An Intranet Reality Check For TREC Ad Hoc *

David Hawking[a], Peter Bailey[b] and Nick Craswell[b]

[a]CSIRO Mathematical and Information Sciences
Canberra, Australia
David.Hawking@cmis.csiro.au

[b]Department Of Computer Science
Australian National University, Canberra, Australia
{Peter.Bailey, Nick Craswell}@cs.anu.edu.au

Abstract

A text retrieval system which has performed well on TREC ad hoc has been modified to provide the production search facility on a university intranet. The applicability of the TREC model to the intranet environment is assessed and search engine design decisions are evaluated by analysis of query logs in the light of insider knowledge of the intranet. It is noted that successful processing of very short (average 2.3 word) queries may hinge on minor lexical differences. Accordingly, indexing should preserve all useful information and operations such as stemming, should, when advantageous, be applied at query processing time. Heuristics and organisational thesauri may be needed to identify the "best" response to a query, rather than a set of "relevant" answers. Duplicate elimination and departmental search are important. Special techniques are needed to process names of people and departments and to deal with frequent misspelling of important query words.

Keywords: Intranet search engines, text retrieval, TREC, evaluation, theory into practice, query logs.

*The authors wish to acknowledge that this work was carried out within the Cooperative Research Centre for Advanced Computational Systems established under the Australian Government's Cooperative Research Centres Program.

1 Introduction

Test collection methodology, best exemplified by the TREC ad hoc task [14], is essentially a laboratory simulation of a "natural" information retrieval environment. Competition among the text retrieval "species" within the TREC simulated environment has led to evolutionary selection for features such as term-weighting schemes, stemming, stopword elimination, case folding, phrase identification, passage retrieval and pseudo-relevance feedback. Buckley et al [2, p. 117, Table 16] show that, over the first five TRECs, the best systems doubled their performance on the primary competitive dimension (average precision).

The benefits of reproducible, collective, independent, blind evaluation offered by TREC ad hoc are indisputable. However:

- TREC ad hoc simulates only a tiny portion of current retrieval activity. In essence, it models the information need of a journalist or college student researching an article or term paper. That need is expressed in a form suitable to be given to a librarian or research assistant. In reality, people seek many different types of information for many different reasons and almost never write verbose descritpions of their needs.
- The assessment of retrieved documents in TREC grossly simplifies the real world situation. In TREC ad hoc, each judged document is assessed independently of all others and labelled either as "relevant" or "not relevant". A retrieval result list in which all documents repeat or paraphrase the same marginally relevant content will score highly on TREC measures but rate very poorly with searchers. In real search, searchers are often seeking a single item that they know or suspect exists. This type of search has been modelled in some TREC tracks but not in ad hoc.
- TREC ad hoc collections cannot accurately represent all of the document sets over which search services are currently offered. TREC ad hoc collections comprise an artificial amalgam of newspaper reports and government data from a handful of disparate sources and do not include exploitable citation links. Certain sub-collections include structural information and author, title and subject metadata but topics do not encourage its use[1].

Our aim is to perform a detailed reality check of the TREC ad hoc simulation in the field. Readers should note that the scope of current TREC evaluations is much broader than the ad hoc task[2] and includes question answering tasks, known-item searches and queries and data taken from the web.

We took a text retrieval system which has performed well in TREC ad hoc and transformed it into an intranet search engine which is now in production use at a university. During the transformation, we re-visited each of the key design decisions in the light of our understanding of the World Wide Web (WWW) and of intranets. Here we present our design decisions and, as far as is possible, evaluate them using query processing logs collected during the first three months of operation.

2 Evaluation methodology

Our goal was to evaluate the "out-of-the-laboratory" merit of retrieval system design features which had evolved in a test collection environment. Because of this the range of evaluation methodologies was limited. Obviously, a test collection could not be used!

An interactive experiment with human subjects (such as the TREC interactive track studies [10]) would take us back into the laboratory. Furthermore, because there were eleven different design dimensions (see Section 4) to study, the required scale of an interactive experiment would have been totally beyond our resources.

Naturalistic observation or video recording of real searchers at work would have been totally impractical due to the spontaneous, sporadic and geographically scattered (mostly off-campus) characteristics of searching behaviour and our interest in studying searching phenomena which may occur only infrequently.

