Anonymous folksonomies for small enterprise webs: a case study

Tom Rowlands
CSIRO ICT Centre and ANU DCS
ACT 2601 Australia
tom.rowlands@ieee.org

David Hawking
Funnelback
ACT 2601 Australia
david.hawking@acm.org

Ramesh Sankaranarayana
Dept. of Computer Science
Australian National University
ACT 2601 Australia
ramesh@cs.anu.edu.au

Abstract *Tags and emergent folksonomies are a potentially rich new source of document annotations, offering query independent and dependent evidence for exploitation by information retrieval systems. Previous research has shown that tags may facilitate improved web search in an environment where each tagging action generates a (user, tag, resource) triple.*

For websites operated by a public institution, operational or privacy concerns may prevent the recording of data capable of identifying individuals. This leads to a simpler anonymous tagging system but is likely to reduce user motivation for tagging, since the user cannot access their own set of tags. It also means that votes for tags are not counted, and a potentially useful joining attribute is not available.

Using webpage, metadata, query, click, anchortext and tag data provided by a public museum, we demonstrate that, despite these limitations, tag data collected by an anonymous tagging system has the potential to improve retrieval effectiveness.

Keywords Information Storage and Retrieval

1 Introduction

'Tagging' a resource is the action of tying a typically short, and often white-space free, string, the 'tag', to a resource. The resource may be a web page, document, picture, person, or a reference. Tags are used on many social networking websites such as `flickr.com`, `delicious.com`[1] and `citeulike.org` and in some blogs, allowing users to read blog posts of a particular tag. There is no restriction on the text a tag may contain; no controlled vocabulary or, necessarily, a particular meaning attached to any particular tag.

[1] Formerly known as `del.icio.us`

Proceedings of the 13th Australasian Document Computing Symposium, Hobart, Australia, 8 December 2008.

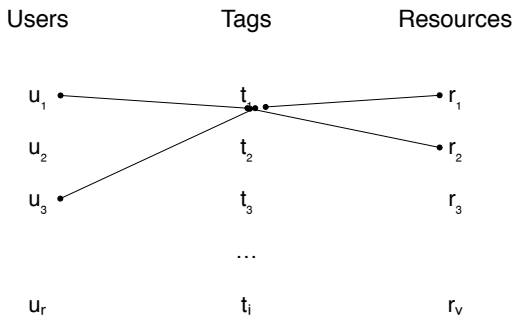

Figure 1: The relationship between users, the tags they use and the resources that are tagged, expressed as a graph. It is possible for the one resource to be tagged multiple times with the same tag, but only by different users.

On services permitting multiple taggers, such as `delicious.com`, users can typically see the tags others have applied, and over time a 'taxonomy of the folk' develops. Importantly, users can re-apply the same tag to objects already exhibiting a tag, thereby reinforcing the tag.

Bao et al. [1] have shown that this type of tagging can be used to improve retrieval effectiveness in web search. It is a more open question as to whether folksonomy tagging can, in practice, deliver retrieval benefits at an enterprise or website level. Please note that, although most folksonomy tagging systems are web-based, tags could potentially be applied to non-web data.

1.1 Anonymous tagging

Anonymous tagging systems are different to those described above in that user information such as user-id, IP address or geo-location is not recorded. All tags are public. Consequently, one of the potential incentives for tagging, organisation and ease of reference for the individual [6], is removed. While an underlying database perhaps permits a resource to be tagged more than once with the same tag, this doesn't tend to happen in practice.

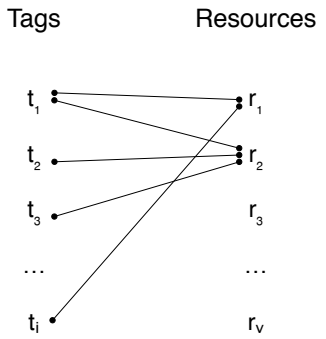

Figure 2: The relationship in an anonymous tagging environment is such that tags are applied to resources, independent of the users. This arrives at a simpler graph than in Figure 1, in which each object can only be tagged with each tag once.

