the user to input the newly found fact or modify previously saved facts.

The difference between two systems is what is presented on their main windows. The main window of the control system (*First20*) is shown in Figure 1. This kind of presentation is quite similar to those web search engines such as Altavista and Excite. The main window of the test system (*AIS3*) is shown in Figure 2. Roughly, the number of words on each page of an *AIS3* window is three times that of the *First20* window. Also, there is a *save* icon next to each answer indicative sentence, with the same function as the "Save Answer" button in the document reading window. If a user finds a fact from the sentence, s/he can save the fact directly by clicking this icon.

The three best AIS are dynamically generated after each query search according to the following procedure:

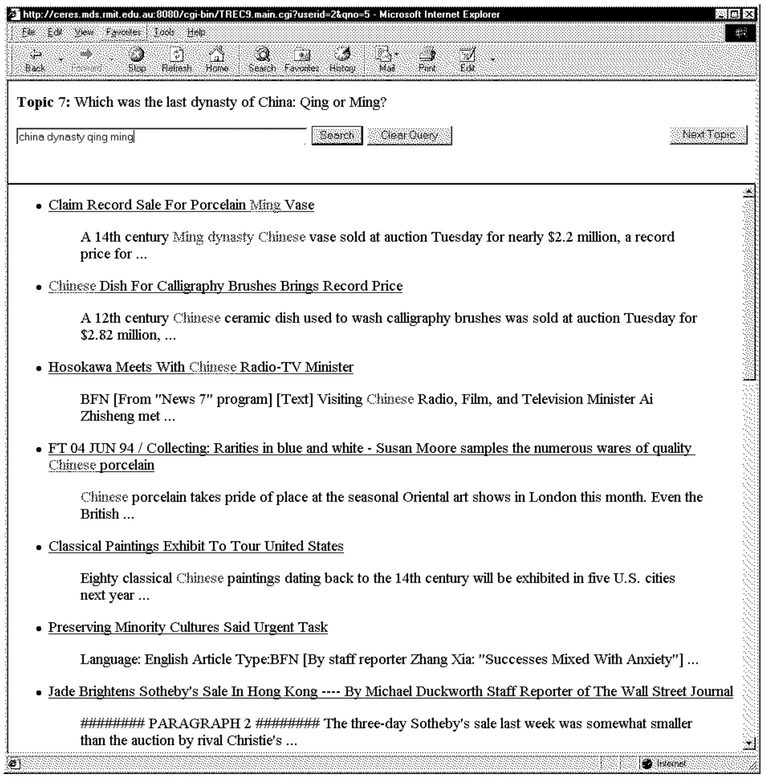
- An AIS should contain at least one query word and be at least ten words long.
- The AISs are first ranked according to the number of unique query words contained in each AIS. If two AISs have the same number of unique query words, they will be ranked according to order in which they occur within the document.
- The top three AIS are then selected.

2.4 Experiment

2.4.1 Procedure

Each subject searched eight topics according to the TREC-9 Interactive Track experimental guidelines, with four topics on each system. During the experiment, the subject followed the following steps:

- Reading the introduction to the experiment.
- Filling in the Pre-Search Questionnaire.
- Demonstration of main functions of each system.
- Hands on practice with both systems.


http://ceres.mds.rmit.edu.au:8080/cgi-bin/TREC9.main.cgi?userid=2&qno=5 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Search
Clear Query
Next Topic
Internet

Topic 7: Which was the last dynasty of China: Qing or Ming?

china dynasty qing ming

- Claim Record Sale For Porcelain Ming Vase

  A 14th century Ming dynasty Chinese vase sold at auction Tuesday for nearly $2.2 million, a record price for ...

- Chinese Dish For Calligraphy Brushes Brings Record Price

  A 12th century Chinese ceramic dish used to wash calligraphy brushes was sold at auction Tuesday for $2.82 million, ...

- Hosokawa Meets With Chinese Radio-TV Minister

  BFN [From "News 7" program] [Text] Visiting Chinese Radio, Film, and Television Minister Ai Zhisheng met ...