We also envisaged problems in using a user survey to evaluate our multi-factorial decisions. Spink et al [13] report a study of on-line feedback provided by users of the Excite WWW search engine. They report a relatively low response rate and indications that responses may not have been representative of the overall searcher experience.

Jansen et al [5] and Silverstein et al [12] present detailed analyses of WWW search engine[3] query logs. This technique allows study of the full population of actual queries without need for simulation or for sampling of either queries or searchers. Analysis can focus on issues of most importance to searchers by concentrating on frequently repeated queries or query types.

The biggest disadvantage of evaluation by query log analysis is that, in isolation, it may be difficult to deduce a searcher's information need from a query submitted to a search engine and thus difficult to judge the quality of answers. This is particularly the case for very short queries. Here, however, the limited scope of a university intranet makes it is possible to deduce underlying information need with a much greater degree of confidence.

For example, it is very likely that searchers who submit `physics`, `anthropology` or `latin` as queries to a university intranet search facility are in fact looking for

[1] Use of subject metadata is actually forbidden.

[2] Indeed there will be no ad hoc task per se in TREC-9.

[3] EXCITE and AltaVista respectively

the departmental websites for those disciplines, rather than encyclopaedia or dictionary entries for them. Similarly, searchers who ask about `scholarship` are very likely to be intending applicants rather than people interested in learning about scholarship through the ages.

Accordingly, we chose to perform our evaluation using analysis of query logs. We used a combination of formal and informal knowledge of the local vocabulary (including people's names, names of departments, geographical entities, locally significant acronyms, important university functions and so on) to interpret the logs. In many cases, we re-ran particular queries with and without a particular system feature to observe whether the availability of the feature made a significant difference to the results.

HTTP is a stateless protocol and our present query log format does not allow us to unambiguously distinguish between multiple users submitting identical queries and a single user requesting multiple result screens. The following analysis assumes that an interval of less than five minutes between submissions of the same query indicates that a single searcher has requested multiple result screens.

3 General observations on the logs

Table 1: Scope of the intranet studied. Also documented are the period covered by the study and the number of queries analysed.

Period commenced	01 Aug 1999
Number of days	99
Number of distinct hosts	177
Size of data	3450.2 MB
Number of pages indexed	336,336
% pages with a last-modified date	33.21
% pages with recognised metadata	82.83
% pages with Dublin Core tags	0.07
% pages with author metadata	2.74
% pages with keyword metadata	6.27
Total queries	45312
External queries	33119 (73.1%)
Average query length (wd.s)	2.31

Table 1 indicates the scope of the particular intranet and of the logs analysed in the present study. The search service was launched in July 1999 and runs on a relatively inexpensive PC system[4]. The period covered spans the first three months of service. However, the logs for the first couple of days were believed to include a high proportion of artificial test queries and have not been included.

[4]dual Pentium III, 450 MHz, 1024 MB RAM

```
542 library
208 exam timetable
198 employment
197 scholarship
192 isis
191 scholarships
168 jobs
145 samba
119 vacancies
116 timetable
112 law
112 accomodation
 97 mba
 96 map
 93 accommodation
 92 summer school
 91 university house
 91 library catalogue
 87 psychology
 85 courses
 84 roger clarke
 75 what is marxism
 75 religion
 75 positions vacant
 74 forestry
```

Figure 1: The 25 most popular queries (after case folding). ISIS is the acronym for an IT support organisation within the university, Samba is open-source software developed within the university and Roger Clarke is a prominent former academic and current visitor.

The proportion of queries submitted from outside the intranet is much higher than expected.

3.1 How does intranet search differ from WWW search?

The Real World Searching panel at SIGIR97 as reported by Lesk [8] outlined the economic fundamentals and logistic difficulties of operating a WWW search engine and gave valuable insights into searcher behaviour.

WWW publishers may stuff invisible parts of their pages with keywords intended to cause presentation of documents unrelated to the searcher's query, resulting in frequent annoyance to the searcher and occasional commercial benefit to the publisher. This activity is known as *spamming* or *stuffing* and causes considerable headaches to WWW search engines. By contrast, on an intranet, search-facilitating content modification is generally in the interest of searchers.

