Challenges in Enterprise Search

David Hawking

CSIRO ICT Centre,
GPO Box 664,
Canberra, Australia 2601
David.Hawking@csiro.au

Abstract

Concerted research effort since the nineteen fifties has lead to effective methods for retrieval of relevant documents from homogeneous collections of text, such as newspaper archives, scientific abstracts and CD-ROM encyclopaedias. However, the triumph of the Web in the nineteen nineties forced a significant paradigm shift in the Information Retrieval field because of the need to address the issues of enormous scale, fluid collection definition, great heterogeneity, unfettered interlinking, democratic publishing, the presence of adversaries and most of all the diversity of purposes for which Web search may be used. Now, the IR field is confronted with a challenge of similarly daunting dimensions – how to bring highly effective search to the complex information spaces within enterprises. Overcoming the challenge would bring massive economic benefit, but victory is far from assured. The present work characterises enterprise search, hints at its economic magnitude, states some of the unsolved research questions in the domain of enterprise search need, proposes an enterprise search test collection and presents results for a small but interesting subproblem.

Keywords: Information Retrieval; enterprise search; search quality evaluation.

1 Introduction

The IDC report entitled "The High Cost of Not Finding Information" (Feldman and Sherman, 2003) quantifies the significant economic penalties caused by poor quality search within enterprises, both in the form of lost opportunities and through lost productivity. CSIRO's observations of a large number of Australasian organisations suggest that these were not isolated or exceptional cases – in fact, very poor enterprise search is the norm and, while employees and customers may complain or laugh about it, the organisation as a whole typically fails either to recognize the seriousness of the situation or the possibility of doing better.

I interpret the term *enterprise search* to include:

- any organisation with text content in electronic form;
- search of the organisation's external website;
- search of the organisation's internal websites (it's intranet);
- search of other electronic text held by the organisation in the form of email, database records, documents on fileshares and the like.

In general, search of non-textual and continuous media is included but here I consider only retrieval mediated by text (e.g. retrieving a video by matching textual annotations associated with it against a text query rather than by measuring its visual or auditory similarity to a video query).

An obvious reason for poor enterprise search is that a high performing text retrieval algorithm developed in the laboratory cannot be applied without extensive engineering to the enterprise search problem because of the complexity of typical enterprise information spaces. Out of the hundreds of search engines on the market, so few are able to work with the range of databases, content management systems, email formats, document formats, operational and security requirements typical of medium scale enterprises that quality of search results is often forgotten when purchasing decisions are made. (See Stenmark (1999) for an illustration.)

Not only does enterprise information complexity restrict the range of applicable commercial search products and increase the cost of deploying them but it makes it difficult to measure the quality of search results obtained and also, we hypothesise, makes it hard to approach the effectiveness level achieved by state-of-the-art whole-of-Web search engines.

For researchers in the Information Retrieval (IR) field and for commercial companies, the problem of enterprise search is a formidable challenge but one for which a solution would deliver enormous benefit.

No solution can be presented here, but it is hoped that a characterisation of the problem, a list of open research questions, a discussion of approaches and a preliminary proposal for an evaluation framework may attract researchers to work in the area and thereby accelerate progress toward a solution.

Section 2 characterises the problem of enterprise search in the light of past work and in terms of some typical enterprise search scenarios. Section 3 enumerates and characterises a number of open research problems, including the development of a test collection for enterprise search. Section 4 addresses the specific problem of searching enterprise data held in web format and reports new results across a range of outward-facing enterprise web sites. It also includes

a discussion of the approach of converting heterogeneous web data into a common web format. Section 5 concludes.

2 Characterising Enterprise Search

Stenmark (1999) is one of the few authors to have addressed the enterprise search problem in its full generality. He distills his experiences with purchasing intranet search engines for two large organisations into a generalised purchasing guide, but none of the 81 criteria in his tables (e.g. What platforms does the product run on? What formats other than HTML can be indexed by default? Is the index updated in real time? Can Boolean type queries be asked? How easy is the product to install and maintain?) relate to the quality of search results.

Abrol et al. (2001) propose a "business portal" as a solution to the full problem of enterprise search and identify the following characteristics:

1. *The need to access information in diverse repositories including file systems, HTTP web servers, Lotus Notes, Microsoft Exchange, content management systems such as Documentum, as well as relational databases.*
2. *The need to respect fine-grained individual access-control rights, typically at the document level; thus two users issuing the same search/navigation request may see differing sets of documents due to the differences in their privileges.*
3. *The need to index and search a large variety of document types (formats), such as PDF, Microsoft Word and Powerpoint files, etc. and different languages (such as, English, European and Asian languages).*
4. *The need to seamlessly and scalably combine structured (e.g. relational) as well as unstructured information in a document for search, as well as for organizational purposes (clustering, classification, etc.) and for personalisation.*

The above characteristics do not fully represent the complexity of the situation. Many enterprises are now building systems in which documents are synthesised from paragraphs or sub-documents held in a database. Which candidate paragraphs are actually presented to a particular searcher may depend upon both their personal interest profile and their access rights.