Anonymous tagging offers some advantages to an organisation operating a website. There is no need to track individual users or securely store their associated logins and passwords. There is less chance that a user will feel ownership over their data, thereby reducing the risk the service provider will need to disseminate data should the service be ceased. Notwithstanding the loss of incentive mentioned in the abstract, casual visitors to the site may be more likely to tag resources, since they do not have to log in to do so.

To the best of our knowledge, anonymous tagging systems have not previously been studied from an information retrieval perspective.

1.2 Aims and scope of the present work

We present a case study of a data collection comprising document content, metadata, anchortext, user queries and clicks, as well as folksonomy tags from the anonymous tagging system of a public institutional website.

We characterise the collection and compare the distribution of number of items tagged per unique tag with that observed in a user-based tagging system.

Our principal aim is to test the hypothesis that tags collected in an anonymous tagging environment are capable of boosting retrieval effectiveness within an institutional website.

Accordingly, we investigate the following questions:

- What proportion of resources are tagged?
- How quickly is the untagged proportion of the collection likely to diminish at currently observed tagging rates?
- To what extent do queries match tags?
- Do tags permit retrieval of documents not retrieved on the basis of text created by author and/or publisher, i.e. content and official metadata.

2 Related work

2.1 Annotations

Significant retrieval effectiveness and efficiency gains have been demonstrated by the use of annotation data.

Craswell et al. [3] demonstrate superior site finding performance with an anchortext surrogate index 'an order of magnitude' smaller than content. Eiron et al. [5] study the utility of anchortext for information retrieval based on the idea that queries, anchortext and titles are created by a similar thought process.

Xue et al. [12] use surrogates and compensate for a relatively small quantity of click data by using co-visitation.

Dmitriev et al. [4] invite users to add 'explicit' annotations to an intranet. They argue that such annotations are expensive to produce as users have to be asked to produce them and often do not find the time. They also examine 'implicit' annotations, such as queries associated with documents through clicks.

2.2 Tags and folksonomies

Mathes [10] proposes some attractions of user generated textual metadata, arguing the relative simplicity of tagging systems, their 'low cognitive cost', rapid feedback and development of communities, all as attractions to potential taggers. Golder and Huberman [6] discuss how folksonomies are distinct from taxonomies in their being non-hierarchical and inclusive. They analyse data from `delicious.com` and demonstrate some interesting effects by comparing users. They also report that tags are not always used as a description of the document content and so document and tag vocabularies can be expected to be different.

Bao et al. [1] define *SocialSimRank*, estimating the similarity between queries and web pages based on the tag graph. They also define *SocialPageRank*, estimating the query independent value of a web page based on the tag graph. They use both in a whole of Web search task, combining with other forms of evidence using a support vector machine, and show good results.

The challenge of tag segmentation due to lack of explicit boundaries (e.g. 'informationretrieval') in tags, discussed in [1], is not a problem in our case.

Halpin et al. [7] suggest a generative model for tagging that arrives at a power law (see [2]) based on preferential attachment. The central idea is that users are more likely to tag a resource with a tag that has already been used for that resource. They demonstrate tags on `delicious.com` following a power law pattern.

3 Data

The primary data used in this paper has been gathered from an online museum catalogue. Pictures and descriptions of the museum's artifacts are published, along with various metadata, as HTML pages and are accessible via a standard web interface. We collected the document data using a commercial web crawler. The museum provided tag and click data as a database dump. The site is particularly interesting because it offers non-trivial quantities of content, anchortext, click associated query and tag evidence.

The tagging data covers only those items within the museum's 'online collection', which comprises 132327 of the 135216 documents on the museum's website. We consider only the online collection. The majority of documents describe a single museum exhibit and include Title, Description (usually identical to title), Keywords, and content.