The most obvious difference between the WWW and intranet search logs is the several orders of magnitude difference in both volume of queries and quantity of data. It seems obvious that intranet search engines can afford to use better retrieval methods, even if this significantly increases computational cost. The query load on the search service under study averaged only 458 queries per day. Anecdotal evidence suggests that response time is consistently good. Consequently, there is little need to analyse efficiency issues here.

Perhaps surprisingly, the average query length of 2.31 reported in Table 1 is almost identical to the 2.35 reported in both [12] and [5]. However, the profile of queries submitted is very different. Comparison of the 25 most popular queries presented by [12, p.8] with those in Figure 1 reveals not a single word in common.

Not surprisingly, since "adult content" sites are not supposed to operate within universities, the proportion of sexually oriented queries is two orders of magnitude lower than on the WWW. We found that only

97 queries (0.22%) contained a word from a list of 108 adult content words.

3.2 What was sought?

Examination of the data in Figure 1, shows that all of the top 25 queries from the logs relate to the everyday business of the University or to issues covered in teaching or research (Marxism and religion). Not surprisingly, this is true for almost all queries submitted. People do not often search for what is not likely to be there.

49 roger clarke	12 Vince Craig
33 Roger Clarke	11 simon grant
30 karl marx	11 peter
24 Karl Marx	10 snyder
16 Simon Grant	10 pagan
12 baume	10 clarke

Figure 2: The 12 most popular queries involving a person's name (case preserved).

Searchers very frequently submit people's names as queries. Using the university's on-line internal telephone directory as a source of family names and of forenames, and a set of simple heuristics, we selected all the queries which appeared to contain a reference to a person's name. Such queries numbered 5483 (12.16% of all queries). The list includes a number of false hits, such as `Bruce Hall` (the name of a student residence) but also misses names of people not listed in the telephone book such as `Karl Marx`. Often names are specified as single family names (`baume`, `snyder` and `pagan` are most frequent) or as single forenames (`peter` and `phil` are most popular).

Our intuition is that people's names may be submitted as queries for a variety of reasons:

1. The searcher is looking for contact details such as telephone number, or email address.
2. The searcher is looking for some sort of biographical information about the person. The searcher may be trying to locate a former colleague or trying to find background information for a newspaper article.
3. The searcher is trying to access one or more of the person's publications.

Our intuition is that, in each of these cases, the searcher will be well served if the person's home page is presented at the top of the results list. It would be desirable to also present the results of a lookup on the university's phone/email directory database.

Given the frequency of occurence of queries consisting of just a person's name, it may be worthwhile to attempt to identify people's names in documents and queries and to recognize that Professor Carmody, John Carmody, J.B. Carmody, Professor J. Carmody, etc. are all possibly references to the same person. Lawrence et al [7] mention the use of methods for achieving this. Clustering of results might allow different people (or different personas) sharing the same name to be distinguished.

Many queries give the name of internal university organisations. Examples from the top 25 include `library`, `ISIS`, `law`, `university house`, `psychology` and `forestry`. It is presumed in these cases that the searcher will usually be satisfied if the search highlights the home page for the relevant organisation. It may also be useful to present results of organisational directory lookups.

4 Evaluation of design decisions

This section presents the salient features of the retrieval methods used by the intranet search engine as designed and highlights differences from the system used in TREC ad hoc. Both systems are based on inverted file indexes.

The presentation of each feature records the design decisions taken for both TREC ad hoc and for our intranet search system. It is followed by a description of the results from the relevant query log analysis and conclusions about the merits of the decision taken.

4.1 Case folding

294 (515) ANU	20 (397) MARX
268 (1926) AND	19 (732) LAW
71 (129) MBA	19 (26) APAC
61 (203) ISIS	18 (75) ANUTECH
42 (63) IT	17 (31) OPRS
32 (1890) OF	16 (22) RSSS
29 (59) CRES	16 (24) ITA

Figure 3: The 12 most popular apparent acronyms. The first number is the frequency of occurrence in the logs of the all-uppercase version of the acronym. The second is the total query frequency of all case variants. It is fairly clear that `AND` is not an acronym but is rather intended to be a Boolean operator, despite the fact that that query syntax is not supported. Similarly, `OF`, `MARX` and `LAW` do not appear to be true acronyms.

