Does Topic Metadata Help With Web Search?

David Hawking
CSIRO ICT Centre, Canberra ACT 2601, Australia. E-mail: david.hawking@acm.org

Justin Zobel
School of Computer Science and Information Technology, RMIT, Melbourne, Australia. E-mail: jz@cs.rmit.edu.au

It has been claimed that topic metadata can be used to improve the accuracy of text searches. Here, we test this claim by examining the contribution of metadata to effective searching within Web sites published by a university with a strong commitment to and substantial investment in metadata. The authors use four sets of queries, a total of 463, extracted from the university's official query logs and from the university's site map. The results are clear: The available metadata is of little value in ranking answers to those queries. A follow-up experiment with the Web sites published in a particular government jurisdiction confirms that this conclusion is not specific to the particular university. Examination of the metadata present at the university reveals that, in addition to implementation deficiencies, there are inherent problems in trying to use subject and description metadata to enhance the searchability of Web sites. Our experiments show that link anchor text, which can be regarded as metadata created by others, is much more effective in identifying best answers to queries than other textual evidence. Furthermore, query-independent evidence such as link counts and uniform resource locator (URL) length, unlike subject and description metadata, can substantially improve baseline performance.

Introduction

Many organizations maintain public or intranet Web sites to support their activities. To facilitate use of these Web sites, site managers provide site maps and search interfaces. Although whole-of-Web search engines can be used to search individual sites, a local search engine can make use of organizational knowledge, can provide more frequent update, can additionally index non-Web content, and can provide appropriate search of internal information with complex access rules. In a well-maintained site, the pages are likely to have a consistent format, and in many cases contain metadata in addition to the text that is visible in a browser.

A range of arguments has been made in favor of inclusion of metadata. For example, it is claimed that it can be used to search in a uniform way across heterogeneous information sources. In the present work we examine the claim that *topic metadata*—subject and description metadata explicitly applied to a document by its author or publisher, such as Dublin Core `dc.subject` and `dc.description` metadata elements—can be used to enhance text retrieval. That is, we examine the claim that search in the presence of metadata can be more accurate and reliable than is otherwise possible.

Several issues are presented by topic metadata, such as:

- The difficulties of creating it consistently, of ensuring that it is accurate, and ensuring that searchers use matching terminology;
- The fact that metadata search requires matching the query against a possibly incomplete or inaccurate surrogate rather than the original;
- The fact that matches of query terms within metadata are invisible to users;
- The problems presented by metadata spam.

Despite these issues, some institutions have decided to invest significant resources in providing metadata on their Web site. For example, Anon University[1] maintains a large Web site created under policies that strongly encourage inclusion of metadata. The Web site is built on a document management tool that enforces uniform style and ensures that, for example, page content must be approved before being included. The search interface is designed to make use of metadata—pages with metadata are explicitly weighted more highly than pages without it—and academics with a background in document management oversaw the Web site implementation. This Web site is an instance where substantial, informed steps were taken to ensure that appropriate metadata would be present.

In this article, we examine the metadata on the `anon.edu.au` Web site and investigate whether metadata is useful for identifying the best answers to queries posed via a simple

Received December 20, 2004; revised August 9, 2005; accepted August 11, 2005

 • Published online 2 February 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20548

[1]An Australian university. We do not give the real name of Anon University, to avoid risk of embarrassment to Web publishers and authors.

search interface. As part of this investigation, we examine how much metadata is in place at Anon University, whether there is agreement between query and metadata vocabulary, and whether the metadata seems to accurately describe the pages to which it is applied.

Anon University is one of the larger Australian Universities and has a significant international profile. It employs several thousand staff and teaches tens of thousands of students. Its Web sites comprise hundreds of thousands of pages. Anon University was chosen for detailed study primarily because of its significant commitment to metadata publishing. However, our ability to conduct the study was critically dependent on being granted access to query logs and to a means of identifying the best answers for the selected queries. Fortunately, these resources were available for Anon University.

Our method for evaluating the usefulness of topic metadata in search extends the experiments described by Hawking (2004). In those and the present experiments, a large set of queries with corresponding known best answers was processed against the same crawl of an organizational Web site, each time allowing the retrieval system (held constant) to use a different part of the document, such as content, title, or subject or description metadata. Effectiveness scores computed for each run are used to determine the relative usefulness of different parts of documents in answering queries.

In Web and Web site search, it is particularly important to searchers that a search engine be able to return the best answer to a short query (for example, `www.microsoft.com` to the query "Microsoft") at the top of the list of results. The evaluation methodology we employed concentrates exclusively on this ability. Performance on each query is determined only by the ranking of the known best answer; It is not affected by the retrieval of other relevant (but less generally useful) pages. For example, in response to the query "library," we look only at the ranking of homepage of the main University library and do not give credit for retrieving any of the tens of thousands of pages that contain the word library.

In previous work, query sets and judgments were derived from the site map published by the organization in question. The judgments are thus made by the organization's own Web publishers who decide on the list of important topics and decide which are the best resources for those topics. For Anon University, approximately 40% of queries were derived by this means.

For Anon University, we also have a complete query log covering an extended period of time and have been able to obtain judgments for significant subsets, made by an employee of Anon University (not an author of this article). We were thus able to observe that findings relying on the site map method were broadly valid for real queries, but that the effectiveness of anchors was exaggerated.

The present experiments and analyses relate only to evaluating the contribution of subject and description metadata to the problem of locating the best answers to simple Web queries. We do not consider retrieval of items such as images, videos, or museum artifacts that are not text documents. Nor do we consider the use of metadata for nonretrieval purposes such as content management and preservation. Hunter (2003) provides an overview of research into broader uses of metadata.

We have not evaluated the potential usefulness of relational metadata such as author, publisher, and date in searching. If applied accurately and comprehensively, such metadata can be used to support scoped search. Participants in the HARD (High Accuracy Retrieval from Documents) Track[2] of TREC-2004 investigated retrieval taking into account familiarity, genre, and geographical metadata constraints as well as topical relevance. Comprehensive and accurate faceted category metadata can also be used to provide an integrated search and browse interface (Hearst et al., 2002).

Neither have we evaluated the potential benefit of metadata in presenting and organizing sets of search results. For example, in certain types of information-seeking tasks, it may be very useful to present internal metadata such as author, date, or description (for example, "This is Anon University's official handbook on Microbiology for 2001"). Furthermore, when the task requires the retrieval of many documents, it may be helpful to group the search results according to values of key metadata fields such as year, source, genre, or subject category.

With these caveats, our conclusions on topic metadata are straightforward. Answer ranking based on topic metadata was inferior to that based on other evidence, in particular visible content or anchor text. Despite the effort spent creating the metadata, the quality was low; quite how the quality might be improved with realistic resourcing remains an open question. Moreover, the failures in retrieval highlighted contradictions in the assumptions underlying metadata: it cannot simultaneously be specific, unambiguous, rich, and drawn from a limited vocabulary. Metadata that is appropriate for document management may well be inappropriate for search. It is not a solution to authorial individuality, and the cost of creating it institution-wide cannot be justified by the marginal search benefits observed in even the best cases.

Some of the sources of information that have been shown to be helpful in locating the best answers to Web queries—such as document uniform resource locators (URLs) and anchor text—are in fact forms of metadata. It is perhaps curious that many organizations pay little attention to these while investing heavily in topic metadata.

Background

Searches can be conducted across the whole Web, institutional Web sites, or just individual computers. Currently, there are only limited numbers of environments in which it is possible to search via metadata, as it is not sufficiently widely used. Here we confine our attention to an institutional Web site used for dissemination of nonconfidential information to the institution's stakeholders, and consider how metadata and search might interact.

Institutional Web Sites

Enterprises—including educational and government organizations as well as private companies—use Web sites for

[2] `ciir.cs.umass.edu/research/hard/`

DOI: 10.1002/asi

promotional purposes, for e-commerce, and for restricted-audience internal communication. Like other similar institutions, Anon University relies heavily on Web publishing to promote and conduct its research, teaching and community outreach activities, to attract students, and to enhance its image with the public and with allocators of funding. Web technology has become the dominant means for communicating materials to students such as class details, course notes, and lecture and examination timetables.

Such material is of considerable importance to individual students and to the university. Navigational aids and search facilities are integral to the Anon University Web site as they can increase the benefit derived from Web publication. If students can quickly find the information they need, they will be better prepared and better informed and will have less need to seek help from members of administrative or academic staff. The consequences of inaccurate, incomplete, or inaccessible course and timetable material are potentially serious and, accordingly, Anon University has adopted publication policies designed to ensure that its key Web materials are accurate, legally defensible, and recognizably official.