The click data covers the period 14 July 2008 to 14 August 2008. The content of sixty documents for which we have tag data were not yet downloaded by the time our crawl was terminated. The crawl was tempered to reduce load on the museum's servers. It started on 20 August 2008 and lasted just under five days.

There are 7221 distinct tags with 11 509 applications of those tags. Each application included the date on which the tag was applied to the document. There were no cases of the same tag being applied more than once to the same document.

A conventional query log was not available. We have no information about queries which were submitted but which did not lead to any click. Instead, we have a click log which shows 10 747 distinct user-composed queries[2] associated with a total of 364 310 clicks. It is known that the site search facility makes use of the tags.

The average length of the tags, queries and anchortext is 1.5, 1.4 and 8.1 words respectively. The distribution of lengths is shown in Figure 6. Anchortext tended to repeat the title, often truncated, of the target page.

For the study, all tags and queries have been case folded and leading and trailing whitespace has been removed. In our data, tags are associated with document identifiers rather than URIs. Sometimes multiple URIs share the same docid e.g. `http://museum.com/getdoc?docid=1&image=1` and `http://museum.com/getdoc?docid=1&image=2`. In such cases, the tag or click has been considered to 'apply' to both URIs.

4 Experiments

4.1 Experiment 1: Characteristics of anonymous tags

In this section, we investigate the frequency of applications of tags to particular resources. It has been previously established that, in non-anonymous tagging systems, sufficiently popular tagged resources and entire folksonomies from the one system yield power law-like distributions. An example of such a distribution is shown in Figure 4.[3] The anonymous tagging system that is the focus of this paper is shown in Figure 3. This shows graphically that, like queries and non-anonymous tagging systems, anonymous tagging systems yield a few tags occurring a relatively large number of times, with many tags occurring very rarely.

[2]We eliminated a large number of records in which the query was generated by a user clicking on a navigational link within the site.

[3]This example data is taken from the `citeulike.org` facility.

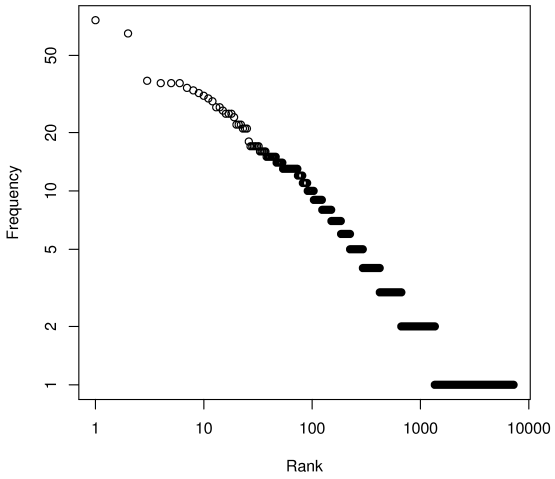

Figure 3: A log-log graph of the number of times each tag has been used across the corpus, ordered by that number. For example, the most frequent tag, ranked 1, has been applied to 76 objects. As the tagging system is anonymous, each tag can only be applied to each object once. The distribution is similar to that in the more traditional case, shown in Figure 4, despite the lack of reinforcement.

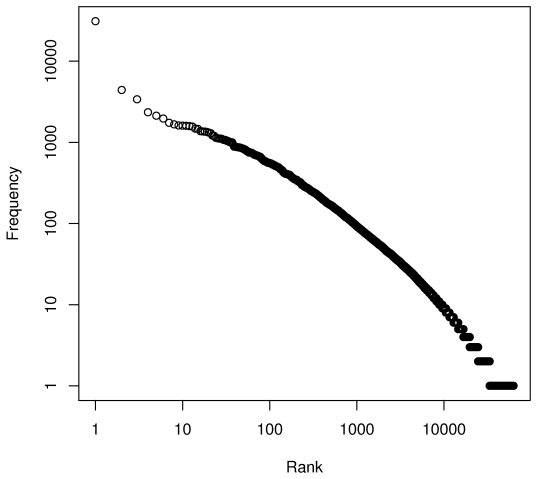

Figure 4: A log-log graph, similar to Figure 3, but from a non-anonymous tagging system where tags can be reinforced by other users. This data is from `citeulike.org`.

