Individual Freedom and Enlightened Self-Interest: Prerequisites for knowledge worker productivity?

David Hawking, Funnelback Pty Ltd, Canberra, Australia. david.hawking@funnelback.com

Tom Rowlands, CSIRO ICT Centre, Canberra, Australia. tom.rowlands@csiro.au

Ramesh Sankaranarayana, School of Computer Science, ANU, Canberra, Australia. ramesh@cs.anu.edu.au

Keywords: metadata, effective search, folksonomies, microblogging, document annotations, click-associated queries

Abstract

Numerous industry studies document the serious productivity cost incurred by knowledge-reliant organisations when employees are unable to locate the information resources they need to do their jobs, or take too long to do so. The traditional approach to avoiding this problem has been to impose an organisation-specific taxonomy and to tag documents with subject metadata describing and classifying it. A study of a university with a strong organisational commitment to metadata, explains why this approach fails and shows how little such metadata contributes to satisfying searches people actually conduct.

If present in sufficient quantity, metadata applied freely by “the masses” can contribute strongly to resource discovery. Hyperlinks and their anchortext provide popularity/importance signals as well as weighted topical metadata reflecting a useful diversity of perspectives and language among web/intranet authors. Queries are another useful source of weighted metadata and can be associated with resources through the clicking and other behaviour of searchers. Results from a number of organisations will be presented illustrating the search gains achievable from these sources.

Employees and stakeholders can easily be provided with the means to apply folksonomy metadata through tagging but a study of a public museum website shows that, in the absence of incentive and motivation, metadata volumes will be very small.

Finally, timeliness of metadata is critical to usefulness. We have shown how microblog posts can be interpreted as metadata describing the web resources they link to and can be used to retrieve very recently created content. Since microblogging systems are now available for the enterprise, there is potential for improvement in enterprise resource discovery from this means too, *provided that employees and stakeholders can be motivated to provide the data.*

1. Introduction

From our own experience as searchers we all know the inefficiency and lost productivity which arises from ineffective search within organisations and on their websites. How much time have we wasted searching and not finding when we are sure that the resource we need must be there? How many times have we:

- Gone to a competitor's website because the retailer/bank/airline we first approached couldn't find the answers to our needs?
- Created a report or presentation from scratch and later discovered that one already existed which we could have easily updated?
- Given up on a hopeless search facility and telephoned or gone to an enquiries desk instead?
- Completed a project to a lower standard because we didn't find the resources which would have made it better?
- Been embarrassed in a meeting or in communication with a client because we have not been able to get across all the details of our relationship with them?
- Made poor decisions because we failed to find important information?
- Taken a lot longer to get up to speed when starting a new job because of being unable to effectively search the corporate memory contributed to by a predecessor?

A number of industry analysts have attempted to quantify the financial costs of poor search. Taking just one example, a study of 600 U.S. companies by IDC [4] found that searching for information consumes about 24% of a

typical information worker's time, costing the company USD14,000 per worker per year. Furthermore, the survey revealed that employees waste three hours per week re-creating content which already exists, adding an additional USD4,500 to the burden.

With such clearcut incentives to improve the effectiveness of search, organisations have attempted to overcome the problems in a variety of ways:

- By “wheeling in” a well-known-brand search product, connecting it to all the information sources and repositories in the organisation and hoping that it will magically solve the problem.
- By providing search facilities which are specific to each separate information repository and relying on people to know which repositories to search.
- By abrogating responsibility and hoping that employees will take copies of all the documents they need and use desktop search.
- By embarking on an exercise of re-organising their content to match a taxonomy and/or attempting to impose universal markup of content to an agreed metadata schema.

Organisations differ widely in the quantity and nature of their information. Different types of business and different corporate cultures inevitably lead to different types of information need, different search patterns, and different features signalling which documents and resources are likely to be most useful. Furthermore, a great deal of useful information exploited by the major Web search engines, such as unaffiliated linking, and huge volumes of user interaction data is not available within an organisation.

Because of differences between one organisation and another and between organisations and the Web, the first approach has typically not been successful.

The second and third approaches represent the status quo with its documented shortcomings. Let us focus on the fourth approach in which some organisations have invested very heavily.

In Section 1, we summarise the results of a 2006 research study by Hawking and Zobel [6] conducted at Anonymous University, an organisation with a strong commitment to metadata and one which has invested heavily in it. Metadata can be used for a whole range of purposes; At this university an explicitly stated motivation was to improve the accuracy of text search. Hawking and Zobel analysed the extent to which the topical metadata then extant was able to contribute to answering queries submitted by visitors to the University's main website. They found it contributed very little and were able to identify a number of reasons why this was that case.