As noted in the introduction, policies adopted by Anon University require the creation of appropriate metadata prior to the publication of each official Web page. A major motivation for subject and description metadata is to improve search effectiveness, but publication details such as author, title, publisher, and date information are also required.

Intuitively one would expect that queries submitted to the `anon.edu.au` search interface would strongly reflect the purposes of the Web site and would overwhelmingly tend to be submitted by the obvious stakeholder groups. Given the large populations involved, one would expect that the vast bulk of queries would be submitted by current and prospective students seeking administrative and academic information. Casual inspection of lists of most popular queries seem to confirm this expectation. The most popular queries overwhelmingly relate to results, exams, short courses, graduation, employment, timetables, enrollment, and various academic subjects.

Metadata Standards

Much work has been done on the definition of metadata standards. The Dublin Core Directorate initiative defines a set of 15 core metadata elements (ISO Standard 15836-2003) that form the nucleus of many organization-specific metadata standards such as the Australian Government Locator Service (AGLS, National Archives of Australia, 2003). The Web publication system in use at Anon University typically inserts eight different Dublin Core metadata elements, such as `DC.Creator` and `DC.Subject`, into each HTML (hyper text mark-up language) page as attributes of `meta` tags. Much of this information is repeated in non-Dublin-Core form in the `title` element and in other attributes, such as `Author` and `Keywords`.

Metadata standards continue to develop. The RDF Resource Description Framework (Miller, Swick, & Brickley, 2005) is an XML DTD (document-type definition) designed to permit the description of resources including both Web pages and physical artefacts. According to Candan, Lu, and Suvarna (2001), RDF is intended to allow creation of metadata that can "be used by information access and integration engines to increase their efficiency and precision" (p. 7).

The *Semantic Web* (Berners-Lee, 1998) is a vision of a future organization of the Web in which the meaning of Web resources and their interrelationships is captured in the form of metadata descriptions. Advocates of the semantic Web believe that it will solve problems such as search ambiguity (Ding, van Rijsbergen, Ounis, & Jose, 2003); however, the semantic Web has not yet been widely adopted and there remains debate about what form it may eventually take. We hope that our study sheds some light on the subject.

Motivation for Metadata Encoding

Metadata has been embraced by organizations wishing to aid search and management of their online document collections. One such organization is Anon University, where the pages explaining how to create Web content discuss metadata at length. They state, for example, that "metadata tags facilitate precise retrieval,"[3] a point of view that is echoed on hundreds of other government and organizational websites. A typical example is:

> By using AGLS Metadata, it is easier for users to find the government resources they require. Quality Metadata provides reliable, detailed descriptions of the key concepts of a document or the key purpose of the service. By all agencies using the same Metadata standard similar items in different agencies will be described in a similar fashion. This makes it more likely that the search results will be sufficiently refined and at the same time exclude material that is not required. Additionally, quality Metadata records assist agencies to be sure that their users will find these relevant resources.[4]

The AGLS is another example of a large organization—in this case, the Australian government—making a considerable investment in the use of metadata. The primary aim of this metadata appears to be to assist searches. "The AGLS Metadata Standard is a set of 19 descriptive elements which government departments and agencies can use to improve the visibility and accessibility of their services and information over the Internet."[5]

Other governments have Web sites describing similar aims, as do many universities and other organizations. Curiously, some of these sites have poor internal topic metadata, or none at all! When accessed in December, 2004 a government guide to minimum Web site standards included no description

[3]A precise reference for the quote is omitted to preserve anonymity.

[4]Why Use Metadata? Retrieved December 9, 2004, from `www.agimo.gov.au/practice/delivery/checklists/metadata`

[5]AGLS, Retrieved December 2004, from `www.naa.gov.au/recordkeeping/gov_online/agls/summary.html`. This site is maintained by the Australian Government's National Archives of Australia.

DOI: 10.1002/asi

field and an unhelpful subject field: "Government publications; Government information."

Many of the claims in favor of metadata are predicated on the belief that full-text searches are frequently unsuccessful. For example, again considering an Australian Web site,

> Full text searches often return a very large number of results, many of which are not relevant. This can occur because: searchers usually prefer to search using a very general word or phrase with no concept of using advanced search functions; and free text search indexes consist of a "dumb" index of all words contained in a document regardless of the importance of the word/concept to the document.[6]

There are standards devoted to the problems allegedly presented by metadata-free search, such as the Dublin Core (`www.dublincore.org`) and the GILS Global Information Locator Service. "The Internet provides access to an amazing quantity of information. Internet-wide search services index hundreds of millions of Web pages. However, people cannot discover what they need unless the information is somehow organized."[7]

Although such claims are undoubtedly accurate, it is far from clear that, for full-text search, topic metadata is the right solution. Indeed, other forms of organization inherent in Web site structure and hyperlinking may already be sufficient. Paepcke, Garcia-Molina, Rodriguez-Mula, and Cho (2000) state that "similarity-based techniques . . . are increasingly being overwhelmed by the amount of data they are confronting," but note that link and popularity information can play a role as metadata.

Claims for metadata continue to be made in research articles.

> Automatic indexing based on a webpage's full-text has been widely used by internet search engines. However, automatic indexing techniques are most effective in a relatively small collection within a given domain. As the scope of their coverage expands, indexes succumb to problems of large retrieval sets and problems of cross-disciplinary semantic drift (Weibel, 1995). *It is obvious that automatic indexing techniques are not enough to handle internet information because the internet is huge, dynamic, and diverse.* These features call for a simple, compatible, and convenient internet information description standard to assist and facilitate automatic indexing of internet information effective [sic.] and efficiently. Since creators of webpages are usually not experts in information retrieval, in fact many are content specialists with only the technical skills needed to transfer content to this medium, an information retrieval standard for improving accessibility should be designed for use by web designers and publishers with varying backgrounds. The introduction of metadata may become such a standard.
>
> Metadata attempts to facilitate understanding, identifying, describing, utilizing, and retrieving internet information sources and their contents. [Italics added for emphasis.] (Zhang & Dimitroff, 2005, p. 692)

Zhang and Dimitroff (2005) created a set of artificially constructed Web pages and submitted them to a range of Internet search engines to measure the effect of Web page metadata on the visibility of that page in search engine rankings. They found that pages which only contained the query word in internal metadata were almost totally invisible in search engine results. This is, presumably, due to the use of spam rejection techniques by the search engines.

Zhang and Dimitroff found improved visibility when the query word was present in metadata as well as in title or content, but do not appear to have controlled for the consequent increase in overall term frequency, or for the fact that some search engines may give higher weight to term occurrences early in a document. They did not compare the importance of metadata query matches with the importance of Web evidence (such as hyperlink in-degree, PageRank, and URL structure) known to be heavily relied on in result ranking by many Internet search engines, such as Google (Brin & Page, 1998). It is not clear that they controlled for these effects in their study.

Agosti, Crivellari, and Melucci (1999) were among the first to test the value of metadata in search. Using 15,166 Web pages from the Library of Congress, they observed that content alone was slightly more useful than content plus metadata, but that title plus keywords was more effective still. As confounding factors, the baseline retrieval mechanism was poor (a simple form of cosine measure) and only a small fraction of the Library of Congress pages had metadata.

There is a wide literature that cites the desirability of metadata. For example, Desai (1997) states that, on the Internet, "search and discovery would become difficult without some well thought out discovery mechanism built around adequate metadata" (p. 192) and argues that metadata is needed to support retrieval by content. (The examples in this article, intentionally or otherwise, highlight the difficulties of use of metadata: the need to develop expert systems to gather the metadata, the complex range of information that such metadata standards require, and, perhaps most surprisingly, the fact that the suggested metadata for this article does not include the term "metadata.")

Candan et al. (2001), as justification for Resource Description Framework (RDF),[8] assert that users "increasingly find it difficult to retrieve relevant information" on the Web and that, because "the heuristics used in the search process are not perfect, search engines and other information access tools cannot provide highly efficient access to information on the Web (p. 6)." The implication is that metadata will rectify these issues.

The Semantic Web is expected by some authors to improve document retrieval. For example, Newby (2002)

[6]"AGIMO—Metadata," `www.agimo.gov.au/practice/delivery/checklists/metadata`, maintained by the Australian Government Information Management Office.

[7]"Overview—ideas behind the GILS approach," `www.gils.net/about.html`. The metadata on this page too was strangely lacking, consisting of a small number of words drawn from a toolbar of links. It added little to the visible content.

[8]A W3C Semantic Web activity; see `www.w3.org/RDF/`.

DOI: 10.1002/asi

writes that "the Semantic Web will let IR systems generate a candidate set of documents from all those known based on exact and unambiguous criteria" and will provide "a match between the information need and the actual content of a document"; Newby notes that such search can only work once ambiguities are removed from the metadata, but regards such problems as solvable by further research.