A number of presentations at the 2003 Infonortics Search Engines Meeting (`www.infonortics.com/searchengines/`) addressed selection, implementation and integration of search engines within enterprises.

Like Stenmark, Abrol et al and the Search Engine Meeting presenters paid little attention to search result quality.

From an IR perspective, many of the challenges inherent in the above are of an engineering rather than a research nature. However difficult it may be to build a text extraction filter for some proprietary document format or an adapter for a commercial CMS system or a record-level, profile-driven right-to-view search filter, the exercise is unlikely to lead to generalisable IR or IS research insights. Where then does the research interest lie in Enterprise Search?

The ultimate goal of an Enterprise IR system is to respond to a request by searching all the documents which may possibly contain a useful answer (and which the searcher is entitled to see) and to present search results in a form or order which is of *maximal utility to the searcher*[1]. What this means in practice depends very much on the nature of the organisation, the identity of the searcher and the characteristics of the task to which the search results will contribute.

Broder's three search types: informational, navigational and transactional (Broder, 2002) are all represented in enterprise search. His "third-generation search" features ("answering the need behind the query") are also very applicable in Enterprise search. Examples include corpus-moderated spelling correction (did you mean ...?), query-targeted "advertising", transfer of search context when the searcher accesses search results, presentation of useful results from sources other than the corpus being searched (e.g. results from database lookup of a person's name, and recent news items relating to the topic of the query.) Finally, source diversity is a generally desirable characteristic of enterprise search, as it is for whole-of-web search; It is undesirable to present dozens of consecutive results from the same website or email thread at the expense of similarly rated results from different sites or threads.

Consider the following scenarios:

2.1 Scenario 1: External visitor to enterprise website

In this simplest of scenarios, the complexities introduced by privacy, security, heterogeneous formats and the like are avoided and normal Web search techniques are appropriate. In general, priority should be given to site entry pages and highly referenced pages over ones which are merely relevant. For example, the query "media releases" should return the home page of the companies media releases site (or a list of most recent releases) rather than an arbitrary list of ancient issues. It may be necessary to map the language of the query to the language of the site (e.g. "press release" to "media releases").

2.2 Scenario 2: Intranet search

Employees within a company with a well-developed intranet are able to use intranet search to locate company policies, financial information for their project, and client histories and also to locate online services where they can generate invoices, claim expenses, log effort and renew their staff card. Relative to Scenario 1, this type of search may be characterised by fewer problems of language mismatch (because employees are indoctrinated in the company's way of doing things), but significant problems occasioned by the company security model.

2.3 Scenario 3: Internal multi-source search

A project manager in an oil exploration company wants to identify all wells previously drilled by the

[1] Note that enterprises generally prefer a utility function which reflects what the enterprise wants the searcher to see!

company in the Black Gold field where the problem known as "stuck pipe" was experienced. She types the query 'Black Gold stuck pipe'.

Although the company retains drilling reports for all the wells it has drilled, a variety of different language is used in the reports, and the reports are held in a surprising variety of formats and locations: Some have been entered into the official database system, some exist as email messages, some as spreadsheets on a shared hard drive in company headquarters and some as OCR-ed text files. The problems here are the language gap between the query and the reports (there are many different ways in which the stuck pipe problem and the location of wells is represented in the reports), the difficulty of accessing all the data, the need to provide a consistent and effective usefulness ranking across the different repositories, and the need to present a list of distinct wells (not all documents) with links to well charts.

2.4 Scenario 4: Searching for other than documents

Within an enterprise it is possible to intersect structured (such as stafflists or lists of company divisions) and less structured sources of information (such as web pages and email messages) to provide lists of people, divisions, geographic areas which are related to a topic. This could be used to identify people or groups with relevant expertise or experience. Craswell et al. (2001b)

2.5 Scenario 5: Task-integrated corporate memory search

A newly recruited sales manager SM is about to attempt to sell the enterprise's services to a target customer XYZ Pty Ltd and types "XYZ Pty Ltd" as a query to his company's internal search engine. In many organisations, the search interface would not give access to the databases, fileshares or email in which most of the company's interactions with XYZ are logged and would not go beyond the company's own data. Instead it might return a few intranet web pages ranked in an order which strikes SM as arbitrary.

One could envisage a better search service which searched all the relevant data and generated a high quality ranking (in which recency may well be a key component), weighting the different types of document appropriately.

Even better would be an engine which searched all the relevant data both inside and outside the company and automatically synthesised a carefully presented result page including:

- XYZ contact details, including details and pictures of key people if available,
- A picture of XYZ's current financial status obtained by accessing stock exchange reports and shareprice trends, articles in financial newspapers, and reports from industry analysts like Dunn & Bradstreet.
- A synopsis of recent financial transactions between XYZ and the company - what is the nature and scale of the existing relationship, are there outstanding debts?
- General information about XYZ obtained from news sites and the general Web,
- A synopsis of email conversations between XYZ and the company with due emphasis on recent traffic. Ideally the email threads could be summarised in such a way as to identify major issues and the current state of play.
- A list of company staff (with contact details) who have recently dealt with XYZ.