In subsequent sections, we report on four different types of social metadata which can potentially be applied to documents and assess their advantages, limitations and potential for success.

2. Conventional Topic Metadata at Anonymous University

Anonymous University was chosen for the Hawking and Zobel (2006) study because it was firmly committed to the role of metadata in supporting information discovery and search, and because we had access to query logs for the university, and cooperation from staff to provide relevance judgments.

Like many universities, Anonymous University publishes a central corporate website www.anonymous.edu.au and tolerates the publication of hundreds of other websites within the anonymous.edu.au domain. As at other universities corporate policies regarding web publishing are more strictly adhered to on the central site.

Anonymous University uses (and did so at the time) web publishing as an essential tool for promoting and facilitating research, teaching and outreach activities. Its websites are used to attract students and research funding and for image-building with the public. They are the principal means for communicating timetable information, course descriptions and lecture notes to students.

Effective earch across university websites can contribute substantially to the success of the institution. For example, the enquiry handling burden on academic and administrative staff would be exepected in increase in inverse proportion to the ability of students to discover learning resources and administrative information for themselves. Furthermore revenue generated by student enrolments would be expected to increase in some relation to the ease with which potential students can find enrolment and course information.

With the intention of improving resource discovery, Anonymous University promulgated a mandatory metadata policy and provided resources to support metadata markup. Many academics and practitioners at the time believed that this was a very sound approach, as illustrated in the following quotes:

> *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 internet information effective 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.* -- Quoted from [14]
>
> *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.* -- Quoted from [1]

But how useful was the topical (subject and description metadata) in answering the queries that searchers on www.anonymous.edu.au actually submit?

Experiments at Anonymous University

Here we describe only the experiments central to the theme of the present paper. The interested reader is referred to Hawking and Zobel [5] for further experiments, further detail and follow-up discussion.

Our crawl of anonymous.edu.au websites contained approximately 180,000 pages after we eliminated large numbers of automatically-generated, empty pages from the University calendar of events. We indexed the crawl using a search engine capable of dividing the content of documents into separate fields for title, URL, referring anchortext, and various types of metadata and querying over them separately. The search engine was configured to use the well-known and highly performing BM25 relevance scoring function [9].

From accumulated query logs for the Anonymous University search service we extracted three separate query sets:

1. Top101 - The 101 most popularly submitted queries
2. Bottom100 - The queries ranked 901-1000 in popularity
3. Random75 - 100 queries were selected at random and those which could not be reliably interpreted or answered within the domain were eliminated, leaving 75 queries with low submission frequency.

For each of these 276 queries a knowledgeable employee at Anonymous University located the best answer for each query. They used a combination of searching, navigating and their own knowledge to identify the best answers. For many queries this is very easy -- e.g. for queries like exam results, timetable, and jobs there is a single URL created to serve the information need behind the vast majority of submissions of that query.

We also added 187 artificial queries from the university's sitemap, on the basis that each topic in the sitemap (e.g. admissions, alumni and visitors) is one which the university expects will be of interest to its stakeholders. Each topic in the sitemap links to a URL which the website publisher has judged to be the most useful link for that topic.

The value of different types of evidence in finding the best answers to the 463 test queries was determined by (logically) indexing that type of evidence (e.g. title or subject metadata) in isolation, running the queries against that index and observing the rank at which the desired answer (or a known equivalent) was found. For each query a score was assigned which was the reciprocal of that rank. Thus the score is 1 for rank one, 0.5 for rank 2, 0.1 for rank 10. The average of these scores is called the *mean reciprocal rank (MRR1)* and it is used as the basis for comparison.

Results and discussion

<<< Insert Figure 1 near here >>>

<<< Insert Figure 2 near here >>>

Figure 1 shows the MRR1 for many different representations of the original web documents. As can be seen, the subject and description metadata is of less value than title and only slightly better value than the words in the document's URL. Adding subject and description metadata to title and content makes things slightly worse!

Perhaps the reason is that the results in Figure 1 relate to all of the university's hundreds of websites, where the metadata policy may not be universally applied? This was checked by indexing only the www.anonymous.edu.au website and removing the queries for which the best answers were not from this central site.

The results are shown in Figure 2. As can be seen the pattern is very similar. The improvement in value of the URL words evidence is greater than that for the topical metadata. Topical metadata alone or in combination continues to be outperformed by titles.

The most dramatic result in both figures is that all the types of evidence directly associated with the document itself --- title, content, URL words, subject and description metadata --- are strongly outperformed by anchortext.