Ding et al. (2004) describe Swoogle, a search engine for the Semantic Web, and propose a measure for estimating the importance of individual documents, but do not attempt to evaluate the benefit of Semantic Web metadata in retrieval.

The value of metadata in retrieval has been questioned. Marshall (1998) identified a range of problems associated with metadata, such as the difficulties of keeping it consistent and up-to-date, and noting that different user groups use it in different ways; for example, document metadata that is appropriate to users accessing it from within an organization may be inappropriate for users who access it from outside. Doctorov (2001), in a sardonic rather than academic perspective, raises similar issues: that people won't take the time to create such data when the benefit is obscure, and that different people describe the same thing in different ways.

However, there has been little other exploration of the quandary of metadata for search: that it needs to be rich and semantics-laden, with precise expressive power, yet must be unambiguous. Standardization of formats does not address this issue. Given the different meanings that different people bring to words, it seems plausible that the problems presented by authorial individuality are not formally solvable.

Prior Work on Metadata and Enterprise Search Effectiveness

Wilkinson et al. (1991) compared indexing methods in processing 50 natural language queries against a relatively small set of government press releases. Contrary to expectations, they found that retrieval based on manual assignment of indexing terms, both with and without use of a controlled vocabulary, was substantially less effective than when automatic full text indexing was employed.

Stephenson (1999) studied the effectiveness of metadata on the U.S. Environment Protection Agency's (EPA's) public access Web site in answering 24 known-item queries formulated from real reference questions posed by members of the public to EPA librarians. Only eight queries retrieved a responsive answer when the metadata repository alone was searched, compared with 22 for full-text search. Precision was also found to be higher for full-text search.

Using several specific criteria, Sokvitne (2000) assessed the quality of title, author, publisher and subject metadata in web documents published by 20 government and educational organizations known to practice metadata embedding. He concluded that:

> There seemed to be a fundamental and consistent misunderstanding or lack of awareness and training in what the DC.Subject field is for and how it should be used effectively. On current indications, an increase in metadata records would not improve the recall/precision ratio exhibited by freetext search engines, but merely duplicate it. Strategic decisions to adopt DC or AGLS will not provide any return to the organizations and sectors involved unless there is an accompanying development of the policies and skills to populate the DC.Subject element.

and that

> the Dublin Core standard will have questionable value as a discovery tool unless the elements are able to be populated and used correctly.

Smith (2002) studied the effect of metadata presence on "Web Impact Factors" (Ingwersen, 1998) for both electronic journals and New Zealand university Web sites. In both groups the prevalence of metadata was quite low. In the latter case, Smith reports a small positive correlation between metadata prevalence and Web impact factor, in the former, there was a small negative correlation.

Abrol et al. (2001) and Fagin et al. (2003) assessed the value of various types of evidence in ranking enterprise web search results but did not include topic metadata.

Drott (2002) examined 60 websites operated by companies included in the Fortune Global 500 list for the presence of "keywords" and "description" metatags and found that they were present in only about a third of cases.

Hawking (2004) compared the relative contributions of subject and description metadata, title, content, URL words, and referring anchortext to navigational search within the Web sites of six different organizations. In each case, both subject and description metadata were found to be less useful than title, content, or anchortext. Relative to an investigation of the usefulness of metadata in search, there were two limitations of this previous study: first, it was not known how seriously committed to metadata were the six organizations studied; second, queries and corresponding best answers were derived from the site map of the organization in question and consequently may not correspond to queries actually posed.

Hawking's experiments used site map entries as queries, a strategy that has the advantage of eliminating the need for explicit relevance judgements. However, with site map entries the same organization is creating both queries and data, reducing the likelihood of vocabulary mismatch.

For the `anon.edu.au` Web site we have a complete query log, of queries presented both internally and externally. This log, and relevance judgements made by Anon University staff, allow us to undertake a fresh evaluation of the relative importance of different components of Web pages for retrieval, as well as to evaluate whether the previous use of site map entries led to realistic results.

Experiments

Our experiments had several aims: to investigate the value of metadata in supporting searches, to evaluate the

DOI: 10.1002/asi

contribution of individual document components to effectiveness, and to identify whether the use of site map entries as queries is a valid strategy for determining effectiveness.

We therefore had to gather data and segment it into fields, gather queries, and undertake relevance judgements. We used multiple sources of queries and multiple aspects of the data, as described later.

Document Set

The data used for most of these experiments comprised the Web sites of Anon University. The complete `anon.edu.au` Web site has two components: a *managed* site that has been carefully organized into a hierarchy, with approval processes as outlined above, and a large number of *uncontrolled* pages maintained by individual departments that are outside the managed site. Access to some of the pages—both managed and uncontrolled—is restricted to internal users only; the others are visible to the external Web.

We collected our data with an external-view crawl commencing on April 15, 2004. The number of pages was 285,051, after elimination of duplicates. Subsequent analysis revealed that 104,089 of them came from a dynamically generated calendar site, `www.calendar.anon.edu.au`, which happily supplies information for days or months (such as January 6194) for which no events have yet been scheduled. These pages seem to have no subject, description, author, or publisher metadata and have been excluded from our calculations of metadata prevalence. (The Google query `site: anon.edu.au` estimates that there are 204,000 Anon University pages in the Google index but only four from `www.calendar.anon.edu.au`.) We found 102,247 pages (out of 285,051 − 104,089 = 180,962, that is, 56.5%) with subject or keywords metadata in our external crawl.

Retrieval Mechanism

The Anon University dataset was fetched and indexed using the Panoptic v5.2 search engine.[9] Query terms occurring in metadata fields, content, titles, and anchor text—the words highlighted in a browser to indicate a Web link—are indexed separately from each other, allowing queries to be processed using only a single field or a combination of fields.

The Panoptic crawler records a file of URL redirects, which enables anchor text and link counts to be applied to the page actually stored by the crawler. Anchor text is associated with the target of a link, but is also counted as part of the content of the source document. The crawler records a redirect when a fetched page is rejected because it is a near duplicate of another already stored. The indexer also notes onsite and offsite link counts and URL length for each indexed page.

[9] See `www.panopticsearch.com`. Note that one of the authors, Hawking, is the leader of the Panoptic development team.

Task and Measures

The retrieval task modeled was that of finding the known best answer (or one of the best answers) to each of a long series of queries. It is a form of *known-item search* task. In each run, the first 10 documents retrieved were examined and the rank of the first best answer (or a duplicate URL) was recorded. Each run produced a list of queries and ranks. For example, in one of the runs we observed the following right-answer ranks for six queries:

```
admission → 2
security → 1
multimedia → 1
event management → 1
semester results → 2
postgraduate studies → 11
```

A rank of 11 means that no correct answer was found in the first 10 results. To perform well on this task, a retrieval system must be able to distinguish the best resources on the topic from those which contribute limited or very specialized information.

From these ranks we computed three measures:

1. P@1. The proportion of queries for which a best answer occurred at rank 1.
2. S@10. The proportion of queries for which a best answer occurred in the first 10.
3. MRR1. The mean reciprocal rank of the first best answer.

Of these measures, P@1 and S@10 correspond directly to user experience. However, MRR1 takes into account more ranking information, is more stable, and is in our experience a better basis on which to compare systems and runs. We used the Wilcoxon signed rank test for statistical significance tests, which takes into account the same information as MRR1.

In addition to the best answer to each query, there are typically many Web pages published by Anon University that are in some way relevant. For example, 4500 hits (3890 on Google) are reported for the query "security," including material on Cisco and Oracle, security audits, copyright on the Internet, secure installation of FTP, social security benefits, the political situation in Jakarta, course offerings in security management, and Computer Science technical reports. However, it is unlikely that people seeking information about the political situation in Jakarta would do so by posing the query "security" to the main search interface on `anon.edu.au`. The people who submit that query are most likely to be students or staff wishing to contact the University's security service or to obtain access to buildings—and pages on this topic are the most likely to be managed pages in the university Web site for which appropriate metadata has been created.

If users were looking for information about research in computer security, they would probably use a more specific query or submit the general one to a more focussed search

DOI: 10.1002/asi

TABLE 1. Forms of evidence and corresponding scoring methods used in our experiments.

Mode	Evidence used	Scoring function
URL words	Resolved URL	BM25
Content	Visible words (excl. title) plus image tags	BM25
Title	Words in <title> or DC.title	BM25
Subject	Keywords, subject or DC. subject metadata	BM25
Description	Description or DC.description metadata	BM25
Anchors	Combined incoming anchor text	BM25 / AF1

service such as an index of Computer Science technical reports.