2.6 Scenario 6: Forensic search

Search of corporate records, including unstructured email communication plays a vital role in legal matters ranging from patent litigation, insider trading investigation, liability actions, and analysis of the causes of bankruptcy.

2.7 Summary

As may be seen in the preceding scenarios, there is considerable opportunity to employ advanced techniques in solving enterprise search problems. Key research questions in the area of actual search are outlined in the following section. Beyond the scope of the present paper, the value of high quality search may be greatly enhanced by intelligent delivery, possibly involving document synthesis, information extraction, and the tailoring of presentation to meet the interests, requirements and privileges of the searcher.

There may be great value in integrating enterprise search within the software applications used by employees to do their jobs.

3 Key IR research problems in the area of enterprise search

1. Defining an appropriate enterprise search test collection.
2. Effective ranking over heterogeneous collections characteristic of enterprises.
3. Building an employee portal - A distributed IR problem.
4. Effective search over collections of e-mail.
5. Estimating document importance for documents which are not part of a web.
6. Exploiting search context within enterprise searches.
7. Providing effective search over foreseeable future enterprise collections of interlinked continuous media.

Open Problems in Enterprise Search

The box contains a list of research problems arising within the area of enterprise search. It is no doubt incomplete and deliberately omits many problems which have considerable practical significance (for example extraction of text from binary formats and record-level security) but which seem amenable to reasonably straightforward engineering. In the rest of this section, the research problems are discussed in some detail.

3.1 An enterprise search test collection

The development of an enterprise search test collection sophisticated enough to model the whole range of interesting research problems and to serve as a benchmark by which algorithms may be tuned and improved and by which products may be compared, is a lofty goal whose achievement is likely to accelerate forward progress. Achieving it will require a great deal of intelligent observation of the data holdings of a range of organisations and, particularly, the range of information and service needs which could be satisfied if a suitably capable retrieval system were available. Looking at query logs on existing internal search systems is likely to be fruitless as search failures quickly discourage staff from attempting to use the system for purposes it is not capable of supporting.

To support investigation of open problems 2, 4 and 5, an enterprise test collection should ideally include a realistic combination of different data types, for a range of different enterprises. A hotch-potch collection of unrelated documents in the required formats will not do; the information contained should be naturally inter-related and it should be possible to obtain and/or synthesise realistic information/service needs over the data. Ideally, the test collection should contain complete sets of documents for a number of prototypical enterprises but there are obvious problems inherent in trying to obtain complete data for companies. Perhaps it would be possible to obtain complete snapshots of [failed] companies – e.g. Enron?

It seems quite feasible to abstract the enterprise search problem so as to focus on the interesting research questions while eliminating many of the practical difficulties. Researchers should be able to work on the relevant research problems without needing to acquire or implement adapters or filters for the whole gamut of proprietary databases, word processors, spreadsheets, content management systems and presentation packages. The obvious way to achieve this goal would be to pre-convert the proprietary formats into XML documents in such a way as to exactly preserve document structure and collection inter-relationships.

I envisage an enterprise test suite, consisting of at least three collections, each comprising the near-complete holdings of real (or at least realistic) enterprises plus a large set of corresponding realistic information or service needs. There would be of the order of 1-10 GB of data in each collection, divided into three parts: external websites, internal websites, and XML documents extracted from database, email, word processing, presentation, spreadsheet, and CMS files.

The searcher need statements could be similar in form to TREC[2] ad hoc topic statements with fields to specify the form and nature of the required results. To ensure generality of results, a wide range of different need types would need to be represented, each with sufficient examples to permit statistical validity.

The envisaged suite would not adequately represent personalised document generation but would provide enough complexity to be going on with! Even without personalisation, creation of such a collection will be an ambitious undertaking, fraught with legal and political difficulties.

[2] `trec.nist.gov`

3.2 Effective ranking over heterogeneous collections

Assuming it is appropriate to present results as a single ranked list, work is needed to solve the problem of effective retrieval across heterogeneous document types such as web pages, email messages, database records, spreadsheets, presentation slides and word processing documents[3]. Different types of documents differ considerably in degree of explicit structure (e.g. fields in database records), distribution of lengths (a database may contain records whose length is almost constant, while word processing documents vary considerably in length), presence of links, nature of the relationships between one document and another (e.g. websites), presence of repeated content (such as navigational elements in web documents) and the way in which language is used (the dot points in a set of presentation slides and the cells of a spreadsheet are likely to use language in a different way to the paragraphs in a letter. Application of a ranking function designed and tuned on one type of document only may result in a bias either for or against documents of that type. Past work has addressed the issue of how to tune retrieval functions to the length characteristics of particular collections (e.g. Singhal et al. (1995)) but has not addressed the problem of the heterogeneity introduced by very different document types.