TREC ad hoc: All query and document words were converted to lower case on the grounds that the resulting increase in the number of query term occurrences found might on average improve both recall and precision. Case sensitivity is an advantage in very few TREC topics.

Intranet Design: The same case-insensitive approach was taken on the same grounds.

Intranet Results: Case-sensitivity would distinguish some frequently used acronyms from common homographs. Examples include `IT`, `ITS`, `ACT`, and `ARC`.

Figure 3 lists the most frequently used acronyms. It is interesting to observe that searchers type certain

acronyms such as `MBA` and `ISIS` more often than not in lower or mixed case.

Table 2: Use of case in queries. Any query word consisting entirely of two or more upper case letters is classified as an acronym.

All letters lower case	33126 (73.48%)
All letters upper case	1674 (3.71%)
At least one apparent acronym	2452 (5.44%)

Further inspection of the logs confirms that searchers are quite cavalier in their use of case. Examples abound in which proper names are typed entirely in lower case, or in which entire queries are typed in upper case. Table 2 documents the frequency of use of upper and lower case.

Conclusion: Examples have been identified in which case sensitivity would improve results, but it is also clear that searchers do not currently give reliable signals. We conclude that indexes should include case information, that case insensitivity should be the default, and that case-sensitive search should be provided as an option.

4.2 Stemming

TREC ad hoc: The well-known Porter rule-based stemmer was employed both during indexing and when processing queries. The intention was to increase the number of query term occurrences found and thereby improve both recall and precision. On some topics, the broadening of meaning (`schooling` → `school` as in `home schooling`) was observed to do harm but the harmful effect in most cases was limited by the additional context provided by the large number of terms in the TREC topics.

Intranet Design: In a web environment however, the combination of huge document collections and very short queries often gives rise to examples where stemming does obvious harm. Accordingly, the design of the intranet search engine features an unstemmed index and the ability to perform stem-matching at query time[5].

Intranet Results: The main area in which stemming was expected to cause problems in intranet search was the processing of people's names. As noted elsewhere, (Section 3.2 and Figure 2) searchers often specify only a single name, meaning that ambiguity introduced by stemming could not be corrected by contextual information. Furthermore, searchers frequently do not use capitalisation to indicate that a word is in fact a proper noun, making it difficult to restrict stemming to non-proper nouns.

[5]The stemming operator is not advertised on the search interface.

Of the estimated 5483 occurrences of a person's name within queries, 1905 (34.74%) are affected by stemming. Potentially the results for all these queries could be made worse by stemming, but problems only occur when the use of stems actually increases ambiguity.

A number of family names represented in the university become much more ambiguous when stemming is applied. There is only one person in the ANU phone book with the family name of "Lees" but when stemming is applied there are seventeen people whose family name matches. Furthermore the stem also matches the ordinary word "lee".

In our intranet study, stemming increases the number of matches for "Davy" from two to twenty-one. Other names for which ambiguity increases significantly due to stemming include "Clarke", "Davis", "Woods", "Welling", "Cooke", "Lowe" and "Burne". Several of these stem to commonplace words.

Conversely, failing to use stemming sometimes reduces the quality of results. For example, several of the queries in the top 25 list are either singular or plural (eg. `jobs`, `vacancies`, `scholarships`). A better (more comprehensive) result list is obtained if stemming is applied i these cases.

Conclusion: It seems sensible to retain an unstemmed index but to apply, when appropriate, stemming or at least singular/plural conflation at query processing time. User education and/or interface modification are required to increase the reliability of signals as to when stemming is inappropriate at query processing time.

4.3 Truncation

TREC ad hoc: Truncation operators were not used.

Intranet Design: Manually generated query word truncation allows query simplification when seeking multiple targets and may be used to cope with uncertain spelling.

Intranet Results: Table 3 shows that the truncation operator is used infrequently. Inspection of the logs shows that it is most often used to simulate stemming or to match words with a common prefix. The most common valid examples were `student*`(4), `rainforest*`(3), `multicultura*`(3), `impair*`(3), `fingerprint*`(3), `eff*`(3) and `australia*`(3).