Query Processing

In our experiments, we used a testing script to submit queries via a Web interface to the search engine, collect result pages via the Lynx browser,[10] and extract the result links. Result URLs were canonicalized in the same way as the recorded best answers, for example, by removal of port numbers and default page names from the end of the URLs.

To cope with the fact that the best answer for a query might be published under several equivalent URLs (for example, `xyz.com`, `www.xyz.com`, and `xyz.com/index.html`), without manual judging, each best answer was looked up in the crawler redirects file. If found on the left hand side of a redirection rule, the right hand side of that rule was added to the list of best answers. A search was considered successful if it returned any one of the equivalent URLs and the rank of the first such occurrence was recorded. Note that the redirects file records cases of duplicate content as well as redirections.

The testing script was given access to the search engine configuration file, allowing it to control how queries were processed. Table 1 documents the primary query processing modes. The Okapi BM25 relevance scoring formula is due to Robertson, Walker, Hancock-Beaulieu, and Gatford (1994) and is widely used in TREC and other information retrieval experiments. The version used here has not been specifically tuned for the data and is as follows:

$$w_t = tf_d \times \frac{\log(\frac{N - n + 0.5}{n + 0.5})}{k_1 \times ((1 - b) + b \times dl/avdl) + tf_d} \quad (1)$$

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 document d, 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. The length of a document is a count of the number of indexable content words in the particular field or combination of fields being indexed.

(Negative values of log $(\frac{N - n + 0.5}{n + 0.5})$ are mapped to a small positive constant ε in experiments reported here.) Following Robertson et al., we used the parameter settings $k_1 = 2.0$ and $b = 0.75$.

Results are presented for two different query processing modes.

> *TextOnly.* The relevance scores used in ranking are derived only from the text of the appropriate fields. The scores shown correspond to the Okapi BM25 scores, which would be achieved if documents in the collection were replaced by surrogates consisting of only the type of text (such as anchor text, or content + title) being considered. No weightings are applied when multiple fields are combined (that is, term occurrences count equally regardless of whether they were in the title or the body of a document.) The *WebMix* bar takes WebEvidence into account and is shown as a comparative yardstick.
>
> *WebEvidence.* Inclusion of the WebEvidence conditions allows comparison of the value of query-independent evidence (link counts and URL length) with the value of metadata. The single TextOnly scores derived as described above are normalized on a per-query basis, then linearly combined with scores derived from URL length, offsite indegree, and onsite indegree. Coefficients in the linear combination are held constant across all conditions. Anchor text is scored using the non-BM25 anchor text weighting (AF1) described by Hawking et al. (2004).

The *Webmix* scoring function in the figures records the default behavior of the search engine in use, and was a linear combination of a content, title, and anchortext score with URL length and onsite and offsite indegree. It is a WebEvidence condition in which anchortext is taken into account along with title and content. Although it is not a text-only condition, it is included in the TextOnly figures to facilitate comparisons.

Scoring functions were tuned neither to Anon University data nor to the specific metadata fields. It is quite possible that overall performance levels could be improved by tuning to Anon University data. However, the retrieval system can be seen to be working quite well as the levels of performance achieved using all available evidence are quite high, exceeding $MRR1 = 50\%$, (that is, the best answer is found at rank 2 on average) on all but the randomly chosen query set.

It is possible that results for each particular form of evidence could be improved by individual tuning or by choosing a ranking model specific to the type of data (such as anchor text). However, we suggest that results obtained from the use of exactly the same well-performed ranking function across all the types of evidence provide the most useful baseline, and leave subsequent tuning to future work.

We did repeat the experiments using the same document length (the number of indexable words in the document's visible content) regardless of what types of evidence (e.g., title or

[10] `www.lynx.browser.org`

DOI: 10.1002/asi

anchortext) were being used to process the queries and found that the resulting changes in performance were very small.

Site map queries. The main Web site of many organizations includes a site map in the form of an alphabetized list of named links to key subsites. In the case of Stanford University, for example, the site map includes more than a thousand entries, starting with "Academic Calendar," "Academic Computing," and "Academic Council Senate." Each entry is a link to what the authors of the site map believe is the most useful resource on that topic.

Such a site map can be used as a source of queries with corresponding best answers, and, as discussed earlier, has been used in this way in previous work by Hawking (2004). From the point of view of IR research, it is a considerable advantage that both the topics and the best answers are chosen by staff of the organization being studied and that the list is created to assist visitors to the site rather than artificially for an experiment.

We downloaded the main site map page for Anon University plus the entry pages for 10 top-level subsidiary sites such as "Staff," "Students," "Careers," and "News," and extracted links and labels (anchors) from both the main content pane and the site map pane. If the label was an image, the image `alt text` was used as the label. Labels which were empty or were more than 60 characters were rejected; the rest were used as test queries, using the corresponding link target as best answer. We subsequently rejected 22 queries for which there was no known best answer within the crawl.

By this means we extracted 187 *site-map* queries from the site map. Here are some examples of site-map queries.

```
2003 research publications collection
a-k
about
about Anon University
about the library
academic policy academic registrar
administration
admissions
alumni and visitors
```

An advantage of the site-map queries is that it should be safe to presume that all the identified best answers lie among the managed content. It is possible that, for queries taken from query logs, some best answers identified by our judges may lie outside the managed content, which could lead to a bias against metadata; official metadata is less often present in unmanaged pages. On the other hand, given that the site-map and the pages have (broadly speaking) the same authorship, the match in terminology between the site-map and the pages it indexes is likely to be better than the match between queries and pages sought by a user.

Top101 and MediumFreq queries. The managed pages on the Anon University Web site can be searched with a standard text query box provided on all of the managed pages. We have a log of 848,837 query submissions posed through this interface, covering approximately 2 years. The most popular 1000 queries account for 228,203 query submissions (approximately 27%). The frequency of submission of these queries is plotted in Figure 1. The distribution is approximately Yule–Pareto–Zipfian.

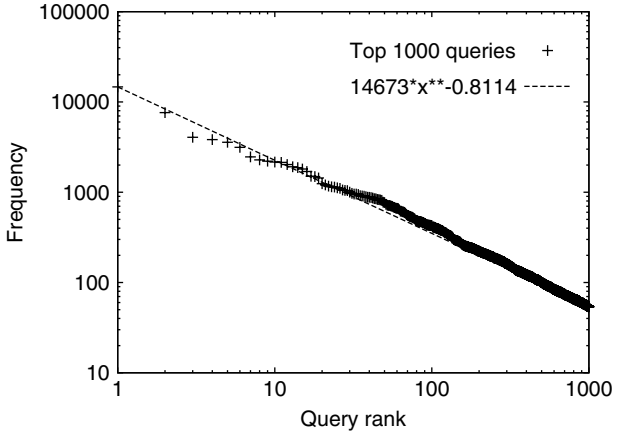


FIG. 1. Distribution of query submission frequencies for the most popular queries submitted to the main search interface at Anon University. Note the use of logarithmic scales.

We wished to compare the performance of search methods on both popular queries and others because the site may have been constructed with the most popular queries in mind, and thus effectiveness on these queries may not be representative of effectiveness overall. We chose the 101 most popular queries (Top101, frequency range 427–14,673) and also queries ranked 901–1000 (MediumFreq, frequency range 54–60).

For each query, we used Anon University staff to identify best answers, working on the assumption that the searches were made with internal knowledge and that a particular, known page was being sought. Experience with other University sites shows that the great majority of searches are made internally; most of the queries themselves appear to be directed to the finding of key pages rather than part of a general search for information. A page was judged to be relevant if it was the obvious or best starting point for browsing to answers to the query.

Random queries. One hundred queries were randomly selected from the query log. Because of the nature of the distribution, the queries chosen were low frequency queries (frequency range 1–131; mean 7.32). Some contained spelling errors (such as "englihs") and some contained errors such as mismatched quotes. Our judge was only able to identify best answers for 75 of these queries; the rest were eliminated.

Summary of Experimental Conditions

The design of the main experiment is an $8 \times 2 \times 4$, Evidence-Type $\times$ Processing-Mode $\times$ Query-Set Matrix. Table 2 lists all the conditions. Results for the main

DOI: 10.1002/asi

TABLE 2. Main experiment: Summary of conditions.

Evidence types	Processing modes	Query sets
URL words	Text only	Top101 (101)
Title	Web evidence	MediumFreq (100)
Title + content		Rand (75)
Description		Sitemap (187)
Subject		All (combined query sets)
Subject + description		
Cont. + title + desc. + subj.		
Referring anchor text		

TABLE 3. Examples of different sets of queries extracted from the query logs.