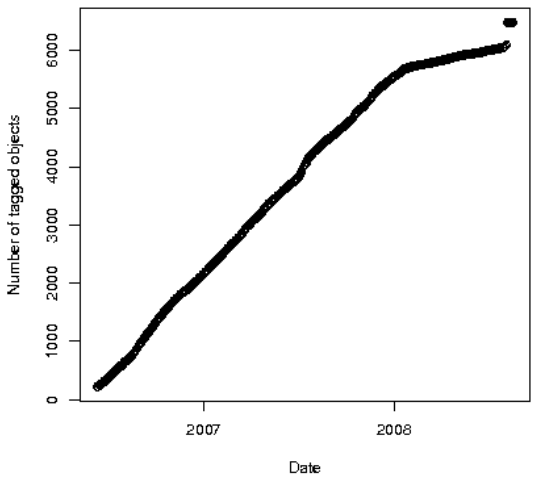

Figure 5: The number of tagged objects over time. The sudden jump in the number of tags at the far right reflects an import of tags made available to users through the Flickr site. Flickr is very popular. Note that this is the number of tagged objects, not the number of tags; if an object is tagged multiple times it is only counted once.

4.2 Experiment 2: How prevalent are tags?

We counted the number of documents (museum objects) to which tags had been applied and observed how that increased with time.

4.2.1 Results

The number of tagged objects plotted against time is shown in Figure 5. The number of tagged objects is far fewer than the number of separate articles in the museum, and reflects only around five per cent of the collection. The overall average rate of tagging over two years is three and a half thousand objects tagged per year.

It is not clear why there is a 'knee' in the plot around the beginning of 2008. The museum suggests this may be due to a declining number of people interested in tagging. Note the sudden jump in the second half of 2008 due to the import of tags from Flickr.

4.2.2 Discussion

If, optimistically, the average rate of tagging were to be maintained, it would be approximately another thirty five years before all the items in the *present* collection received at least one tag.

A substantial increase in the number of tagged objects is provided by making the objects available on the external Flickr site.[4] All of the tags added by users of Flickr were to objects that were previously untagged, and in some cases there were multiple tags added to the same object.

[4]See `http://www.flickr.com/commons/`.

4.3 Experiment 3: Do tags match queries?

To assist in retrieval, tags must match queries actually received by the information retrieval system. We consider three different matching models:

- Exact match: the query and tag strings are identical.
- AND match: all the words in the query are present in the tag.
- OR match: at least one query word is present in the tag.

Note that each of these models is applied to tags individually, not to the collection of tags applied to an object. Obviously, the second and third models are the same when the query consists of only one word.

For contrast, a similar calculation is conducted for anchortext. 'Canned' or automatically produced links containing query text or tags have been filtered out to avoid inflating the results in this case. An example of such links is the automatically generated list of 'recently applied tags'.

4.3.1 Results

Table 1 shows the percentage of query instances that might be answerable by tags and anchortext using three different matching schemes.

Overall, 88% of the query instances share at least one term with a tag and, as a consequence, *might* be at least partially answerable by that tag. This percentage drops to 69% for AND match and 61% for Exact match.

Table 2 looks at queries and annotations the other way around—what percentage of annotations are useful in answering at least one query? The percentages are reasonably high for tags. Overall, 81% of tags are potentially useful (OR match) in answering at least one query and 57% of tags exactly match a query.

A high proportion (88%) of anchortext annotations achieve an AND match with at least one query, but there are no exact matches.