One possible approach would be to divide the overall collection into relatively homogeneous subcollections, perform separate retrieval operations on each and to merge the results. Unfortunately, result merging is a difficult problem and seldom attains the performance of search over a single unified collection. Voorhees et al. (1995)

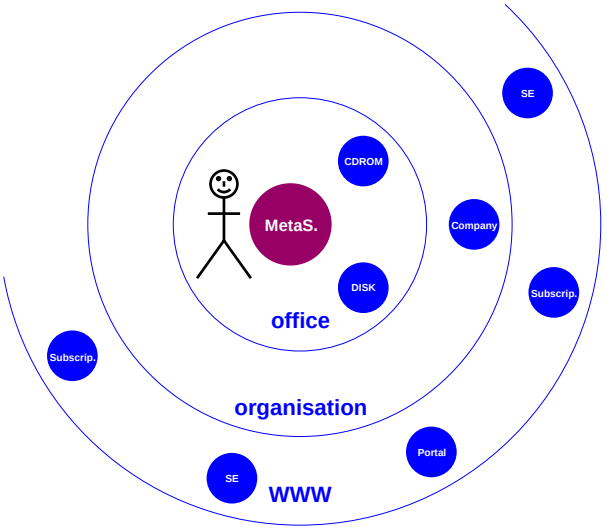

Figure 1: An employee within a company has a unique view of the information world, centred on their own computer and extending to information sources within their department, their company and the world. Building a personal portal enabling effective search over the entire collection is an interesting research challenge.

[3] The data may be suitably abstracted as noted in the discussion of an enterprise test collection.

```
To: boss
Cc: finance
From: minor_manager
Subject: Conferences
Date: 21 Oct 2003 12:07
I'd like to send Smith to ADC2004.  She's entitled under section whatever on
p.27 of the corporate manual.  Jones wants to go but she already went on
that junket to Maui.

ps. How was your break?
Attachment: ADC2004 CFP
```

```
To: minor_manager
From: boss
Subject: re: Conferences
Date: 21 Oct 2003 22:08
How much will it cost?

BTW: where's Smith up to on the project with CFP in Brisbane?
I had email from their CEO which poured a bucket load over our
original plan:

>From: CEO, CFP
>Date: 20 Oct 2003 0901
>Subject: re: your plan
>A load of rubbish if I ever saw one.  Who did you get to
>write it - the cleaner?   Where's the justification for the
>extra costs?
>
>No approval to proceed.

Auckland was great - lots of sailing and booze!
```

```
From: minor_manager
To: boss
Date: 22 Oct 0730
Subject: re: 1. Conference 2. Brissie proj
> How much will it cost?

$720NZ + MEL<->DUN rtn + 4 nights PD.

>>A load of rubbish if I ever saw one.  Who did you get to

Typical Bruce!!!  Got it all wrong.  See Smith's detailed cost estimates
-- intra/cfp/2003/cost_update.xls
```

```
From:
To: minor_manager
Cc: bigboss, hr, finance, denise.smith
Date: 27 Oct 2003 19:43
re: Denise
> $720NZ + MEL<->DUN rtn + 4 nights PD.

Approved.  Finance: use budget code 27.001.34
```

Query: Will anyone from R&D be in Dunedin in January?

Figure 2: A hypothetical email conversation between an employer and their superior illustrates several problems inherent in email search. The four messages and their attachments together tell the whole story of Denise Smith's proposed trip to Dunedin but another discussion is intermingled. Individual messages may not make good search results and may be hard to understand in the absence of the context of the overall thread. Identification of which messages actually constitute a thread is complicated by the evolution of subject lines. References to external documents exist but the target and "anchortext" of those references is often not clearly defined. A certain amount of metadata is present in the form of From:, To:, Subject: etc.

3.3 Building a personalised employee portal - A distributed IR problem.

It may be useful for a company employee to be provided with a search service which includes all (and only) the information sources he or she may access. These may include private files on a local hard disk, documents on a departmental shared hard disk, the heterogeneous corporate data referred to in 1 above, plus results from external sources. It may be appropriate to upweight or downweight results from certain sources. This problem is an interesting variation on collection selection and merging. See Figure 1.

One of the great benefits of hyperlinking with anchor text is the consequent ability to retrieve documents which are not internal to the enterprise and not actually indexed, using the descriptions provided by anchor text and target names. For example, an employee of BHP may be able to retrieve the home page of the London stock exchange or the US Geophysical survey on his/her own intranet search because pages on that intranet link to important external resources.

3.4 E-mail search.

Techniques for effective retrieval within a collection of email messages, taking into account threads of communication, the tendency of some people to quote some or all of earlier messages in a thread, the presence of attachments, structure within messages (subject, to, from etc.) is a problem currently not well solved See Figure 2 for an illustration of problems posed by the nature of the medium.