Top101	MediumFreq	Rand
results	admission	CONVEYANCING
exam results	Security	steve mckillon
short courses	Multimedia	uni news
graduation	Event Management	administration and reception
student results	subject	mr rodney jane
result	semester results	FS321
jobs	rat	040966C
employment	portal	Certificate Engineering

experiment are reported in Figures 2 and 3. Table 3 shows examples of the different types of query.

As described below, two follow-up experiments were also conducted using the same eight evidence types and the same two processing modes, but each involving only one query set. The first (Figure 4) investigates generalizability to a completely distinct (governmental) organization and 99 queries constructed by a resident of the jurisdiction. The second (Figure 5) uses only the main Web site at Anon University and the subset of queries whose best answers are found on the main Web site.

Experimental Hypotheses

We hypothesize that:

- Many of the common queries cannot be effectively resolved using metadata.
- Even when the document set is restricted to those documents that have appropriate metadata fields, search using metadata only is ineffective, and addition of metadata to the other text does not improve effectiveness.
- There is low intersection between the metadata vocabulary and the query vocabulary.
- Poor metadata maintenance is prevalent, as for example demonstrated by common use of strategies such as metadata copying, poor choice of author names, and poor choice of descriptive terms.
- The metadata is often misleading.

In addition, we explore the value of each of the forms of evidence.

Results

Figure 2 shows search effectiveness results for the combined Anon University query sets based on different types of evidence and for three different measures. The bottom row shows results for the TextOnly mode of processing while the top row of results factors in Web evidence not related to text content. The following observations may be made on the TextOnly set:

- No combination of textual evidence scored using the Okapi BM25 formula reached the performance of the Webmix

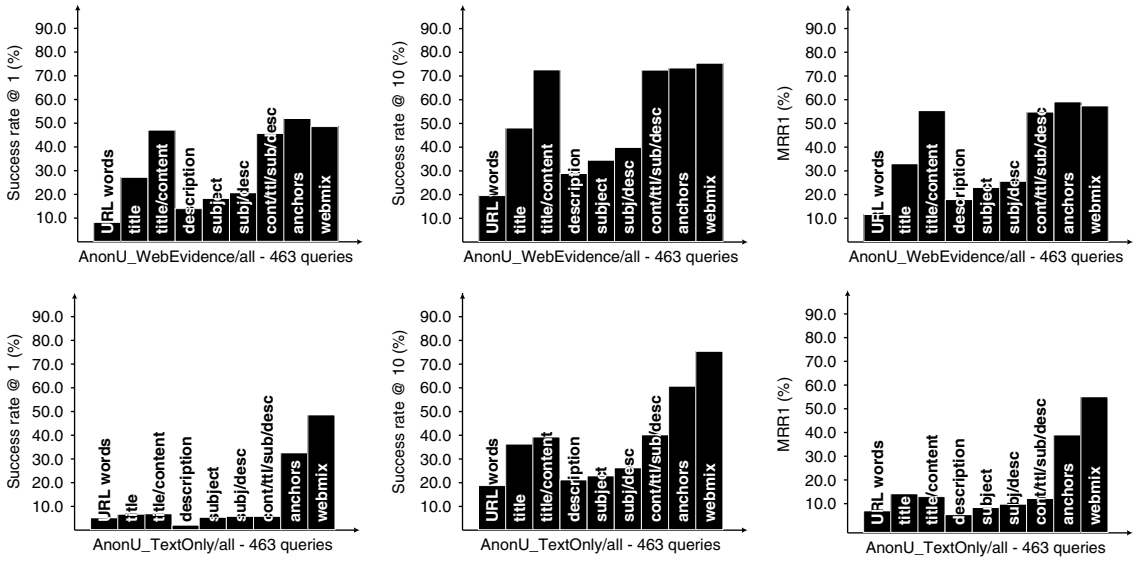
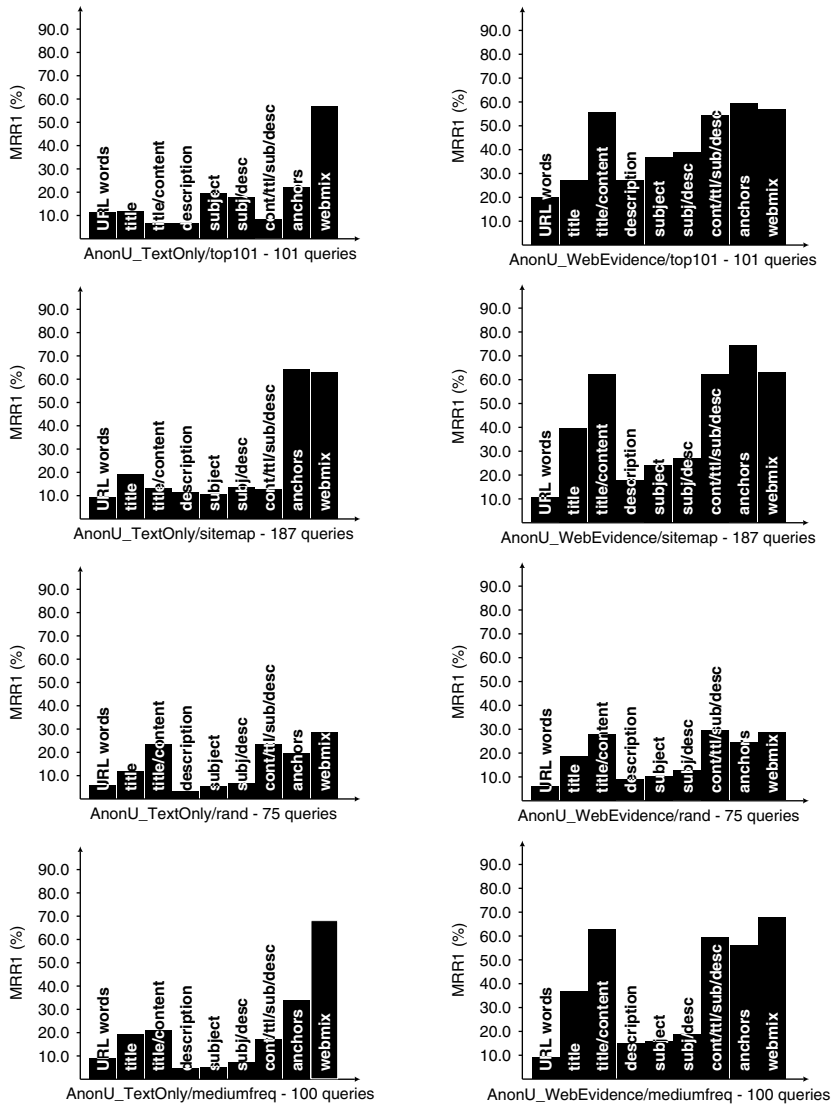


FIG. 2. Results for all Anon University queries combined, using three different measures.

DOI: 10.1002/asi



(a) TextOnly (b) WebEvidence

FIG. 3. The relative value of metadata evidence in identifying best answers to queries relevant to Anon University. MRR1 (mean reciprocal rank of the first occurrence of the best answer) is the measure used. Each graph within a column corresponds to a different query set; each bar in a graph shows the performance for a particular combination of evidence.

yardstick. Webmix outperformed anchors on MRR1 by 39%. The difference was significant ($p \leq 2 \times 10^{-12}$).

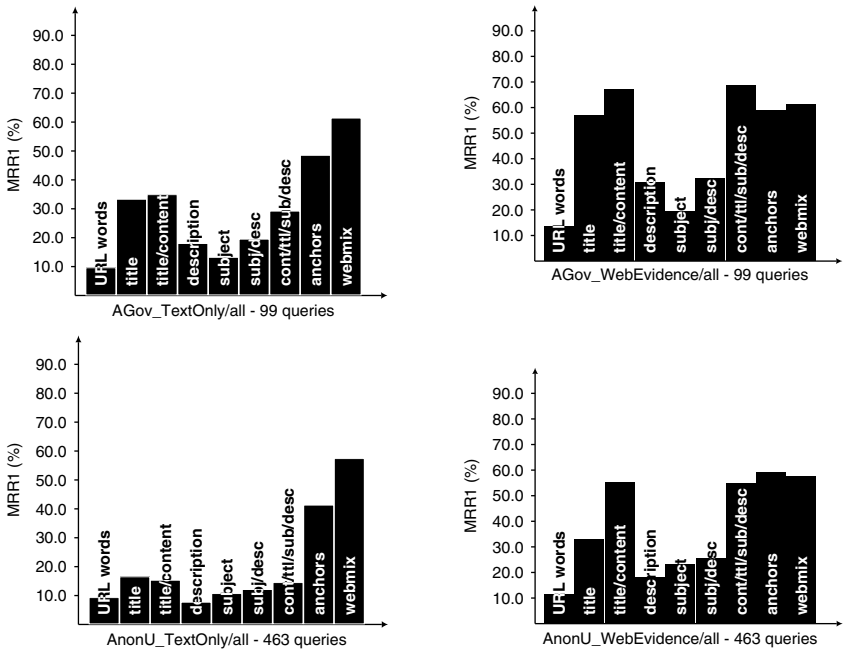
- Anchortext was 2.5 times as effective on MRR1 as that of the second best text-only source of evidence (title) ($p \leq 2 \times 10^{-26}$).
- The best metadata-only combination (subj/desc) was significantly outperformed both by anchortext ($p \leq 6 \times 10^{-31}$) and by title ($p \leq .0005$).
- The best metadata-only run (subj/desc) failed to find an answer within the top 10 results in 74% of cases, compared with 25% for Web mix.
- Addition of metadata to title/content caused an apparent deterioration in performance, but the difference was not significant.
- The advantage of anchors over other forms of text evidence is greatest on the P@1 measure.