- FT 04 JUN 94 / Collecting: Rarities in blue and white - Susan Moore samples the numerous wares of quality Chinese porcelain

  Chinese porcelain takes pride of place at the seasonal Oriental art shows in London this month. Even the British ...

- Classical Paintings Exhibit To Tour United States

  Eighty classical Chinese paintings dating back to the 14th century will be exhibited in five U.S. cities next year ...

- Preserving Minority Cultures Said Urgent Task

  Language: English Article Type:BFN [By staff reporter Zhang Xia: "Successes Mixed With Anxiety"] ...

- Jade Brightens Sotheby's Sale In Hong Kong ---- By Michael Duckworth Staff Reporter of The Wall Street Journal

  ######## PARAGRAPH 2 ######## The three-day Sotheby's sale last week was somewhat smaller than the auction by rival Christie's ...

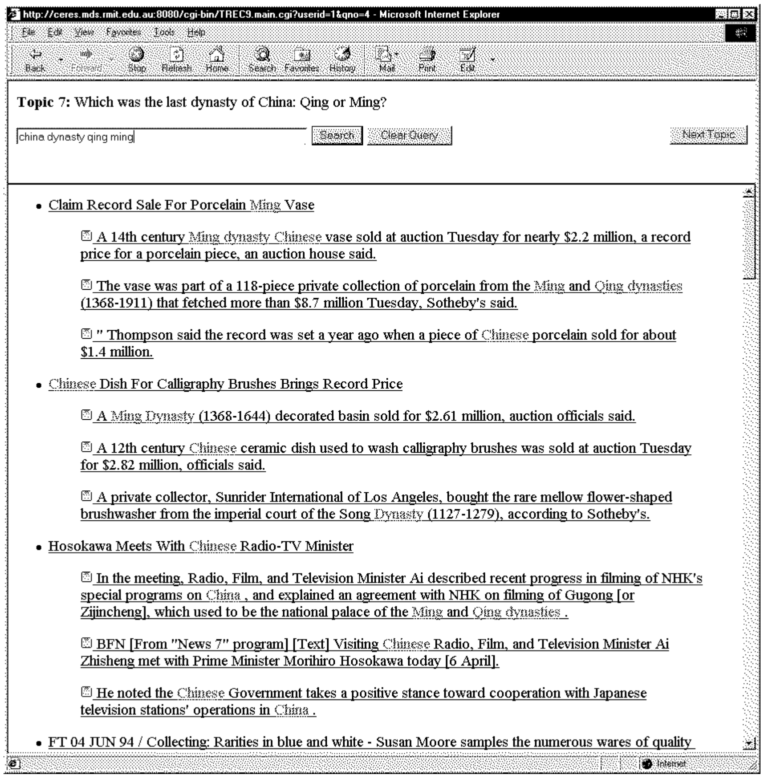
Figure 1: The main window of the *First20* system.

- Search four topics on each system with Pre-Search questionnaire and Post-search questionnaire per topic, and a Post-System questionnaire per system.
- Filling in exit questionnaire.

It takes about 1.5 hours for a subject to finish the whole procedure.

2.4.2 Subjects

Sixteen paid subjects were recruited via an RMIT internal university newsgroup. There are five females and eleven males. The average age of sixteen subjects is 23, with the youngest 19 and oldest 39. They have 5.1 years online search experience on the average. All subjects are the students from the Department of Computer Science, nine of them are undergraduate students, the other seven subjects already had a Bachelor degree and are studying for a higher degree (3 on graduate diploma and 4 on master degree).


http://ceres.mds.rmit.edu.au:8080/cgi-bin/TREC9.main.cgi?userid=1&qno=4 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Search
Clear Query
Next Topic
Internet

Topic 7: Which was the last dynasty of China: Qing or Ming?

china dynasty qing ming

- Claim Record Sale For Porcelain Ming Vase

  ☐ A 14th century Ming dynasty Chinese vase sold at auction Tuesday for nearly $2.2 million, a record price for a porcelain piece, an auction house said.

  ☐ The vase was part of a 118-piece private collection of porcelain from the Ming and Qing dynasties (1368-1911) that fetched more than $8.7 million Tuesday, Sotheby's said.