3.5 Estimating importance of non-web documents

A key lesson from Web search is the success of the "search-and-browse" paradigm where searchers submit broad queries (or the name of some entity) and rely on the search engine to deliver a list of key sites from where they can browse or search. Ranking by text similarity alone does not adequately support this paradigm. Much more satisfying results are achieved in web search if URL, link and anchortext evidence is used in ranking results. In general, such evidence is not explicitly available in fileshares, databases and email. Can similarly effective techniques be devised for non-web formats?

There is an increasing tendency for corporate websites to be generated from content management systems and for each generated page to include a set of automatically created navigational links which are repeated with variations across every page on the site. Link algorithms may need tuning to properly accommodate this phenomenon.

3.6 Exploiting search context

Many ambiguous web search queries (such as "restaurant") can be very effectively disambiguated with the addition of a small amount of information about the context in which the search was initiated. In the restaurant example, knowing that the searcher was in Carlton when they typed the query, would allow the search engine to return search results relating to restaurants within that suburb. Such results have a higher probability of being useful than those relating to restaurants in Morocco, London or Denpasar.

All sorts of factors may provide information allowing the search result set to be more appropriately ranked: geographical location, user profile (reading age, first language, interests, etc.), recent search history, and the nature of the task being performed. The key research question relating to search is how to extract and compactly represent the aspects of context which will make a beneficial difference when the query is processed, without unacceptably slowing the processing of the query.

3.7 Search of continuous media assets

Certain current organisations hold significant collections of continuous media, ranging from audio and video to time-series data. Techological advances make it likely that the number of enterprises holding such data in digital format will increase significantly. Proposals for a *continuous media web* (see `www.annodex.net/`) mean that such holdings may be hyperlinked in the manner of the current static web, leading to the need for effective search techniques over this type of data. Although this is likely to be a global Web problem it seems certain that there will be enterprise-specific dimensions, particularly for media organisations.

4 Search of enterprise webs.

The literature relating to enterprise web search is somewhat richer than for the more general enterprise search problem. Fagin et al. (2003) consider the important sub-problem of enterprise webs and note a number of key differences from the world wide web. In particular, they state the following "axioms":

1. *Axiom 1. Intranet documents are often created for simple dissemination of information, rather than to attract and hold the attention of any specific group of users.*
2. *Axiom 2. A large fraction of queries tend to have a small set of correct answers (often unique), and the unique answer pages do not usually have any special characteristics.* This seems to suggest that narrow queries submitted on the IBM intranet are not aimed at finding site entry pages, but discussion of this point doesn't make it clear.
3. *Axiom 3. Intranets are essentially spam-free.*
4. *Axiom 4. Large portions of intranets are not search-engine-friendly.* This suggests that there is considerable scope for improving search performance first by improving search algorithms but particularly by improving the presentation of the information so as to promote searchability.