Considering the WebEvidence plots (top row):

- The average MRR1 for the WebEvidence conditions excluding Webmix is 0.3485, more than double the corresponding TextOnly average (0.1571).
- The difference between Webmix and anchors conditions is not significant.
- The difference between anchors and title/content conditions is not significant.
- The best metadata-only condition (subj/desc) was significantly outperformed by both anchors ($p \leq 2 \times 10^{-33}$) and by title/content ($p \leq 2 \times 10^{-31}$).

Comparing WebEvidence against TextOnly plots shows that query independent evidence can contribute strongly.

DOI: 10.1002/asi



(a) TextOnly (b) WebEvidence

FIG. 4. Comparison of MRR1 results obtained at Anon University with those obtained from a crawl of the "AGov" government jurisdiction. The AGov query set and the corresponding correct answers were generated by a resident of the jurisdiction in question.

Combining query-independent evidence (link counts and URL length) with a title + content baseline triples MRR1. In sharp contrast, adding subject and description metadata to the same baseline causes an apparent drop in performance.

In summary, subject and description metadata performs worse than all other forms of evidence examined, other than the text of the URL.

Validation on Data From Another Organization

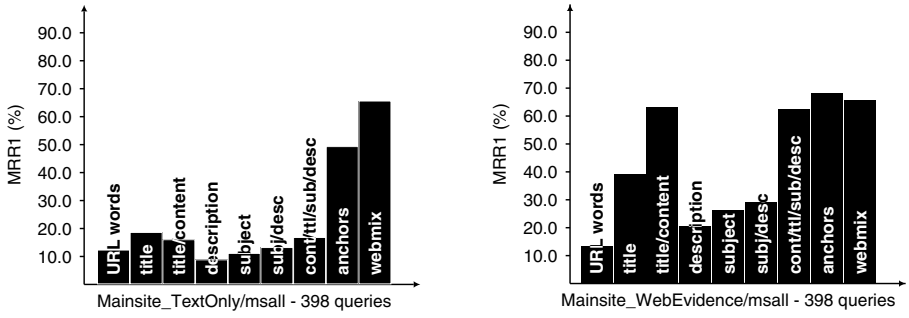
To confirm that our findings were not specific to Anon University, we repeated the same study with almost identical methodology for the set of websites published by a particular government jurisdiction (AGov, an Australian State/Territory government).

AGov also has a commitment to metadata tagging and operates many distinct Web sites. The sites were crawled in the same way as those of Anon University and a total of 174,000 pages were fetched. In this case, we didn't have access to query logs or to a comprehensive site map. Accordingly, a set of 99 queries with corresponding best answers were compiled by a resident of the jurisdiction.

Performance levels might be expected to be higher on this collection because the person compiling the queries browsed the AGov Web sites extensively while choosing and validating queries.

Figure 4 presents the AGov results. They confirm the main findings on the Anon University data:

- Anchortext is substantially more useful in finding best answers than any other text-only measure.
- The best metadata-only combination (subj/desc) was significantly outperformed by title/content, both in the TextOnly ($p \leq 8 \times 10^{-5}$) and in the WebEvidence ($p \leq 5 \times 10^{-8}$) cases.
- The addition of Web evidence brings a substantial improvement (69%) in MRR1 averaged across the evidence-types (excluding webmix).



(a) TextOnly (b) WebEvidence

FIG. 5. MRR1 results obtained for the main Anon University Web site only.

DOI: 10.1002/asi

- In the TextOnly condition, the addition of subject and description metadata to title/content causes a significant deterioration ($p \leq 0.0005$) in effectiveness. In the WebEvidence case, the same change appears to cause an improvement, but the difference is not significant.

The main contrast between the institutions is that, in the Agov WebEvidence case, cont/titl/subj/desc significantly outperforms the use of anchors ($p \leq 0.01$).

Effect of Query Type

Each row of graphs in Figure 3 shows the MRR1 effectiveness results for one of the four different Anon University query sets. The left hand column shows the TextOnly condition and the right shows the effect of adding nontextual Web evidence.

If one set of queries were significantly easier than the others we would expect to see a large difference in the mean heights. If one type of evidence were more suited to a particular type of query we would expect to see different ratios of bar heights in the different graphs.

Considering the MRR1 measure, we observe that:

- In the WebEvidence case, the mean MRR1 for the site map queries is almost identical to that for the most popular queries from the query log and about 11% higher than that for the queries ranked 901–1000.
- Considering the TextOnly case, the MRR1 ratio of anchors to subj/desc is 1.24 for the Top101 set, 3.01 for the rand set, 4.81 for the MediumFreq set, and 4.72 for the site map set. (Based on past results, we would expect the performance of anchors on the site map queries to drop slightly if the anchortext from the pages from which the site map queries were derived were excluded from the index.) On three of the sets there is an overwhelming advantage to anchors, but on the most popular queries the advantage is much smaller.
- Comparison across the columns shows that for each query set there is a substantial advantage gained from taking Web evidence into account, though the advantage is much less for the randomly selected queries.
- For both TextOnly and WebEvidence cases, the plots for the Rand query set differ markedly from those for the other three sets. Web evidence brings much smaller benefit and anchors are outperformed by title/content. However, the usefulness of metadata is also depressed for the Rand set.

Subject and description metadata is of relatively greater value on the Top101 queries than elsewhere; plausibly, the metadata has been tuned to improve performance on the most popular search queries. For example, on the entry page for student results, one metadata entry includes the keywords "result" and "results" several times each, as well as "grades," "marks," and numerous other such terms. The site map queries and the query-log queries do differ in one important respect: the performance of anchors relative to Webmix. On the site map queries, anchors yielded much greater effectiveness than on the other queries, presumably because the way in which these queries were created—by selection from anchors—introduces a bias in favor of anchors as queries.

The depression of performance of anchortext and other Web evidence on the Rand query set suggests that there is a tendency for the best answers to queries that are unpopular with searchers to be documents that are not popular link targets.

A Follow-Up Investigation on the Anon University Main Site

It is possible that any potential contribution of topic metadata to effective search in our experiments has been clouded or nullified by the presence of sites within `anon.edu.au` that do not follow metadata guidelines as well as they should. Metadata policies are believed to be more rigorously practiced on the main, centrally managed `www.anon.edu.au` site than on the many other sites operated by departments and other groups.

Accordingly, we built an index of the main site only and ran all the queries for which a best answer lay on the main site. Results are shown in Figure 5. Subj/desc evidence does show improvement over the same evidence on the full site (see Figure 4): 9% in the TextOnly case and 13% in the WebEvidence case. However title/content results also improve and anchors results improve even more.

The contribution of subject and description metadata to effective search is no stronger when less well-managed sites are excluded.

Why Isn't Topic Metadata More Useful in Search?

As many people expect subject metadata to be useful in supporting effective search, it is worth investigating why our study fails to confirm this belief. In our opinion, the main reason is that it is difficult to indicate via metadata tagging the relative importance of a page to a particular topic.

An Example: "Microbiology"

Consider the query "Microbiology," which occurred 331 times in our log.[11] There is a well-defined best answer for this query, namely the home page for the School of Microbiology and Virology, `www.smv.anon.edu.au`. It is the best answer because it is the entry page to the most authoritative site on the topic of Microbiology within Anon University. It is Anon University's portal to that topic and provides a carefully organized overview of the information and services relating to it.

Within our crawl, the word "Microbiology" occurs within the content of 4233 documents, in the title of 340, and in the subject metadata of 1914, and in anchortext

[11] This query has been altered to disguise the institution.