One of the uses of `impair*` was within the phrase `"hearing impair*"` (57 results). It produced a useful merging of result lists for `"hearing impaired"` (49), `"hearing impairment"` (24) and `"hearing impairments"`. Other examples examined achieved similar benefit but some, like `eff*` and `p*` generated results which were too broad to be useful. Care is needed to avoid excessive false hits.

17 queries used truncation in ways not supported by the present system. Eg. `vas*il*ate`, `*`, and `**`.

Conclusion: On the whole, not much can be

learned from these observations. In very rare cases, searchers gained benefits they could not have gained through use of the stemming operator.

4.4 Stopword elimination

TREC ad hoc: Stopword elimination in TREC is usually motivated by a desire to reduce index space.

Intranet Design: We considered that stopwords sometimes contribute useful information, particularly within phrases such as quotations (eg. "to be or not to be") and titles (eg. "As you like it"). However, in many contexts, stopword processing results in excessive and unproductive computation. It was decided to index all words but to process stopwords only within query phrases and within queries containing two or fewer non-stopwords.

Intranet Results: Acronyms such as "IT", and the initial "A" in people's names are used frequently enough and would normally count as stopwords. However, searchers used very few phrases including a stopword.

Conclusion: We found no evidence that indexing of stopwords caused harm. However, benefit was gained in only a very small percentage of cases.

4.5 Phrases

```
 9 "network services"
 9 "gated pipe"
 7 "stuart hay"
 7 "peter buckingham"
 6 "social identity theory"
 6 "mutual reciprocity"
 6 "manfred pienemann"
 6 "creative writing"
 6 "computer science"
 5 "undergraduate handbook"
 5 "student exchange"
 5 "strategic directions"
 5 "puppy sculpture"
 5 "northern territory"
```

Figure 4: The 14 most popular explicitly specified phrases (case-folded). Three test phrases submitted by the present authors have been removed from the list.

TREC ad hoc: Leading participants in TREC have reported mixed results from use of phrases automatically extracted from TREC topics.

Intranet Design: We thought that phrases explicitly identified by searchers would be more consistently useful than automatically extracted ones, especially in the context of very short queries. Accordingly, phrases are supported.

Intranet Results: Table 3 reports that 2143 queries included explicit phrases. However only 1969 of these included apparently well-formed phrases (those with balanced quotes and at least two words.) There was also some evidence of inappropriate use of phrases. Some were so long as to be unlikely to produce any matches.

There were many examples of queries whose results were significantly improved by use of explicit phrases. For example there are 1199 documents in which the query words `green` and `paper` co-occur but only 89 containing the phrase `"green paper"`. Similarly, inspection of results for queries like `"contract law"`, `"university secretary"`, `"external courses"`, and `"family leave"` reveals dramatic improvements due to use of the phrase operator.

Figure 4 lists the most popular phrases used. Surprisingly, except for `"creative writing"`, all of these examples produce very similar top 10 results whether or not the phrase operator was used.

Conclusion: Explicit phrases were used quite frequently and often made a positive difference.

4.6 What to index?

```
24 stat1007
14 comm1101
12 phil1003
12 paper 1074/1999
12 macroeconomics 3
11 y2k
10 stat2003
10 stat1006
10 integrity teaching - had a 1994
   address do you have a current article
10 biol1002
 9 forestry 1003
 8 handbook 2000
```

Figure 5: The 12 most popular queries involving digits. The ten occurrences of the long query were almost certainly submitted by the same person.

Table 3: Percentage of queries using operators such as quotation marks, truncation, disjunction and so on.

Any operator	3818	8.47%
Disjunction operator	214	0.47%
Negation	23	0.05%
Mandatory inclusion (plus)	462	1.02%
Mandatory exclusion (minus)	85	0.19%
Truncation operator	143	0.32%
Phrase operator	2164	4.8%
At least one digit	1626	3.61%
At least one number	797	1.77%

TREC ad hoc: It is possible to obtain reasonable

performance on the TREC ad hoc task while indexing only words (sequences of letters).

Intranet Design: We suspected that the intranet search service would receive a certain volume of queries in which a numerical item constituted a highly discriminating feature. For example, the annual report or degree rules for a particular year, a particular telephone number, a university subject code such as `ECON3004` and so on.