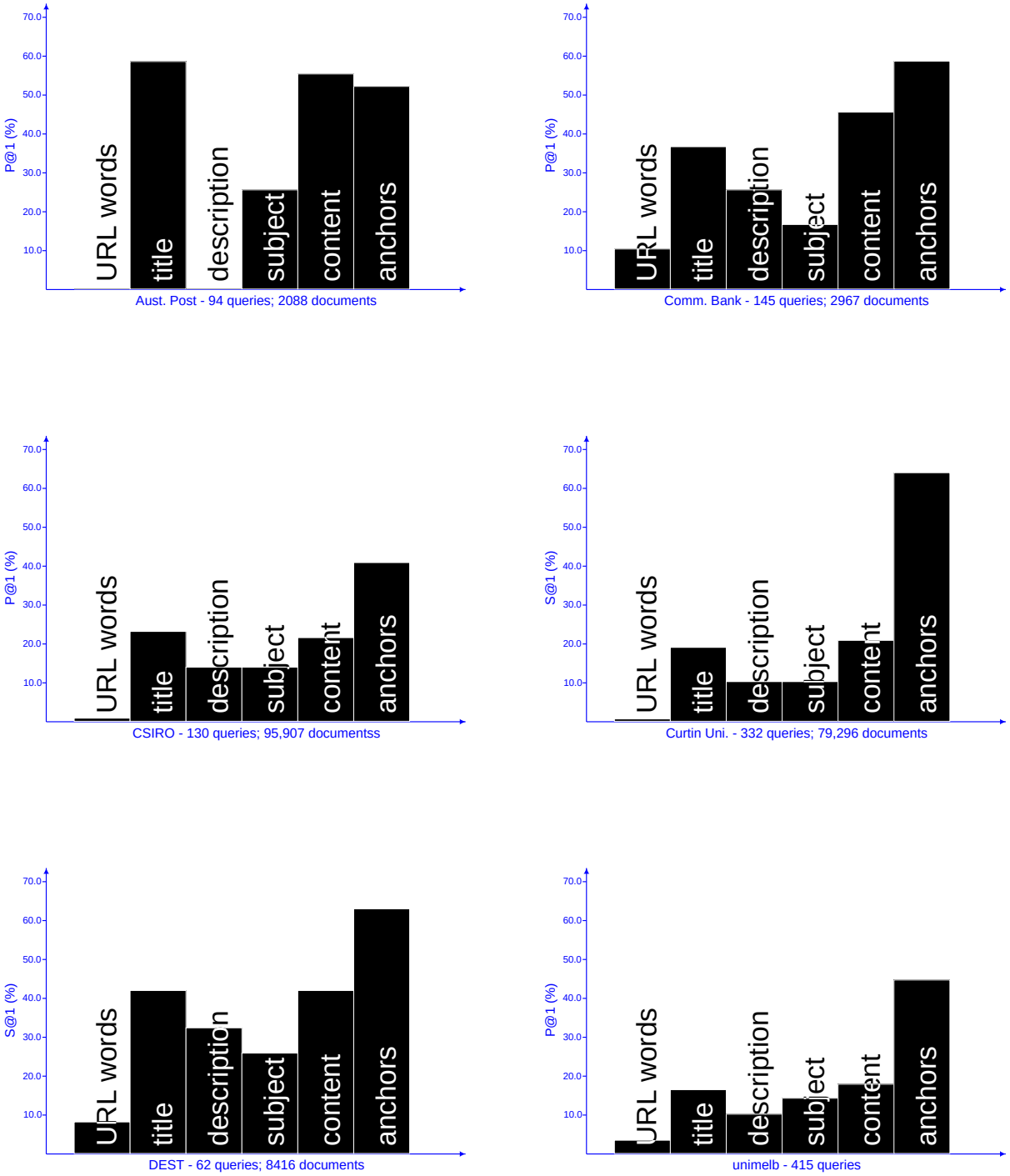

Figure 3: The relative value of different types of query-dependent evidence for navigational search on the external web sites of six different enterprises. Within a graph, the height of a bar reflects the effectiveness achieved on a navigational search task, when the Okapi BM25 scoring function is applied to pseudo documents containing only parts of the available text: content, title, URL words, subject and description metadata and propagated referring anchortext. In each case the sitemap from which the navigational queries were derived was excluded from the index. Exact duplicates of correct answers were accepted as equally correct.

4.1 Converting heterogenous enterprise data into a web

An obvious approach to solving the problems described in Sections 3.2, 3.4 and 3.5 is to convert the organisation's non-web data into a web format. Email messages may be converted into html documents using readily available converters such as hypermail[4]; Many word-processing packages allow documents to be saved in HTML[5] or XML[6] formats; and it is possible to provide web interfaces to many databases.

However, simple-minded format conversion will not by itself make a difference – unless the converted documents are interlinked and organised into "sites" in patterns resembling those of the normal Web, there will be no useful grist for the web search mill. The problems described in Sections 3.2, 3.4 and 3.5 still need to be solved, even if documents are converted into more web-friendly formats.

4.2 What evidence is most beneficial in enterprise search?

Craswell et al. (2001a) show dramatic benefits from the use of anchor text on a task requiring the location of site homepages within the Australian National University (ANU) web. Hawking et al. (2004) (in the present proceedings) conclude that link evidence from the external web is unnecessary for good performance in navigational search tasks on enterprise websites.

Upstill et al. (2003) investigate the value of query-independent evidence such as indegree, two variants of PageRank, and URL-type in homepage finding tasks on three test collections and the ANU intranet data. The methodology employed reranks content and anchortext baselines above both optimal and realistic cutoff points, using one of the query-independent dimensions in each conditions. Findings are: Indegree and PageRank provide improvement if the reranking cutoff can be chosen optimally; PageRank gives no discernable benefit over indegree for collections up to 18.5 million pages; URL-type reranking can bring substantial benefit on some collections even with realistically chosen cutoffs.

Fagin et al. (2003) use a new method called rank aggregation to combine rankings based on ten different types of evidence: title, anchortext, content, URL length, URL depth, words in URL, Discriminator (URL type), PageRank, indegree and position in the crawl order (based on the assumption that important pages will be discovered early in the crawl). They argue for computing uni-dimensional partial rankings of documents and combining them with a function which permits the large differences between one organisation and another to be taken into account.

Fagin et al test their methods using two sets of queries for which the experimenters laboriously located the correct answers (sometimes more than one) by using a combination of browsing and search using the incumbent search engine (not the system being studied)[7]. Queries were mined from the logs of the incumbent search engine. The first set comprised the 200 most frequently submitted and the second was a set of queries submitted with near median frequency. They called the effectiveness measure they employed *recall at position p*. This nomenclature is confusing as the measure does not correspond to recall as conventionally defined. In fact, the measure used is identical to *success rate at rank n* (abbreviated as S@n), the proportion of queries for which a correct answer is obtained by rank n.

Fagin et al were also able to note the relative contributions of the individual ranking factors to the overall combination and were amazed at the efficacy of anchortext, although this has been previously reported by others. They do not report evaluations of the effectiveness of individual rankings in isolation. In contrast to Craswell et al. (2001a) they found that the value of anchortext increased as n is increased.

[4] www.hypermail.org

[5] www.w3.org/MarkUp/

[6] www.w3.org/XML/

[7] All of the experimenters were employees of the organisation being studied.