DOI: 10.1002/asi

referencing 165. Considering only www.smv.anon.edu.au, the word "Microbiology" occurs three times in its title metadata (because the title is presented three times, once in an HTML title element, once in DC. Title metadata and once in Title metadata), twice in description metadata (DC.Description and Description), once in actual document text, twice in outgoing anchor text, twice in image alt text, and 11 times in the URLs of outgoing links. The word "Microbiology" does not appear in keyword or subject metadata. Subject metadata ("Anon University, Homepage, processing") is presented identically three times as Keyword, Keywords, and DC.Subject metadata elements.

In our crawl, www.smv.anon.edu.au receives 2134 incoming links, of which 2120 include the word "Microbiology." More than half—2120 of 3886—of the occurrences of "Microbiology" in anchor text refer to this page. The next most favored page receives 43 links containing the word "Microbiology." As far as anchortext is concerned, there is an extremely strong signal that www.smv.-anon.edu.au is the best answer for the query "Microbiology."

We suspect that the absence of the word "Microbiology" from the subject metadata of the best answer is due to an accidental omission (despite the fact that the same phenomenon occurs in other Anon University pages). Even if the mistake were rectified, at best three occurrences of the word "Microbiology" would be added, insufficient to reliably distinguish this page from the 1914 others that already include "Microbiology" in their subject metadata.

Removing "Microbiology" from the subject metadata of less-important pages would help with this search task, but doing so would harm other types of search. For example, if the searcher was actually trying to find a complete list of resources on Microbiology or if the query were more specific, such as "Microbiology 101 recommended texts."

This reasoning suggests that the problem is not the implementation of metadata at Anon University, but is inherent in the assumptions underlying it. Metadata can be developed or tuned for a particular task, but may then be inappropriate for other tasks. It also follows that having a restricted vocabulary means that many pages will share descriptors. Thus typical one or two word queries will match many pages and searches cannot be specific.

In the kinds of queries we have studied, there is typically one page (or at most a small number) that is particularly valuable. There are many other pages which could be said to be relevant to the query—and thus merit a metadata match—but they are not nearly so useful for a typical searcher. Under the assumption that metadata is needed for search, all of these pages should have the relevant metadata tag, but this makes the particular page harder to find.

Note that we have focussed on a class of queries that should be easy for metadata—queries where users are seeking an authoratitive page. Many information needs are much less easily supported through topic-specific annotation.

MediumFreq in Detail

To further investigate the benefit of metadata, we analyzed in detail 85 of the MediumFreq queries.[12]

For each of the queries, we inspected the relevant pages to see whether the metadata was of value in retrieval.

The results were not positive for topic metadata. In only one case was a query term present in the metadata but absent from the content. In all other 84 cases, all the terms were in the content, and in 57 cases all the terms were in the title. Considering these 84 in more detail, in 40 of the queries, all of the terms were in the metadata, but were also present in the title; in 7 queries where the terms were in the metadata, one term was missing from the title; in 4 queries, at least one term was only present in the content (and not the title or metadata); in 16 queries, all terms were only present in the content; in 17 queries, all terms were in the title and content, but not in the metadata.

Limiting the collections to only the documents with metadata would have had little impact on these results; that is, even if every document had metadata, problems would still arise.

Summarizing, metadata was essential in 1 query, of use in 7 queries, of potential but minimal use in 40 queries, and of no use in 37 queries. The last 20 of the Top101 queries were just as poor; in no case was the metadata clearly of value. One of our hypotheses, that there would be mismatch between query vocabulary and metadata, is substantiated by this evidence.

Another illustration of the same issue is the wide range of forms of common queries. There were over 50 occurrences of each of the queries: results, academic results, course results, exam results, "exam results," examination results, exams, grades, marks, on line results, online results, records, semester results, student records, student results, transcript of results, and transcripts (in addition to numerous variations due to plurals and use of case), each of which is a reasonable query for the same information need. A clever metadatician might be able to brainstorm all of these forms of the query—but if every reasonable page had all of these metadata terms, it would be extremely difficult to find the key results pages.

Other Observations on Anon University Metadata

For metadata to be useful in search, it needs to be accurate, and to add something to the data that cannot be deduced from the visible text. Note too that the metadata cannot stray too far from the visible text: otherwise, users will not understand why a particular page has been retrieved, as the metadata is not displayed. Absent, nonspecific, repeated, or inaccurate subject metadata reduces the effectiveness of metadata-based search.

[12] For some of the other 15 queries, the original judge had expressed some uncertainty about whether the nominated page was in fact the best answer; in other cases, the nominated page was inaccessible at the time of the follow-up analysis.

DOI: 10.1002/asi

TABLE 4. Proportion of pages with metadata, by field, among noncalendar pages.

Code	Metadata type	%
a	author	34.58
b	rights	31.52
c	description	24.44
e	type	1.95
f	format	2.25
g	relation	0.45
j	identifier	42.66
l	language	1.16
n	source	0.71
o	coverage	0.30
p	publisher	24.76
s	subject	56.50
t	title	36.15

Yet, among 180,962 noncalendar pages from Anon University, there were 78,715 (43.5%) that contained no subject metadata at all. A search service that relies exclusively on subject metadata cannot find pages that don't have any. The percentage of pages with metadata is detailed in Table 4. Although the frequent omission of metadata does not directly mean that the concept of metadata is flawed, it is surprising in a climate where the creation of metadata is strongly encouraged.

There are numerous cases of nonspecific metadata. In 34,211 pages (43.6% of all pages with some subject metadata) the subject metadata contained only the name of the university. There were many other cases of exact repetition of subject metadata, as illustrated in Table 5. In many others, these items were included as substrings. Overall, the subject metadata within 51,854 pages (50.7% of those pages with subject metadata) was repeated verbatim in at least 20 other pages. Also, the metadata was frequently redundant in other ways; for example, the title, description, and subject were often identical.

The issue of repetition highlights another paradox of metadata: Including the string "Anon University" helps to distinguish pages on, say, music studies at "Anon University" from music studies elsewhere, but makes it harder to find key pages about the institution "Anon University." The guidelines are not at fault; instead, the issue is—again—that the same metadata is being asked to serve different needs.

TABLE 5. Examples of repeated metadata. The left-hand column is the number of times an exact string was repeated; the right-hand side is a description of the string.

Frequency	Repeated subject metadata string
34211	The string "Anon University"
2205	A list of topics to do with short courses
1859	A list of education words, plus a list of 26 discipline names
1319	A list of media terms with a very idiosyncratic string of punctuation
1134	Four phrases relating to music studies, plus two names
515	The string "Anon University engineering"
487	The string "Anon University library newsitem"
406	The string "phpwebsite"

Anchors, in contrast, neatly solve the problem by using the weight of external evidence to infer these distinctions.

Inaccurate metadata is harder to detect. We did not systematically investigate its quality. However, in the investigation of the MediumFreq queries we found few cases where the metadata was of clear value. It was often vague. Faculty home page descriptions such as "faculty home page, featured information about study programs, research, services, staff and latest news" are hardly helpful—this is a description of what the metadata should be, not in itself useful metadata. This kind of example illustrates that authors struggle to understand the concept of metadata and will inevitably make entries at the wrong level of abstraction.

It can be argued that these issues are a justification for metadata training and page approval processes, but the likely benefit seems at best small, and the likely costs high. Moreover, approval processes appear to be a poor fit with the dynamic content of many pages. Where staff use pages for provision of subject materials, for example, the rate of change of content means that the pages have to be maintained outside the managed Web site.

Despite the effort spent explaining to Anon University staff how to create metadata, it has many flaws. Given the large cost of creating better metadata, and the low benefits observed in our experiments, it is difficult to see how the investment in creating it can be justified. Arguably, when the metadata can't be seen by most users and its purpose is not well understood, it is not surprising that it is often incorrect or useless. And it is not surprising that choice or creation of metadata is not well understood if the underlying concepts are difficult to apply in a consistent way.

Candan et al. (2001) note that "the metadata format used by different authors must be compatible with each other" (p. 7). This remark is part of the justification for RDF, which, it is suggested, can achieve such compatibility; however, our results suggest that authorial individuality is a fundamental problem: metadata (not just its format) must be consistent between authors, and different authors write within difficult assumptions, cultures, and frameworks. We question whether such consistency is achievable, or whether it can be any more specific than text content or evidence such as anchors.

Conclusions

Using a large institutional Web site we have explored the value of topic metadata in search. We found little evidence that metadata was of value for queries extracted from the query log for that site, even when the index was restricted to the central, well-managed site. For the most popular queries, metadata was superior to content, but was inferior to alternatives such as anchor text; some of this metadata had been altered to cater to common queries, demonstrating that metadata can be of some benefit when there is knowledge of how

DOI: 10.1002/asi

users express their needs. For all other queries, metadata was outperformed by title and dramatically outperformed by other evidence.