Intranet Results: Table 3 shows that a few percent of queries do in fact contain digits and several hundred contain a number (ie. an all-digit query term). Figures 5 and 6 illustrate the types of query involving digits and the numbers which are frequently requested. The query `handbook 2000` gives far more specific results than does `handbook`.

Conclusion: Indexing numbers and letter/digit combinations has indeed turned out to be useful. It might be useful to extend the word definition to allow recognition of currency amounts and strings such as `C++`. The resultant increase in index size would be negligible.

4.7 Departmental search

TREC ad hoc: The TREC ad hoc task does not require the ability to restrict searches to subcollections.

Intranet Design: Initially, the intranet search did not provide this ability either. However, no sooner had the new service had been anounced than several webmasters within the university requested provision of a means of restricting searches to their own part of the intranet. For example, the Faculty of Engineering and Information Technology wanted to provide search interfaces on their four servers which searched only the contents of that group of hosts. The Human Resources division wished to provide searches restricted to just one directory on the university's central web server.

Accordingly, a solution based on metadata constraints was engineered. The URL[6] of each page is interpreted as hostname and filename metadata for the page. Departmental search queries are automatically prefixed with hidden search conditions which mandate the satisfaction of constraints based on these fields.

The next section presents results and conclusions for departmental search.

[6]The full web address for the document.

```
191 1999                18 99
187 2000                18 1994
 31 2                   13 1997
 23 1003                12 5
 22 1998                12 1
 19 3                   10 11
```

Figure 6: The 12 most popular numbers.

4.8 Metadata extraction and querying

TREC ad hoc: TREC ad hoc topics do not include relational constraints such as `author = Smith` or `year = 1999`. Therefore there is no need for TREC ad hoc systems to include facilities for querying metadata.

Intranet Design: Small parts of the university intranet are thoroughly marked up with metadata useful for search (eg. Dublin Core [4]) and it is expected that the amount of metadata markup within the University will increase in the future. Accordingly the intranet search engine was designed to support date range and string-within-nominated-metadata-field constraints coupled with content queries. The metadata/content framework is described in detail elsewhere. [3] By default, the title and keywords metadata are considered to be also part of the document content.

As mentioned in the preceding section, the metadata constraint mechanism is used to implement departmental searches.

At the present stage of implementation, two alternative versions of the central search graphical interface are provided. The "simple search" version allows searchers to type in metadata constraints in a not prominently documented and potentially error-prone command language. For example: `a:Carmody d<1999>1995 luminescence` means to rank documents which were published between 1996 and 1998, whose author value includes the word Carmody and whose content relates to luminescence. As may be seen in Section 4.11, few searchers are likely to have the patience to learn how to type this form of query correctly.

Currently, the "advanced search" option provides a graphical interface to hostname, filename and date metadata only.

Table 4: Summary of the number of queries including various types of metadata constraint. The total number of non-date metadata fields used was 9327.

Any sort of metadata	5225	11.53%
Including date constraints	303	0.67%
Host metadata constraints	4880	52.32%
Filename metadata constraints	4428	47.48%
non-host, non-filename metadata	19	0.04%

Intranet Results: Table 4 documents the use of metadata constraints in queries. As may be seen, a proportion of queries do include one or two metadata constraints. The vast majority of these relate to hostname or filename parts of the document URL. We surmise that they are departmental searches but it is not possible to be certain. Date constraints are applied in a small proportion of cases but the use of metadata constraints not supported by the "advanced search" interface is negligible.

Conclusion: The metadata constraint facility turned out to be very useful for implementing de-

partmental searches. The other applications envisaged have not yet eventuated but may do so when:

- metadata is more widely used and more widely understood within the university; or
- the "advanced search" interface is extended to give access to the other metadata fields; or
- the search system is deployed in a more metadata-focussed environment.

4.9 Duplicate elimination

TREC ad hoc: There are few duplicate documents and no incentive for systems to ignore them, since each document is judged independently.

Intranet Design: On the Web, duplicate documents are an unavoidable result of spidering[7]. They arise from the existence of multiple names for hosts, directories and files through links and aliases, as well as through content mirroring. Our intranet search engine calculates a checksum on the content of each document at indexing time. At query time, duplicates are identified by checksum comparison and presented as a single document with multiple URLs.