4.3 A cross-enterprise experiment on the relative value of different types of query-dependent evidence.

This experiment compares the value of the types of query-dependent evidence considered by Fagin et al in their study, and also adds two more (subject and description metadata. The methodology, data and query sets are described in Hawking et al. (2004) (elsewhere in these proceedings). In summary, extensive sets of queries and correct answers are derived from the site maps of the organisations being studied. These were then processed (in effect) against six different indexes of the data, each indexing only a subset of available text, respectively: document content, document title, document URL, document subject metadata, document description metadata and referring anchortext propagated from all other documents in the collection except the site map. Documents were ranked using the Okapi BM25 Robertson et al. (1994) formula with common parameter settings:

$$w_t = tf_d \times \frac{\log(\frac{N-n+0.5}{n+0.5})}{2 \times (0.25 + 0.75 \times \frac{dl}{avdl}) + tf_d} \qquad (1)$$

where w_t is the relevance weight assigned to a document due to query term t, tf_d is the number of times t occurs in the document, N is the total number of documents, n is the number of documents containing at least one occurrence of t, dl is the length of the document and $avdl$ is the average document length (both measured in indexable words).

The effectiveness measure was S@n (the proportion of queries for which a correct answer was found by rank n).

4.4 Results and discussion

Figure 3 shows the S@1 results for six different enterprise collections ranging from 2088 documents (Australia Post) to 171,922 (University of Melbourne). In all cases except for Australia Post, anchortext is more effective than content on this task. There appears to be an approximately linear relationship between the log of number of documents indexed and the relative advantage (expressed as the ratio of the respective S@1 results) to anchortext over content. See Figure 4.

Hypothesis: In a collection with very few documents, the set of documents whose text fully matches

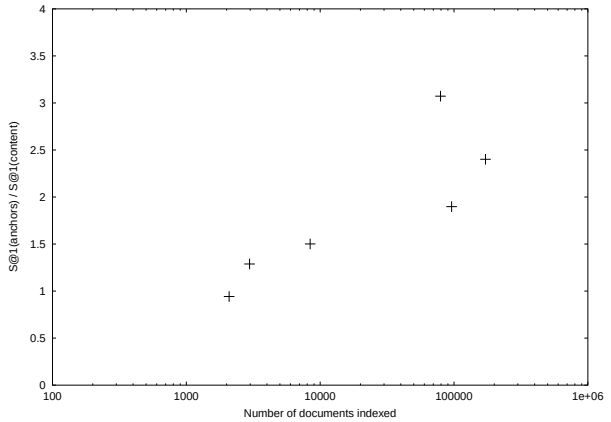

Figure 4: The ratio of anchortext effectiveness to content effectiveness plotted against collection size (log scale). The effectiveness measure is S@n.

a navigational query is relatively small and the probability that the desired site entry page is displaced by an incorrect page is low. Consequently, retrieval based on fixed text elements such content, title, metadata and URL words performs well provided that the element in question contains the query words. As the collection grows the number of pages matching the query is likely to increase and there is an increasing probability that the desired page will be pushed down the ranking. By contrast, as more pages are added to a collection, the number of links to site entry pages is likely to increase and the strength of the anchortext signal is likely to increase.

It is not surprising that the observed relationship is not exactly linear since the characteristics of the sites and their sitemaps are very different. In the case of Australia Post, the sitemap is presumed to have been generated by a content management system (Vignette) as all the URLs it references are Vignette-style URLs (e.g. `www.auspost.com.au/BCP/0,1080,CH3345~M019,00.html`) even though those pages constitute only a third of the crawl.

An experiment to test the hypothesis stated above, while controlling for differences between one organisation and another, would involve testing effectiveness over a series of partial crawls of the same large organisation, each successive crawl fetching more of the available pages and all of them large enough to contain the correct answers.

Between-enterprise differences in levels of performance observable in Figure 3 are likely to be due to differences in publishing practices between organisations. However, further study is needed to confirm this as no check has been made for errors in the sitemaps and it has not been confirmed that all the answers listed in the sitemaps are actually present in the crawls.

Results reported here essentially confirm those for ten other organisations, previously examined. Results for crawls of Australian Broadcasting Corporation, Australian National University, Microsoft, Robert Gordon University, Université de Neuchâtel, and National Institute of Standards and Technology are reported in Hawking et al. (2002) and results for crawls of US Government, RMIT university, Centers for Disease Control and Monash University have been presented in seminars but not published.

In the present set of experiments, the advantage of anchortext over document content on this type of task is much lower for Australia Post and Commonwealth Bank than in previous experiments. Indeed the Australia Post case is the first encountered where anchortext is inferior (to both content and title).

This could partly be because the present study is more rigorous than some of the earlier ones because the sitemap page from which queries and correct answers were derived was not indexed, meaning that its anchortext was not available. However, in four of the previous experiments (Australian Broadcasting Corporation, Microsoft, Robert Gordon University, Université de Neuchâtel) there was no need to apply this control as the test queries and their answers were not generated from sitemaps but rather manually with assistance from staff members of the respective organisations.