We found that query independent evidence (link counts and URL length) was effective in boosting performance of a title + content baseline, whereas addition of subject and description metadata caused a deterioration. For the 85 queries we examined in detail, metadata was of clear value in only one.

These results arose while exploring specific hypotheses. We found that topic metadata was of limited value for common queries, even when only pages with metadata were considered; there was mismatch between query and metadata vocabulary; and much of the metadata was inaccurate or misleading.

We found that in most respects results obtained from site map derived queries were not markedly different from those obtained using real user queries. However, the site map queries did tend to exaggerate the effectiveness of anchors. These results increase our confidence in the metadata effectiveness results reported for six other organizations by Hawking (2004).

We identified a range of causes for the shortcomings of metadata. Some of these are related to difficulties that users at Anon University experienced when creating metadata. It is apparent that many authors did not or were not able to create appropriate metadata, but instead took shortcuts, such as copying metadata from one document to another, copying it from one field to another, or simply omitting it altogether. Because the metadata is not easily visible to authors or viewers of metadata, these shortcuts are of little obvious consequence. For the same reason, metadata is often not updated when pages change.

Other issues of this kind arose through inconsistent interpretation of the metadata fields. As a simple example, `author` was variously used to mean the institution, a faculty, a department, a role (such as course coordinator), or an individual. All of these would be reasonable in different contexts, but the differences cause problems in search. More subtle problems arose from confusion between `title`, `subject`, and `description`; in many pages the content of these fields were identical.

Even with better metadata creation, many issues would still remain. If topic metadata is drawn from a restricted vocabulary it is inevitable that many pages will share descriptors, making it difficult to locate any specific page using those terms. A related problem is that inserting topic terms to facilitate one type of search may cause harm to other types of search.

Another issue is vocabulary mismatch; this issue is widely cited as a motivation for adopting metadata, but it is clearly an issue even when metadata is present. Authors have no obvious mechanism for identifying appropriate search terms in advance. We found remarkable variation in queries seeking the same page. We also found only one example of a query where the topic metadata in the best answer contained query words, but the content did not.

The fact that metadata is usually invisible to readers can create confusion. During our experiments, the relevance assessor was occasionally puzzled as to why a page had been retrieved; the usual cause was a match on metadata. Reliance on invisible and possibly erroneous metadata is not consistent with good user interface design.

We earlier noted other issues with topic metadata: vulnerability to spamming, misrepresentation, and inaccuracy. We have found that metadata is not well created even in a regime where metadata use is strongly encouraged, does not match search needs, is unreliable, and is almost never helpful. It appears that the problems are not primarily due to the particular implementation at Anon University, but to inherent issues with metadata itself.

Additional experiments with the AGov domain confirmed the main findings from Anon University. Based on our case analyses of metadata at Anon University and on search effectiveness results for a total of eight different organizations, we are confident that our findings will generalize beyond the specific organizations studied. We conclude that topic metadata is of little value in processing Web queries of the type that dominate enterprise query logs.

Acknowledgments

This work was supported by the Australian Research Council. We thank Anon University for providing the queries.

References

Abrol, M., Latarche, N., Mahadevan, U., Mao, J., Mukherjee, R., Prabhakar Raghavan, P., et al., (2001). Navigating large-scale semi-structured data in business portals. Paper presented at the 27th VLDB Conference, Rome, Italy. Retrieved December 2004, from www.vldb.org/conf/2001/P663.pdf

Agosti, M., Crivellari, F., & Melucci, M. (1999). The effectiveness of metadata and other content descriptive data in web information retrieval. In Proceedings of the Third IEEE Meta-Data Conference (META-DATA '99). Piscataway, NJ: IEEE.

Berners-Lee, T. (1998). Semantic web road map. Retrieved, from www.w3.org/DesignIssues/Semantic.html

Brin S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Paper presented at the *WWW7*. Retrieved, from www7.scu.edu.au/programme/fullpapers/1921/com1921.htm

Candan, K.S., Liu, H., Suvarna, R. (2001). Resource description framework: Metadata and its applications. SIGKDD Explorations, 3(1), 6–19.

Desai, B.C. (1997). Supporting discovery in virtual libraries. Journal of the American Society for Information Science and Technology, 48(3), 190–204.

Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y. (2004). Swoogle: A search and metadata engine for the semantic web. In Proceedings of CIKM'04 (pp. 652–659). New York: ACM.

Ding, Y., van Rijsbergen, K., Ounis, I., & Jose, J. (2003). Report on the ACM SIGIR 2003 Workshop on the Semantic Web. SIGIR Forum, 37(2), 45–49.

Dublin Core Directorate (2005). Dublin core metadata initiative. Retrieved August 4, 2005, from dublincore.org/

Doctorov, C. (2001). Metacrap: Putting the torch to seven straw men of the meta-utop. Retrieved July 13, 2005, from http://www.well.com/~doctorow/metacrap.htm

DOI: 10.1002/asi

Drott, M.C. (2002). Indexing aids at corporate websites: the use of robots.txt and meta tags. Information Processing and Management, 38(2), 209–219.

Fagin, R., Kumar, R., McCurley, K.S., Novak, J., Sivakumar, D., et al. (2003, May). Searching the workplace web. Paper presented at the *WWW2003*, Budapest, Hungary. Retrieved December 2004, from www2003.org/cdrom/papers/refereed/p641/xhtml/p641-mccurley.html.

Hawking, D. (2004). Challenges in enterprise search. Paper presented at the Australasian Database Conference *ADC2004*, Dunedin, New Zealand. Retrieved December 2004, from http://es.csiro.au/pubs/hawking_adc04keynote.pdf

Hawking, D., Upstill, T., & Craswell, N. (2004). Towards better weighting of anchors. In Proceedings of SIGIR'2004 (pp. 512–513). New York: ACM. Retrieved December 2004, from http://es.csiro.au/pubs/hawking_sigirposter04.pdf.

Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., & Yee, K.-P. (2002). Finding the flow in web site search. Communications of the ACM, 45(9), 42–49.

Hunter, J.L. (2003). Working towards MetaUtopia—A survey of current metadata research [Special issue]. Library Trends, 52(2).

Ingwersen, P. (1998). Web impact factors. Journal of Documentation, 54(2), 236–243.

Marshall, C.C., Making metadata: A study of metadata creation for a mixed physical-digital collection. In *Proceedings of the ACM International Conference on Digital Libraries* (pp. 168–177). New York: ACM.

Miller, E., Swick, R., & Brickley, D. (2005). Resource description framework (RDF) / W3C semantic web activity. Retrieved August 4, 2005, http://www.w3.org/RDF/

National Archives of Australia (2003). National Archives of Australia—Recordkeeping—Government Online—AGLS. Retrieved August 4, 2005, from http://www.naa.gov.au/recordkeeping/gov_online/agls/summary.html

Newby, G.B. (2002). The necessity for information space mapping for information retrieval on the semantic web. Information Research, 7(4).

Paepcke, A., Garcia-Molina, H., Rodriguez-Mula, G., & Cho, J. (2002). Beyond document similarity: Understanding value-based search and browsing technologies. SIGMOD Record, 29(1), 80–92.

Robertson, S.E., Walker, S., Hancock-Beaulieu, M.M., & Gatford, M. (1994). Okapi at TREC-3. In *Proceedings of TREC-3* (pp. 109–126). (NIST special publication 500–225).

Alastair G. Smith. A.G. (2002). Does metadata count? A webometric investigation. In *Proceedings of the International Conference on Dublin Core and Metadata for e-communities* (pp. 133–138). Florence: Firenze University Press.

Sokvitne, L. (2000). An evaluation of the effectiveness of current dublin core metadata for retrieval. Paper presented at Victorian Association for Library Automation, *VALA2000*, Melbourne, Victoria. Retrieved December 2004, from www.vala.org.au/vala2000/2000pdf/Sokvitne.PDF

Stephenson, S.L. (1999). An assessment of the effectiveness of metadata as a tool for electronic resource discovery. Unpublished Master's thesis, School and Information and Library Science, University of North Carolina at Chapel Hill. Retrieved December 2004, from ils.unc.edu/MSpapers/2511.pdf

Weibel, S. (1995, July). Metadata: the foundations of resource description. D-Lib Magazine. Retrieved August 4, 2005, from http://www.dlib.org/dlib/July95/07weibel.html

Wilkinson, R., Schauder, C., & Botham, A. (1991). Automatic indexing and searching of full text databases: A pilot study. Paper presented at the 6th Victorian Association for Library Automation Biennial Conference. Melbourne, Australia.

Zhang, J. & Dimitroff, A. (2005). The impact of metadata implementation on webpage visibility in search engine results (Part II). Information Processing & Management, 41(3), 691–715.

DOI: 10.1002/asi