  ☐ " Thompson said the record was set a year ago when a piece of Chinese porcelain sold for about $1.4 million.

- Chinese Dish For Calligraphy Brushes Brings Record Price

  ☐ A Ming Dynasty (1368-1644) decorated basin sold for $2.61 million, auction officials said.

  ☐ A 12th century Chinese ceramic dish used to wash calligraphy brushes was sold at auction Tuesday for $2.82 million, officials said.

  ☐ A private collector, Sunrider International of Los Angeles, bought the rare mellow flower-shaped brushwasher from the imperial court of the Song Dynasty (1127-1279), according to Sotheby's.

- Hosokawa Meets With Chinese Radio-TV Minister

  ☐ In the meeting, Radio, Film, and Television Minister Ai described recent progress in filming of NHK's special programs on China , and explained an agreement with NHK on filming of Gugong [or Zijincheng], which used to be the national palace of the Ming and Qing dynasties .

  ☐ BFN [From "News 7" program] [Text] Visiting Chinese Radio, Film, and Television Minister Ai Zhisheng met with Prime Minister Morihiro Hosokawa today [6 April].

  ☐ He noted the Chinese Government takes a positive stance toward cooperation with Japanese television stations' operations in China .

- FT 04 JUN 94 / Collecting: Rarities in blue and white - Susan Moore samples the numerous wares of quality

Figure 2: The main window of the *AIS3* system.

2.4.3 Data Collection Methods

Transaction logging and questionnaires were used to collect data. During the experiment, every significant event—such as documents read, facts saved and their supporting documents, and queries sent—were logged and time-stamped automatically. The questionnaires used were the standard questionnaires used by participants in the Interactive Track.

2.5 Evaluation

2.5.1 System

Effectiveness

The effectiveness of the each system is evaluated by the number and the quality of the saved answers. There are two types of topics in this year's interactive track. Type 1 topics are of the form "find any n Xs". Type 2 topics are of the form "compare two specific Xs". For the Type 1 topics (topic 1-4), a complete answer consists of n facts. For the Type 2 topics (topic 5-8), two facts are usually needed to make the comparison. We observed that only for topic 7 (Type 2), the answer may sometimes be supported by only one fact.

The saved facts and the saved documents were sent to the NIST for judgement. For each search session, two judgements were made: whether the subject found the required number of facts (for topics of Type 1) or whether the subject answered the question correctly (for topics of Type 2); and whether the saved facts (or answers) are supported by the saved documents. Both judgements have three scores: all, some, or none.

A fully successful session is defined as whether the question is fully answered and whether the answer is fully supported (i.e. the both judgments are "all"). If we give a score of '1' to such a successful session and a score of '0' to any other sessions, then there are 14 successful sessions in total for users of *First20* and 27 successful sessions in total for users of *AIS3*. The difference between the two systems is significant at level 0.01 (two tailed t-test).

Table 1 shows the successful sessions topic by topic. We can see that users of *AIS3* has more successful sessions for all topics except the Topic 3.

Topic	1	2	3	4	5	6	7	8	Total
First20	0	0	0	2	3	3	5	1	14
AIS3	1	2	0	3	4	7	8	2	27

Table 1: The number of the fully successful sessions per topic.

Table 2 shows the number of the fully successful sessions subject by subject. We see that of the sixteen subjects, ten subjects had more successful sessions when using *AIS3* than when using *First20*; only two subjects had more successful sessions when using *First20* than when using *AIS3*.

Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
First20	0	0	1	1	1	2	1	3	2	0	0	1	1	0	0	1
AIS3	2	2	1	1	3	2	2	1	1	1	1	1	2	3	2	2

Table 2: The number of the fully successful sessions per subject.

Although the subjects may not get the full answer in some sessions, the subjects sometimes demonstrated the ability to find a partial answer. We need not simply classify these sessions as failure, but instead may consider them as partially successful sessions. Therefore, we can award an adjusted score in the range $[0, 1]$. For each topic, each fact that is correctly identified and supported by a document contributes $1/n$ toward the score, where n is the number of required facts for the topic. Overall, *AIS3* gets score 0.65 and *First20* 0.47; the difference between the two systems based on the adjusted score is also statistically significant, at level 0.03 (two tailed t-test).