4.3.2 Discussion

No anchortext annotations exactly match any queries, even though there is a very high degree of AND match. The lengths of strings used as anchortext (usually the title or an abridged title of the target document) are quite different to those of tags and queries; this can be seen in Figure 6. An exact match function would not be appropriate for use with anchortext on this site.

4.4 Experiment 4: Do tags contribute useful additional terms?

Here we are interested in queries answerable by tags but not by document content, metadata or anchortext.

Table 1: Percentage of the query workload, of various lengths, matching at least one annotation. For example, twenty seven per cent of query instances of length two AND match a tag.

Match type	Exact					AND					OR				
Query length	1	2	3	≥ 4	all	1	2	3	≥ 4	all	1	2	3	≥ 4	all
Tags	75	21	24	2	61	85	27	27	3	69	85	96	99	100	88
Anchortext	0	0	0	0	0	97	93	90	87	96	97	100	100	100	98

Table 2: Percentage of annotations, of various lengths, matching at least one query instance. For example, twenty three per cent of tags of length three match a query exactly.

Match type	Exact					AND					OR				
Annotation length	1	2	3	≥ 4	all	1	2	3	≥ 4	all	1	2	3	≥ 4	all
Tags	68	39	23	11	57	68	81	86	93	73	75	94	98	99	81
Anchortext	0	0	0	0	0	68	66	98	92	88	72	83	100	99	96

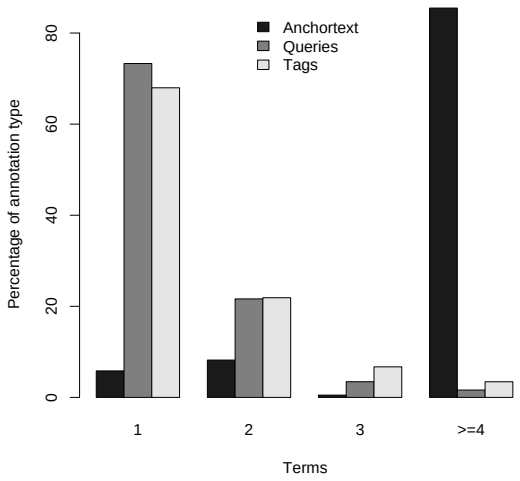

Figure 6: The percentages of different lengths of annotation data. There is a large proportion of anchortext that is four or more terms long, while tags and query instances both tend to be short.

4.4.1 Results

For 48% of distinct queries, a tag matching the query (at least one intersecting term) reveals at least one new document with which the query did not share a term. This represents 54% of the workload. Document content included all text and metadata.

The percentages of query instances matching tags but not the associated document is shown in Table 3.

5 Discussion

We have seen that in this instance, even after two years offering a tagging interface a relatively small number of objects have been tagged. This is a disappointment to the museum which they partially addressed by posting items on Flickr and collecting tags.

Many of the reasons for tagging outlined by [9] and [6] do not apply in the anonymous tagging environment. This may partly explain why objects displayed in Flickr are tagged at a much higher rate than on the museum site.

We note that it may be possible to derive further tags implicitly from website referrer logs. Another possibility might to be to provide tagging incentive by instituting a tagging game (see e.g. [11]).

Not every object must be tagged for the tags to contribute useful evidence. For example, an anonymous tagging system on an intranet may assist staff in finding key pages more quickly even with only a small subset of important pages tagged.

Fifteen per cent fewer queries were potentially answerable by tags than by anchortext, but 'potentially answerable' is an extremely optimistic metric. Examining from the 'other direction', however, it is most often very long anchortext that matches queries.

The rough power law distribution shown by [7] is seen in the folksonomy distribution shown in Figure 3. Halpin et al. suggested in their generative model of tagging systems that the likelihood of a tag being applied to an object was influenced by the tags already applied to that object. In the case of anonymous tagging systems, the distribution of tag application is 1—a straight line—and yet the distribution across the corpus is still a, roughly, power law distribution with a steep decline for the most frequent tags.