Handling of duplicate documents was improved in the present study compared to the previous ones: Duplicate documents were identified and removed, anchor text was redirected from an eliminated document to the surviving duplicate, and known duplicates of correct answers were also accepted as correct. This should have the effect of both raising the general level of scores and of improving the effectiveness of anchortext.

5 Summary and Conclusions

The problem of enterprise search is one of major economic importance. It is associated with a number of unsolved research problems including the six listed in Section 3. Progress in solving these problems would be accelerated by the creation of a suitable test collection. Such a collection would also permit tuning of algorithms and benchmarking of commercial search systems. Ideally, an enterprise test collection would reflect real-world complexity only in the areas of greatest research interest and would radically simplify the tedious engineering problems occasioned by proprietary document formats and complex security models.

One approach to achieving the gains achieved by Web search engines within the heterogeneous enterprise domain would be to convert all documents to web format. However, simple-minded conversion of documents and database records to HTML would not be sufficient as it would not create the links, anchortext, or site structure which underpin the success of Web search techniques.

The sub-problem posed by enterprise webs is an interesting one which has been studied by a number of authors. Here we consider navigational search within six different enterprise webs and compare the relative value of six different types of query-dependent evidence. In five out of six cases, anchortext is superior to each of content, title, metadata and URL words. The advantage to anchortext over content appears to increase approximately linearly with the log of collection size. A hypothesis is advanced to explain this but thorough testing is left for future work.

Acknowledgements

My understanding of enterprise search owes much to the shared knowledge and insights of my past

and present colleagues including: Peter Bailey, Nick Craswell, Francis Crimmins, the late Paul Thistlewaite, Trystan Upstill, Anne-Marie Vercoustre, Ross Wilkinson and MingFang Wu.

References

Mani Abrol, Neil Latarche, Uma Mahadevan, Jianchang Mao, Rajat Mukherjee, Prabhakar Raghavan, Michel Tourn, John Wang, and Grace Zhang. Navigating large-scale semi-structured data in business portals. In *Proceedings of the 27th VLDB Conference*, pages 663–666, Roma, Italy, 2001. `www.vldb.org/conf/2001/P663.pdf`.

Andrei Broder. A taxonomy of web search. *ACM SIGIR Forum*, 36(2), 2002. `http://www.acm.org/sigir/forum/F2002/broder.pdf`.

Nick Craswell, David Hawking, and Stephen Robertson. Effective site finding using link anchor information. In *Proceedings of ACM SIGIR 2001*, pages 250–257, New Orleans, 2001a. `www.ted.cmis.csiro.au/nickc/pubs/sigir01.pdf`.

Nick Craswell, David Hawking, Anne-Marie Vercoustre, and Peter Wilkins. P@noptic expert: Searching for experts not just for documents. In *Poster Proceedings of AusWeb'01*, 2001b. /urlausweb.scu.edu.au/aw01/papers/edited/vercoustre/paper.htm.

Ronald Fagin, Ravi Kumar, Kevin S. McCurley, Jasmine Novak, D. Sivakumar, John A. Tomlin, and David P. Williamson. Searching the workplace web. In *Proceedings of WWW2003*, Budapest, Hungary, May 2003. `www2003.org/cdrom/papers/refereed/p641/xhtml/p641-mccurley.html`.

Susan Feldman and Chris Sherman. The high cost of not finding information. Technical Report #29127, IDC, April 2003. `www.idc.com`.

David Hawking, Nick Craswell, Francis Crimmins, and Trystan Upstill. Enterprise search: What works and what doesn't. In *Proceedings of the Infonortics Search Engines Meeting*, San Francisco, April 2002. `www.infonortics.com/searchengines/sh02/02slides/hawking.pdf`.

David Hawking, Francis Crimmins, Nick Craswell, and Trystan Upstill. How valuable is external link evidence when searching enterprise webs? In *Proceedings of the Australasian Databases Conference ADC2004*, Dunedin, New Zealand, January 2004.

S. E. Robertson, S. Walker, M.M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of TREC-3*, November 1994. NIST special publication 500-225.

Amit Singhal, Gerard Salton, Mandar Mitra, and Chris Buckley. Document length normalization. Technical Report TR95-1529, Department of Computer Science, Cornell University, Ithaca NY, 1995.

Dick Stenmark. Method for intranet search engine evaluations. In *Proceedings of IRIS22*, Department of CS/IS, University of Jyväskylä, Finland, August 1999. `http://w3.informatik.gu.se/~dixi/publ/method.pdf`.

Trystan Upstill, Nick Craswell, and David Hawking. Query-independent evidence in home page finding. *ACM Transactions on Information Systems (TOIS)*, 21(3):286–313, 2003.

Ellen M. Voorhees, Narendra K. Gupta, and Ben Johnson-Laird. Learning collection fusion strategies. In *Proceedings of ACM SIGIR'95*, pages 172–179, 1995.