3. Anchortext as social metadata

Anchortext was first used in Web search by the World Wide Web Worm. [8] By anchortext ("anchors" in the figures) we mean the words in a web page which a web user clicks on in order to follow a link. Frequently this anchortext provides a terse description of the link target. In our experiments, as in Web search engines [2], a document is indexed by the anchortext from other pages which refers to it, rather than its own outgoing anchortext.

Anchortext is the first form of *social metadata* we will consider. It is created by the society of web authors. When an author considers creating a link to another web page, they make a judgment about the value of that potential target and they also try to succinctly summarising its meaning in the context of the document they are writing. For example, "for details of how to set up your computer system, see the *Ubuntu Desktop Users Guide"*, where the italicised words form the anchor.

A web page which is a particularly valuable resource on a certain topic tends to acquire a large volume of incoming anchortext and may acquire a diversity of related annotations: "Einstein", "Albert Einstein", "winner of the 1921 Nobel Prize for Physics", "inventor of General Relativity". These anchortext annotations may include foreign language translations, abbreviations, nicknames and different spellings. They may use phraseology different to that of the document's author. Thus they may permit retrieval even when the query does not match the document or its metadata.

The number of anchortext annotations which match the query provides a means for ranking documents which is not available in the case of internal metadata. Let us explain this difference in the context of an example.

Consider a broadcaster's website for "the Science Show", which contains hundreds of transcripts of individual program segments. It also has a homepage which links to the transcript pages and gives general information about the show and its presenter. When applying traditional subject metadata to one of the transcripts, an author or curator would almost certainly include something like "Science Show, 21 August 2010 - The science of climate change". The phrase "Science Show" would be present in the subject metadata for each of the transcripts. If "Science Show" were not present, a retrieval system based on subject metadata would handle specific queries like "Science Show 21 August" very poorly.

But now we have a ranking problem for the broader query "Science Show". For this query the Science Show homepage is clearly the best answer because it provides a means of accessing all the transcripts in addition to general information about the Science Show. However, a purely metadata retrieval system has no obvious way of distinguishing the homepage from the individual transcripts. The same problem besets a retrieval system based on document titles.

In contrast, an anchortext retrieval system may find that, while the homepage and the transcripts each receive anchortext containing the phrase "Science Show", the homepage is easily distinguished by its much higher annotation count. Links from within the broadcaster's website and from elsewhere on the Web will predominantly target the homepage.

Anchortext voting

As we have seen, anchortext reflects a diversity of views of a document, not just those of the author or curator. Furthermore anchortext can be used for ranking, if we consider each application of an anchortext string as a vote that the document is well described by that text.

An anchortext voting model begs the question of whether each vote should be accorded an equal value. To answer it we need to consider the process by which links are created.

In a common type of webpage author-created content is embedded in branding and navigational "boiler plate" which is repeated with variations on all the pages in the site or sub-site. Links in the content section are deliberately created by the page's author. The links in the boiler plate section are also deliberately created, but for the site as a whole, and not for each individual page on which they appear. Anchortexts such as "Program Guide", "Contact your broadcaster" etc may appear 100,000 times across a site. Do they deserve 100,000 votes or just one, on the grounds that the entire ensemble of repeated anchortexts actually represents a single judgment by a single person (or committee)?

Other findings at Anonymous University

The bar labelled Webmix in Figures 1 and 2 shows that performance can be improved by combining anchortext with title and content and by factoring in non-textual sources of evidence such as link counts and URL length.

In addition to the limitations of topical metadata outlined above, Hawking and Zobel found significant deficiencies in the way it was applied at Anonymous University. There was evidence of large scale replication of metadata text, regardless of its relationship to the document question. This probably arose from authors creating a new document using another as a template. Traditional metadata tends to be invisible to both creators and consumers and hence mistakes and deficiencies tend to go uncorrected. It is interesting to note that document titles (<title> elements in HTML documents) are visible to both authors and readers and that they provided more useful retrieval evidence than did subject and description metadata.

It was not clear that content creators understood how the metadata they included would be used. It was not clear that they were motivated to create good metadata.

In contrast, anchortext has the following generally desirable properties.

Desirable properties of social metadata

1. Not prone to errors which are likely to go undetected;
2. Provides direct value to the creator (e.g. good anchortext provides a more useful site or page);
3. Reflects a diversity of views, not just those of the author or publisher;
4. Supports ranking by voting.

Anchortext can only be applied in a hypertextual environment and only by authors. Let us now consider other types of textual annotation which may be applied to a broader range of documents and by a broader population of annotators.