Table 3 shows the average score across subjects per topic for each system. *AIS3* has the higher score than *First20* for all topics except for the topic 5.

Topic	1	2	3	4	5	6	7	8	Mean
First20	0.08	0.00	0.38	0.71	0.81	0.75	0.75	0.31	0.47
AIS3	0.46	0.25	0.47	0.79	0.75	0.94	1.0	0.56	0.65

Table 3: The comparison of two systems topic by topic based on adjusted scores.

Based on the above results, the hypothesis that the AIS is better document surrogate than the *First20* for fact finding task is supported.

Perception of the system

The Subjects' perception of the systems is captured from three questions in exit questionnaire. The three questions are:

- Question 1: Which of the two systems did you find easier to learn to use?
- Question 2: Which of the two systems did you find easier to use?
- Question 3: Which of the two systems did you like the best overall?

The distribution of subjects' choice is shown in Table 4. We can see that: for question 1, 17% subjects selected *First20* while 50% subjects selected *AIS3*. For question 2, 25% subjects selected *First20* while 69% subjects selected *AIS3*. For question 3, 31% subjects selected *First20* while the other 69% selected *AIS3*. This suggests that the subjects preferred the *AIS3* system.

Questions 1 and 2 were also asked in the post system questionnaire. However instead of asking subjects to compare the two systems, the subjects were asked to judge the systems independently on a 5-point Likert scale. There is not much difference between the two systems on learning effort (*First20*: Mean = 4.0, *AIS3*: Mean = 4.1). The difference between the two systems on user perception of usefulness is statistically significant (*First20*: Mean = 3.5, *AIS3*: Mean = 4.1. two tailed, paired t-test $p < 0.03$). The result for "easy" was unexpected: we thought that the main window of *First20* was simpler than that of *AIS3*, thus would be easier to use than *AIS3*. The subject's selection and judgement may have been influenced by how well they felt they had completed their tasks.

	Easier to learn	Easier to use	Liked the best overall
First20	3	4	5
AIS3	8	11	11
No difference	5	1	

Table 4: Subjective comparison of two systems.

2.5.2 User Effort

The effort required of subjects to determine answers for each topic can be measured in terms of the number of documents they read, the number of title pages viewed, and the number of queries sent for each topic. On the average, the subjects read fewer documents and fewer title pages, and sent fewer queries from *AIS3* than from *First20*, as shown in Table 5. The difference is statistically significant at level 0.01, 0.001, and 0.02 respectively (two tailed t-test). This may not necessary mean that the users of *AIS3* took less effort than the users of *First20*, as the main page of *AIS3* displayed more text than that in *First20*. However, this may indicate that the extracted answer indicative sentences of *AIS3* may have helped subjects to find the answer or find the documents where the answer may be found.

	First20	*AIS3*
	Mean(SD)	Mean(SD)
Number of documents read	3.42(1.22)	2.66(0.77)
Number of pages viewed	2.80(1.64)	1.98(0.97)
Number of unique queries sent	2.14(0.56)	1.73(0.57)
Number of terms per query	3.25	3.26

Table 5: Subject's interaction with the systems.

2.5.3 Perception of the Topics

Before each search, subjects were asked about their familiarity about the topic. As show in the Table 6, overall, subjects have low familiarity with all topics (all under 3 on a 5-point Likert scale). Of eight topics, Topic 7 had the highest familiarity. Nine subjects claimed that they knew the answer before the search, but four of these subjects were wrong. After the search, three of these four subjects got the right answer.

Topic	1	2	3	4	5	6	7	8
First20	1.75	1.25	1.63	1.38	1.38	1.25	2.63	1.38
AIS3	1.25	1.75	1.50	1.25	1.75	1.00	2.00	1.50
All	1.50	1.50	1.56	1.31	1.56	1.13	2.31	1.44

Table 6: Average score of subject's familarity with each topic.

After each topic, subjects were asked about their satisfaction with the search results and certainty about the answer. Generally, the users of *AIS3* had higher satisfaction and certainty (satisfaction: *First20* Mean = 3.16, *AIS3* = 3.56; certainty: ¡S-Del¿*First30* Mean = 3.50, *AIS3* = 3.89), but these differences are not significant.