Intranet Results: Of the 336,336 documents, there are only 221,981 which occur only once. There are a further 45,522 documents whose content is multiply represented in the collection, leading to 267,503 distinct checksums. There is a 68,833/336,336 = 0.205 chance that an arbitrary document will have a duplicate. Thirteen different checksums occur more than 100 times. The most frequent occurs 442 times.

Conclusion: Assuming selection of documents in result lists is unbiased with respect to duplication, one would expect there to be an average of nearly two pairs of duplicates in each result page. It is possible that a query might retrieve more than a hundred copies of the same document. We conclude that duplicate detection is a very useful feature in an intranet search engine.

4.10 Limits to content ranking

TREC ad hoc: Documents are ranked purely on the basis of the "similarity" of their content to the query.

Intranet Design: The initial version of our intranet search was engineered to take the same approach but the use of link-based methods such as those of Kleinberg [6] and Brin and Page [11] was envisaged as a future improvement.

Intranet Results: It transpires that content ranking does not work well for certain key pages. As noted in Figure 1 the most popular query submitted is `library`. The Library web master would like this query to retrieve the Library home page as the first result, but in fact it is ranked 7985! According to the relevance scoring formula, there are thousands of web pages which are more "about" `library` than the library home page. A similar but less extreme situation arises with `forestry`.

Conclusion: Although individual webmasters can address this problem by modifying the text content of their pages, the intranet search service will shortly be modified to introduce a Key Pages result panel above the conventional ranked list of results. When a single word query Q1 is submitted, the Key Pages panel will suggest `http://Q1.<domain>` and `http://www.<domain>/Q1` if these are found to exist within the university's domain. Every query submitted will also be looked up in a site-specific table of query-to-keypage translations. As an example this table will include a translation for "computer science" to `http://cs.<domain>`.

4.11 Spelling correction

TREC ad hoc: Spelling errors within TREC topics are extremely rare. No facilities to cope with such errors were provided.

Intranet Design: This issue was not reconsidered.

Intranet Results: Query terms extracted from the logs were matched against a list of correctly spelled words and misspellings were counted. The spelling list comprised the Unix spelling dictionary `/usr/dict/words` augmented by names from the university telephone directory plus identified university acronyms plus a list of other words from the logs which were considered to be correctly spelled.

On this basis, it is estimated that 6033 or 13% of all queries included a spelling error. There appear to be eleven different misspellings of the most popular query (`library`), and thirteen for `accommodation`. Amazingly, there are 28 different incorrect spellings of `scholarship(s)`. Perhaps this is a preliminary screening test for applicants! Finally, as seen in Figure 1, one misspelled version of accommodation occurs more frequently than the correct one!

Conclusion: Clearly, the search engine needs a better method for dealing with misspelled queries. One approach would look up each query word in an organisation-specific wordlist and, if not found, suggest the closest spelled words in the list, using an approximate matching (eg. those of Manber and colleagues [9] or sound-alike technique.

5 Discussion and conclusions

Space does not permit analysis of many other aspects of the mutation of a TREC ad hoc system into a useful intranet search engine. Some very important ones include spider policy and coverage, user interface design, result presentation, text extraction from encoded document formats and security. A lot of these issues are largely shared by WWW search engines and have been written about by eg. [1] and [7].

[7] The process of finding and retrieving web documents for indexing, by repeatedly extracting and following links.

General search statistics such as average query length, use of query operators, proportion of errors and so on in the university environment were very similar to those observed in WWW search engine query logs. This may reflect the high proportion of externally submitted queries and/or the highly educated demographic profile of users of WWW search engines [13].

Despite the similarities with WWW search, intranet search shows many differences. In our case, the balance of queries is clearly oriented toward the business of the university. Spamming is not a problem because there is no commercial incentive for it to occur. An intranet search engine experiences dramatically lower query processing loads and much smaller collection sizes. Accordingly, it is feasible to use more effective retrieval techniques, even if they are computationally expensive. It is also possible to provide organisation-specific heuristics to more effectively guide searchers to their goals.