4. Folksonomy tags

The del.icio.us (now delicious.com) website pioneered the concept of social bookmarking in 2003. On this website people can create their own set of bookmarks for web resources that they find useful and “tag” them with their own descriptive (e.g. “cold_fusion”) or evaluative (e.g. “must_read”) textual annotations. Although people can easily create bookmarks by other means, they are motivated to use the site because they can leverage the bookmarks and tags applied by other users.

Thomas Vander Wal, described in [7], coined the term “folksonomies” to describe systems like this, in which metadata for a collection is built up by consumers of information as well as by authors and publishers.

We see that the four advantages listed for anchortext can also potentially be applied to retrieval via social tags. However, we found that the social bookmarking system operated by a large public museum [11] did not achieve all these advantages.

Probably in order to avoid problems associated with retaining personal information, the Museum elected to use an anonymous tagging system. Anyone could tag a page on the museum's site, but they couldn't come back later to get a list of the pages they had tagged or the tags they'd applied.

In other words, there is less incentive to tag (point 2 in the desirable properties list above). Not only that, but the museum is unable to keep a count of how many people have applied a tag to a webpage (point 4 in the list) -- Either a particular tag has been applied or it has not.

The museum found that the volume of folksonomy data thus collected was small. Indeed, the rate of tagging was far higher for a subset of the museum's collection of images which were posted on Flickr (flickr.com), where the incentive is higher.

We noted that, despite the paucity of data in the Museum's folksonomy, 54% of queries submitted via the museum's search facility were able to retrieve at least one additional document through the use of the folksonomy tags. That is, folksonomy tags frequently use words which are not present in either a document's content or its metadata.

Although 15% fewer queries at the museum were potentially answerable by tags than by anchortext, folksonomy annotations seem to offer definite retrieval and ranking potential, provided that the system which records them satisfies the attributes listed for anchortext. In particular, it must operate in a way which provides incentive for users to create and access tags.

5. Click-Associated Queries

Another form of social metadata arises from the interactions of searchers with a search engine. People enter queries, scan results lists and click on results in those lists. From logs of these interactions, we can simple-mindedly produce a histogram of click frequency per document and use this in ranking documents which match a query. This type of approach was taken by the DirectHit search engine in the nineteen-nineties [3]. Xue and collaborators at Microsoft Asia [13] have also studied the use of click data in Web search ranking.

Alternatively, we can interpret each click as a vote that the document clicked on is relevant to the query which generated the ranking. For each click we can annotate the clicked document with the originating query.

Arguably, the social metadata consisting of clicked queries satisfies all four of the desirable properties. However, it is only possible to apply query annotations in this way to documents which are already retrievable by the basic search system. So how useful is this type of evidence in facilitating effective search in practice?

Experiments with click-associated queries

<<< Insert Figure 3 near here >>>

In 2006 we [5] used the same methodology as in [6] to answer this question. Figure 3 shows some results obtained during this study for sitemap-generated and popular queries for both a Stock Exchange website and a government portal. The stock exchange site is characterised by a small number of pages (thousands) and a high click volume (hundreds of thousands of clicks in the sample used). In contrast, the government portal contains millions of pages and only tens of thousands of clicks.

In Figure 3 we see that for both organisations there is a big difference in the pattern for the sitemap queries compared to the popular ones. Clicks are outperformed by anchortext on the sitemap queries but the advantage reverses on the popular queries. This observation accords with intuition because obviously more click data is available for popular queries. On the other hand, the anchortexts used within the site are very likely to use the same phraseology as used in the sitemap from which the sitemap queries were generated. (Excluding the anchortexts from the sitemap itself made only a very tiny difference to the performance achieved.) In [10] we argued that test sets should be randomly sampled from the search engine query load to avoid biases in evaluation.

Content matching (including titles) is much more effective for the smaller collection than the bigger one. We hypothesize that this is because when relatively few documents fully match a query ranking them is relatively less difficult.

Readers are refered to [4] for more detailed discussion and for results from a wider range of organisations.

6. Microblog annotations as a potential solution to temporal ranking bias

Sources of ranking evidence external to a webpage are inherently biased against new content. When first published, a webpage has no incoming links (in Google terms, no PageRank), no referring anchortext, no folksonomy tags, and no click-associated queries. Recency can be used as a ranking factor but how can we distinguish between the many recently published pages? Which ones will be considered useful or important?

Microblog services like Twitter (twitter.com) provide a continuous flow of small (140 characters maximum in the case of Twitter) messages about the state of the world and its citizens. Many of these messages (tweets) include URLs of pages on the web, many of which have been only recently created. Major web search engines have begun to index Twitter content while at WWW2010 we demonstrated [12] a retrieval system in which web pages referenced in tweets were retrieved using the content of those tweets.