There is no significant correlation found between the familiarity and the number of successful sessions, the satisfaction, or the certainty.

Most tested topics were very clear to the subjects. There are two topics which had different interpretations among subjects. One is the Topic 1: "What are the names of three US national parks where one can find redwoods?" — many subjects saved state parks. Another is Topic 2: "Identify a site with Roman ruins in present day France?" — some subjects were not certain about the area scope of the site. Some subjects said they did find Southern France, but they did not think that could be counted as an answer, instead trying to find the specific name of the site.

2.5.4 Subjects Difference

It is interesting that our subjects fall into two groups: one whose first language is English, and another whose second language is English and their first language varies. The subjects of the latter type are all international students from Asia. The native language group has 7 subjects while the foreign language group 9 subjects.

We break the subjects into two groups and summarise the data accordingly for each system based on the adjusted scores, the result is as shown in Table 7. No difference is detected between two groups of each system. This may indicate that the language and culture background are unlikely to have influenced subjects' performance for the tested topics.

	Native	Foreign
First20	0.46	0.49
AIS3	0.66	0.64

Table 7: The comparison of two groups (native language and foreign language) based on the adjusted score.

2.6 Discussion

The experiment investigated the role of document surrogate in the interactive fact finding task. The experiment results show that using an AIS3 as a document surrogate is significantly better than the First20 in helping users locate relevant documents and thus find more relevant facts. Subjects also more preferred the *AIS3* system than the *First20* system.

Although our hypothesis has been supported in the experiment, we understand that more topics and wider variety of document collections will need to be tested to further validate the hypothesis.

3 Cross-Lingual Track

This year we participated in the Chinese–English cross-lingual track, drawing on our experience from our involvement in the Chinese track several years ago. We used two approaches to convert the problem to one of monolingual retrieval. First, we tested converting the English language queries to Chinese (run *rmitcl002*), and second, we tested converting the Chinese document collection to English (run *rmitcl003*). In both approaches we made use of the online dictionaries that were made available.

The translations were on a word by word basis. For the English-to-Chinese translation, if a word that contained uppercase letters was not in the dictionary, we converted it to lower case and tried again. The reason for this is that some proper nouns appear in the dictionary with a capitalised first letter, however for words at the start of a sentence it is more appropriate to convert to lower case.

We also tested combination of evidence (*rmitcl001*), combining the results of the two previous runs based on normalised similarity values, that is, $sim_{new} = 0.5 \times sim'_1 + 0.5 \times sim'_2$, where sim'_1, and sim'_2 are the normalised similarity measures from two runs above. We also included a monolingual run as a baseline.

Our monolingual run results were somewhat lower than the median. We are not completely sure why this is the case but suspect it was partly because the run was a straight processing of the data with no special treatment, and because character indexing rather than word indexing was used. Unfortunately the cross language runs produced random results; there is obviously a problem with the software which we are working to resolve.

4 Main Web Track

Four runs were submitted, labelled `rmitWFGweb`, `rmitWFLweb`, `rmitNFGweb` and `rmitNFLweb`. These correspond to two categories of indexes and, in each case, to two filtering protocols. The index categories were *global* (G) and *local* (L), both based on the wt10g corpus. The global index centrally-indexed all documents; the local indexes were based on five, separate subsets of the data source, as per distribution across 5 wt10g CDs. Each of the two index cases were further classified according to the filtering protocols, *no filter* (NF) and *with filter* (WF). Thus, `rmitWFGweb` refers to the filter-based, global index run.

This section is structured as follows. Section 4.1 presents the similarity ranking formulation to score and subsequently rank the documents. Details of the two filtering protocols are presented in Section 4.2. After indexing Title-only fields of TREC topics 451 to 500 were used in the querying process. Manual queries were used, as discussed in Section 4.3. We used tools from the `mg` system [4] to construct and query indexes. Document sources were stopped and stemmed during the indexing process, and so too were queries, prior to submitting them for ranking of documents. Retrieval effectiveness results for various runs are presented in Section 4.4.