The intranet environment differs from TREC ad hoc, in that:

1. Queries are very short. This tends to increase the significance of subtle lexical signals such as case.
2. Spelling errors are common. Methods (probably organisation-specific) are needed to deal with them.
3. Average precision at 1000 documents retrieved (the main ranking measure in TREC ad hoc) is not an appropriate evaluation measure for intranet search. Among the set of relevant documents on an intranet, there is often a single "best answer" which should be presented first. This document may be the only one accessed by the searcher. In many cases the searcher may be presumed to be performing a "suspected item" search. eg. "the page which tells me how to enrol", "the page which lists all the scholarships people can apply for", or "the page which gives me access to the library catalog".
4. There is a need for special processing of names of people and names of departments or offices within the university.
5. Among university web publishers there is a strong desire to be able to restrict answers to particular subsets of the intranet.
6. Automatic query generation does not come into play. Consequently, query operators such as truncation and phrases can be more consistently effective, at least in the hands of skilled searchers.
7. Identification of duplicate documents improves results.
8. In the future, when more university intranet web publishers provide basic metadata and more searchers understand how to use metadata, combined metadata/content queries will provide a means of improving intranet search results. Because metadata/content queries were so infrequently used in our case study (apart from via the departmental search mechanism), it is not possible to draw any conclusions as to the effectiveness of our design for handling them.

Our basic philosophy is to eliminate as little useful information as possible when indexing and to provide query-time facilities for increasing the number of matches, such as stemming, truncation, case folding and query expansion. More efficient implementations of some of these features are needed and mechanisms are needed to intelligently determine when they should be applied.

The apparent lack of consistency and sophistication on the part of many users presents a major challenge. A user interface is needed which is effective in encouraging users to provide reliable signals as to how their query should be best interpreted.

The TREC ad hoc task has provided a powerful and invaluable force for the improvement of text retrieval systems. However, caution is needed in translating its lessons from the laboratory into particular search applications in the field.

References

[1] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. In Helen Ashman and Paul Thistlewaite, editors, *Proceedings of the WWW7 Conference, Brisbane*, pages 107–117, Amsterdam, April 1998. Elsevier.

[2] Chris Buckley, Amit Singhal, and Mandar Mitra. Using query zoning and correlation within SMART: TREC-5. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of TREC-5*, pages 105–118, Gaithersburg MD, November 1996. NIST special publication 500-238, `http://trec.nist.gov`.

[3] Peter Bailey David Hawking and Nick Craswell. Efficient and flexible search using text and metadata. Submitted to IEEE Advances in Digital Libraries, ADL 2000.

[4] Dublin Core Directorate. Dublin core metadata initiative. `http://purl.oclc.org/dc/`.

[5] Bernard J. Jansen, Amanda Spink, Judy Bateman, and Tefko Saracevic. Real life information retrieval: A study of user queries on the Web. *ACM SIGIR Forum*, 32(1):5–17, 1998.

[6] Jon Kleinberg. Authoritative sources in a hyperlinked environment. Technical Report RJ 10076, IBM, May 1997.

[7] Steve Lawrence, C. Lee Giles, and Kurt Bollacker. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67–71, 1999.

[8] Michael Lesk. Report on "Real World" searching panel at SIGIR97. *ACM SIGIR Forum*, 32(1):1–4, Spring 1998.

[9] Robert Muth and Udi Manber. Approximate multiple string search. `http://glimpse.cs.arizona.edu:1994/udi.html`, 1994.

[10] Paul Over. Trec-7 interactive track report. In Voorhees and Harman [14], pages 65–71. NIST special publication 500-242, http://trec.nist.gov.

[11] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford, Santa Barbara, CA 93106, January 1998. http://www-db.stanford.edu/~backrub/pageranksub.ps.

[12] Craig Silverstein, Monika Henzinger, Hannes Marais, and Michael Moricz. Analysis of a very large AltaVista query log. Technical Report 1998-014, Digital Systems Research Center, Palo Alto, 1998. http://www.research.digital.com/SRC.

[13] Amanda Spink, Judy Bateman, and Bernard J. Jansen. Searching the Web: a survey of EXCITE users. *Internet Research: Electronic Networking Applications and Policy*, 9(2):117–128, 1999. ISSN 1066-2243.

[14] E. M. Voorhees and D. K. Harman, editors. *Proceedings of TREC-7*, Gaithersburg MD, November 1998. NIST special publication 500-242, http://trec.nist.gov.