It seems intuitive that microblog annotations used in this way have the possibility to overcome search engine bias toward older web content, but confirming or contradicting the hypothesis has proven difficult.

Please note that microblogging is not restricted to the open Web. There are microblogging tools available for use within organisations and some organisations are in fact using them. Tweets within an organisation could potentially be used in the same way as on the Web.

7. Discussion and conclusions

Each type of textual ranking evidence external to the document has different characteristics and limitations. It seems very likely that retrieval effectiveness can be improved when more sources of evidence are added to the mix. However, it is currently very rare to find organisations where all of the types of evidence are available in useful form and useful quantity. It would be very gratifying to be able to run experiments on the relative value of different types of evidence in an environment where all types of evidence were available in abundance -- along with a

representative set of queries and corresponding right answer judgments.

To conclude:

- The availability of multiple sources of evidence helps to promote effective search.
- Topical metadata mandatorily applied to documents within an organisation contributes little or nothing to effective search. This is partly due to perennial implementation shortcomings and partly to inherent limitations.
- On the other hand, metadata democratically applied to documents through hyperlinks, behaviourally associated queries, folksonomy tags and “tweets” can improve search significantly, provided there is enough of it.
- Social metadata is only available in useful quantities when individuals are motivated to create it, or do so unconsciously through their everyday actions.

References

[1] Australian Government Information Management Office, *Why Use Metadata?*, 2004, (www.agimo.gov.au/practice/delivery/checklists/metadata, accessed Dec 2004)

[2] Brin, S. and Page, L., *The anatomy of a large scale hypertextual Web search engine*, Proceedings of WWW7, pp. 107-117, 1998.

[3] Culliss, G., *User Popularity Ranked Search Engines*, 1999. (web.archive.org/web/20000302121422/http://www.infonortics.com/searchengines/boston1999/culliss/index.htm, accessed 11 Sep 2009)

[4] Feldman, S., Duhl, J., Marobella, J.R. and Crawford, A., *The Hidden Costs of Information Work*, March 2005. (www.scribd.com/doc/6138369/Whitepaper-IDC-Hidden-Costs-0405, accessed September 2010.)

[5] Hawking, D., Rowlands, T. and Adcock, M., *Improving rankings in small-scale web search using click-implied descriptions*, Australian Journal of Intelligent Information Processing Systems. ADCS 2006 spe issue., vol. 9, no. 2, pp. 17--24, 2006. (es.csiro.au/pubs/hawking-rowlands-adcock-ad06.pdf)

[6] Hawking, D. and Zobel, J., *Does Topic Metadata Help with Web Search?*, JASIST, vol. 58, no. 5, pp. 613-628, 2007. (es.csiro.au/pubs/hawking_zobel_jasist.pdf)

[7] Mathes, A., *Folksonomies -- Cooperative Classification and Communication Through Shared Metadata*, 2004. (www.adammathes.com/academic/computer-mediated-communication/folksonomies.pdf, accessed September 2010)

[8] McBryan, O., *GenVL and WWWW: Tools for taming the Web*, Proceedings of WWW1, 1994

[9] Robertson, S.E., Walker, S. and Hancock-Beaulieu, M., *Okapi at TREC-3*, Proceedings of TREC-3, pp. 109-126, November 1994, NIST.

[10] Rowlands, T., Hawking, D. and Sankaranarayana, R., *Workload sampling for enterprise search evaluation*, Proceedings of ACM SIGIR 2007, pp. 887-888, 2007. (es.csiro.au/pubs/rowlandsHS07.pdf)

[11] Rowlands, T., Hawking, D. and Sankaranarayana, R., *Anonymous folksonomies for small enterprise webs: A case study*, Proceedings of ADCS '08, December 2008. (es.csiro.au/pubs/rowlands_adcs08.pdf)

[12] Rowlands, T., Hawking, D. and Sankaranarayana, R., *New-Web Search with Microblog Annotations*, Proceedings of WWW'2010, 2010. (es.csiro.au/pubs/rowlands10.pdf)

[13] Xue, G-R., Zeng, H-J., Chen, Z., Yu, Y., Ma, W-Y., Xi, W-S., and Fan, W-G., *Optimizing web search using web click-through data*, Proceedings of ACM CIKM 2004, 2004. ({http://doi.acm.org/10.1145/1031171.1031192)

[14] Zhang, J. and Dimitroff, A., *The impact of metadata implementation on webpage visibility in search engine results (Part II)*, Information Processing & Management, vol. 41, no. 3, pp. 691-715, 2005.