4.1 Relevance scoring method

The combining function used to establish similarity of documents and queries was the standard Cosine measure [5], where similarity, $S_{q,d}$ between document d and query q is given by:

$$S_{q,d} = \frac{\sum_{t \in q \cap d}(w_{q,t} \cdot w_{d,t})}{W_q \cdot W_d}$$

where the document-term and query-term weights are computed, respectively, as:

$$w_{d,t} = log_2(f_{d,t} + 1)$$

$$w_{q,t} = log_2(f_{q,t} + 1) \cdot log_2 \left(\frac{N}{f_t} + 1 \right)$$

The terms $f_{x,t}$ $(x = q|d)$, f_t and N are, respectively, frequency of term t in x, number of documents containing t, and the total number of documents. Finally, W_x is given by

$$W_x = \sqrt{\sum_{t=1}^{n} w_{x,t}^2}$$

for the n terms in the vocabulary.

4.2 Filtering versus non-filtering

We used two term extraction protocols in the indexing process. For the NF cases the default term extraction policy used by the indexing tool was used. Words are extracted as follows:

- A word is a string of alphanumeric characters delimited by non alphanumeric or space symbols.
- Long digit strings are truncated at every fourth byte, until a non-digit is encountered. Each truncated portion constitutes a word, including the residue, if any.
- Words in tags are ignored.

In the WF cases, data sources were subjected to a filtering process prior to indexing:

- A word is a string of alphanumeric characters delimited by non alphanumeric or space symbols; a word must begin with a letter and may not have more than two digits.
- Words from a long, non-space-delimited string are not extracted beyond the tenth character in the original string.
- Words in tags are ignored, except words inside HTML comments.

Topic-id	Word(s) before change	Word(s) after change
455	`whan`	`when`
463	`tartin`	`tartan`
464	`nativityscenes`	`nativity scenes`
474	`bennefits`	`benefits`
475	`compostion`	`composition`
477	`Carribean`	`Caribbean`
483	`rosebowl`	`rose bowl`
487	`angioplast7`	`angioplasty`

Table 8: *Summary of amendments TREC Topics 451-500.*

4.3 Queries

Title-only fields of TREC Topics 451 to 500 were used. These were manually amended to ensure that at least one document was ranked for each run and query, and to correct spelling inconsistencies between query terms in Title and other fields. Table 8 summarises these changes; it presents the part of the query (prior to stopping and stemming) that was modified.

Note that amendments in the first and last table entries are inconsequential. Prior to indexing `when` is removed because it is a stop word and `whan` is unlikely to appear in the document text. Similarly, `angioplasty` is stemmed to `angioplast`; leaving `angioplast7` as is would cause both indexing protocols to index the term `angioplast`.

4.4 Results

Unfortunately, after submission, the WF result runs were identified as being flawed, due to an erroneous word filter. Nevertheless, corrected runs, while depicting improved performances, revealed that word filtering did not improve performance of the WF cases over the NF cases. The filtering process was motivated by the rationale that removal of URL references and inclusion of words in comments would improve overall performance; this was not the case.

Figures 3 and 4 present the Recall-Precision performances of NF versus WF for the global and local scenarios, respectively. The three-way relevance judgements (not relevant, relevant, highly relevant) were altered to reflect a binary relevance (relevant, not relevant) by re-codifying *highly relevant* documents as *relevant*.