Any system permitting users on the Web at large to add or remove information at will opens itself to the issue of spam [8]. The tags on the site investigated here seemed to be relatively free of spam. By restricting the resources to those available on the site it is made less attractive to spammers attempting to manipulate rankings in commercial search engines. Private systems allowing links beyond the site itself are also possibly immune for the same reason; their impact on commercial search will be small or nil.

Table 3: Percentage of query workload where a query matches a tag but does not match the content of a tagged document (including Title, Keyword and Description metadata) or anchortext pointing to the tagged document. For example, 49% of queries containing two terms matched a tag that was applied to at least one document containing neither of the query's terms.

Match type	Exact					AND					OR				
Query length	1	2	3	≥ 4	all	1	2	3	≥ 4	all	1	2	3	≥ 4	all
Content	28	1	0	0	21	57	1	0	0	42	57	49	37	20	54
Anchortext	36	1	0	0	27	77	2	0	0	58	79	94	97	99	83

6 Conclusions

Anonymous tagging systems, like that deployed at the museum which is the object of the present study, do not provide the same incentive to tag as do user-centric tagging systems on the Web. When objects from the museum are displayed in Flickr, they are tagged at a much higher rate than on the museum's own site. Anonymous tags provide only a binary signal as to the importance of a resource with respect to a tag. There is no voting aspect; either a tag is applied or it is not.

Despite these differences, anonymous tag data from the museum shows a similar distribution of tags to that described by [7].

Although the sparsity of tag data and its slow rate of accumulation mean that a retrieval system for the museum could not be based on tags alone, we found that a relatively high proportion (54%, assuming OR-match) of the query load for which answers could be identified using the tags that were not identified by text or metadata generated by the author or publisher. This suggests that future research on combining anonymous tags with other evidence in a retrieval system would be worthwhile.

Acknowledgements The authors would like to thank the museum who kindly granted access to the data, without which the experiments could not have been run.

References

[1] Shenghua Bao, Gui-Rong Xue, Xiaoyuan Wu, Yong Yu, Ben Fei and Zhong Su. Optimizing web search using social annotations. In Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider and Prashant J. Shenoy (editors), *WWW*, pages 501–510. ACM, 2007.

[2] A. Clauset, C.R. Shalizi and MEJ Newman. Power-law distributions in empirical data. *Arxiv preprint arXiv:0706.1062*, 2007.

[3] Nick Craswell, David Hawking and Stephen Robertson. Effective site finding using link anchor information. In *Proceedings of ACM SIGIR 2001*, pages 250–257, 2001.

[4] Pavel A. Dmitriev, Nadav Eiron, Marcus Fontoura and Eugene Shekita. Using annotations in enterprise search. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 811–817, New York, NY, USA, 2006. ACM Press.

[5] Nadav Eiron and Kevin S. McCurley. Analysis of anchor text for web search. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 459–460, New York, NY, USA, 2003. ACM.

[6] Scott A. Golder and Bernardo A. Huberman. Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, Volume 32, Number 2, pages 198–208, 2006.

[7] Harry Halpin, Valentin Robu and Hana Shepherd. The complex dynamics of collaborative tagging. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 211–220, New York, NY, USA, 2007. ACM.

[8] Monika R. Henzinger, Rajeev Motwani and Craig Silverstein. Challenges in web search engines. *SIGIR Forum*, Volume 36, Number 2, pages 11–22, 2002.

[9] Cameron Marlow, Mor Naaman, Danah Boyd and Marc Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006. ACM.

[10] Adam Mathes. Folksonomies-Cooperative Classification and Communication Through Shared Metadata, December 2004.

[11] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM Press New York, NY, USA, 2004.

[12] Gui-Rong Xue, Hua-Jun Zeng, Zheng Chen, Yong Yu, Wei-Ying Ma, WenSi Xi and WeiGuo Fan. Optimizing web search using web click-through data. In *Proc. ACM CIKM '04*, pages 118–126, 2004.