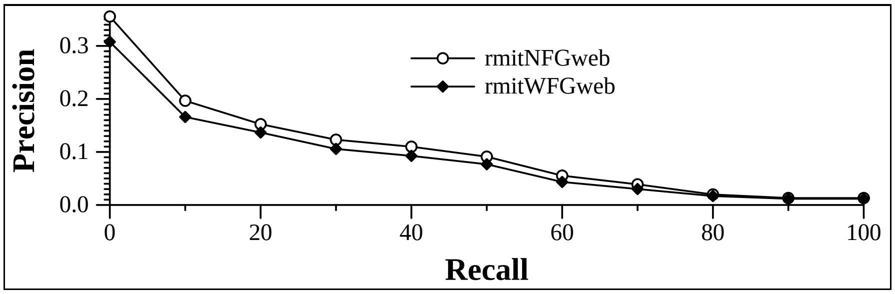


Figure 3: *Main web track: Retrieval P-R performances of global indexes.*

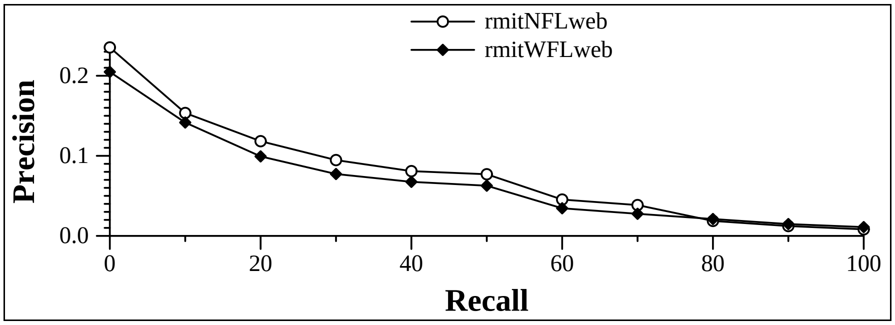


Figure 4: *Main web track: Retrieval P-R performances of local indexes.*

5 Query Track

5.1 Stage 1: Query Variations

Three variations of the TREC Topics 51–100 were manually created:

- `UoM1a`: A 2–3 word query based on the topic statement.
- `UoM1b`: Another 2–3 word variation based on the topic statement.
- `UoM2`: A sentence based on the topic and relevance judgements.

All query variations were created by the same person and roughly 2–3 minutes were spent on each topic for each variation.

5.2 Stage 2: Retrieval Variations

There were 43 different query sets made available by the participants. Prior to retrieval runs, each topic of each query variation had stopwords removed. Two different retrieval systems were used, each based on the full-text retrieval system `mg` [4].

The first system used was the standard document-based version of mg using the following vector-space similarity measure with a normalised-by-maximum-frequency variant of the query-term weights:

$$S_{q,d} = \frac{\sum_{t \in \tau_{q,d}} (w_{q,t} \cdot w_{d,t})}{W'_d} \tag{1}$$

where

$$\begin{aligned} w_{d,t} &= \log_e(f_{d,t}) + 1 \ , & (2) \\ w_{q,t} &= \log_e(\frac{f^m}{f_t} + 1) \cdot (\log_e(f_{q,t}) + 1) \ , \text{ and} & (3) \\ W'_d &= (1 - s) + s \cdot \frac{W_d}{W_{av}} \ . & (4) \end{aligned}$$

Document length normalisation is by *pivoted cosine normalisation* [3] where W_{av} is the average W_d over all d, and s, the slope of the pivoted cosine normalisation function, is taken to be 0.7. Using the Q-expression notation developed by Zobel and Moffat, this formulation is expressed as BD-ACI-BCA [5].

The second system employed was a locality-based version of mg in which term locality is used as a guide to relevance [1]. This run employed the *arc* shape formulation and the *logarithmic* height formulation.

A total of 86 retrieval runs were submitted (43 query sets * 2 retrieval runs). For both the document-based and locality-based versions of mg, the query variations Sab3a, Sab1c and Sab1b were most effective in terms of average precision, precision at 20 documents and reciprocal rank of first relevant document.

Acknowledgements

This work was supported by the Australian Research Council.

References

[1] O. de Kretser and A. Moffat. Effective document presentation with a locality-based similarity heuristic. In Marti Hearst, Fredric Gey, and Richard Tong, editors, *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval*, pages 113–120, University of California, Berkeley, U.S.A., August 1999. ACM.

[2] Gerard Salton. *Automatic Text Processing.* Addison-Wesley, Reading, Massachusetts, 1989.

[3] A. Singhal, G. Salton, M. Mitra, and C. Buckley. Document length normalization. *Information Processing & Management*, 32(5):619–633, 1996.

[4] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and indexing documents and images.* Van Nostrand Reinhold, New York, 1994.

[5] Justin Zobel and Alistair Moffat. Exploring the similarity space. *ACM SIGIR Forum*, 32(1):18–34, Spring 1998.