Finding Similar Pages in a Social Tagging Repository

Alex Penev
NICTA & University of NSW, Australia
alexpenev@cse.unsw.edu.au

Raymond K. Wong
NICTA & University of NSW, Australia
wong@cse.unsw.edu.au

ABSTRACT

Social tagging describes a community of users labeling web content with tags. It is a simple activity that enriches our knowledge about resources on the web. For a computer to help users search the tagged repository, it must know when tags are good or bad. We describe TagScore, a scoring function that rates the goodness of tags. The tags and their ratings give us a succinct synopsis for a page. We 'find similar' pages in Del.icio.us by comparing synopses. Our approach gives good correlation to the full cosine similarity but is hundreds of times faster.

Categories and Subject Descriptors: H.3.5 [Information Retrieval]: Online Services—*Web-based services*

General Terms: Algorithms, Experimentation

Keywords: Web, tagging, social bookmarking, del.icio.us

1. INTRODUCTION

Conceptually, a tag is any simple idea describing an object. A combination of tags describes the object in higher detail. Today's tagging systems are built around users, tags and resources. Their tags are implemented as free-form text keywords and the resources may be any web object (URLs, photos, videos, research articles, blogs, etc).

We are interested in URLs tracked by Del.icio.us, whose users bookmark pages that they find interesting. It is 'collaborative' tagging because users can see the tags used for the URL by the whole community or by other individuals.

So far tagging systems outpace our understanding of how to best support their socially-driven annotation model for searching the repository. In terms of search, a single tag acts as a *filter* because it selects a subset of all objects (a combination of tags selects fewer). As Del.icio.us's repository grows, search becomes more difficult due to noise, inaccuracy, spam or lack of navigational functionality.

TagScore can help in each case. However, we focus this work on currently missing functionality, in particular to 'find similar' pages. Using a page's tags as filters to find similar pages will not work well: a boolean 'and' query will fail for pages with a reasonable number of tags, and 'or' queries retrieve too many results. The user is forced to pick and choose which filters to apply. Unfortunately tags do not support intuitive query refinement and such a behavior fails to make use of tags in intelligent ways. Instead, we apply TagScore. It rates the goodness of a tag and helps us decide how much weight to give tags in a multi-tag search.

Given a page, we have a basic synopsis from Del.icio.us: its tags, how often they were used, and the total number of bookmarks. TagScore enriches this synopsis by also giving each tag a numeric rating (contribution) and giving the page itself an overall 'confidence' score. The overall score is used to prune very large result sets to desired sizes. The contributions are used for the synopsis similarity measure.

2. APPROACH

TagScore. We rate tags with a weighted TFIDF-based approach. It is simple, fast[1], works well for tagging [1, 2] and allows us to score pages independently of each other. The raw scores are weighted to reflect where in the page tags match and the community's notion of their importance.

We address tagging's common criticisms by lessening the impact of idiosyncratic tags using frequencies and combatting word sense ambiguity by also matching related words. Using pre-built global lookups for cooccurrence and WordNet, we obtain up to 100 related words for any word in our dataset (DMOZ100k06 [3]). These are rated the same way as the tags, and their contributions are compounded into a single albeit abstract tag.

Related terms are important for two reasons. An average page has only 5.0 tags and its tag bag is coarse and not an exhaustive description; the terms significantly improve the match rate and overall score (21%). They also smoothen the distribution by reducing the number of ties.

The final contributions of the abstract and user tags are combined (using several normalizations and geometric sums that insist on quality over quantity) to give a single overall 'confidence' score in [0,1). Although we score pages independently, Fig 1(a) shows that TagScore gives a near-uniform score distribution for a randomly-sampled dataset.

The above computations are reasonably fast and scalable. The bulk (71%) of execution time is in disk IO, which can be cut in half with the lookups in memory.

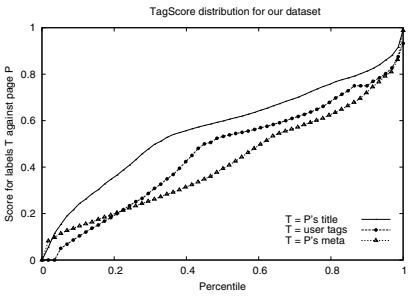
Synopses. Tags are a succinct synopsis of a page. Finding similar pages can be considered a multi-tag search where tag overlap is maximized. We enrich the basic synopsis by attaching a few extra bytes per tag for its rating. There is an obvious space vs accuracy trade-off here, and we use a minimal amount. A comparison between enriched instead of basic synopses is slower, but only by 4% since the enrichment is a matter of bytes. The accuracy gains, however, are far larger (Fig 1(b)).

[1]We determine the enriched synopsis for a page in 0.006s.

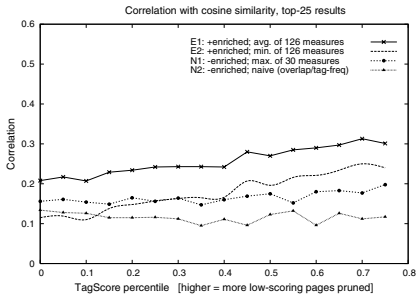
WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

Label source	Samples	Avg/Median TagScore	$\neg CO$	$\neg WN$	No related terms
$T = P$'s meta	2664	0.415 / 0.39 ($\sigma = 0.22$)	0.395 (-5%)	0.417 (+0%)	0.361 (-13%)
$T = P$'s title	4275	0.555 / 0.59 ($\sigma = 0.22$)	0.543 (-2%)	0.554 (-0%)	0.510 (-8%)
T = User tags	4278	0.460 / 0.53 ($\sigma = 0.24$)	0.450 (-2%)	0.454 (-1%)	0.388 (-16%)

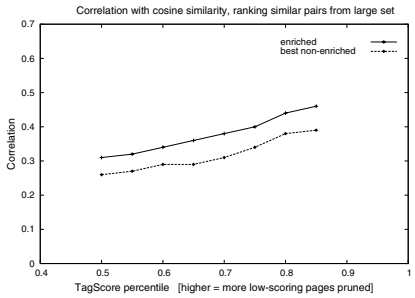
Table 1: Summary of Fig 1(a). Also shows TagScore when not using COoccurrences or WordNet.



(a) TagScore distribution for using the title, tags and meta as page P's labels.



(b) Kendall τ for "given x, find similar to x" compared against cosine.



(c) Kendall τ for "given a large pool, find similar x,y pairs" against cosine.

3. TAGSCORE PROPERTIES

Fig 1(a) shows the TagScore distribution. The smoothness indicates few ties. The linearity indicates a uniform spread. A summary is in Table 1.

We can see that title is a stronger content descriptor than user tags, and both are better than meta-keywords. This has long been suspected by web developers. TagScore captures this sentiment and even quantifies the differences.

Not matching a page's meta reduces the TagScore mean by 15%. Correlation with the original list is very strong (0.74) but does not justify ignoring meta completely because it is inexpensive. We found that taking the first-25 terms is a good shortcut (2% drop in mean, $\tau > 0.9$).

Ignoring title gives a similar mean to ignoring meta, suggesting they are interchangeable for being matched against by user tags. Correlation was very strong (0.76), but title achieves this accuracy with fewer terms and is more ubiquitous, therefore it is the more-useful component of a page.

Taking only the first 2.5KB of the body content produced almost identical mean, median and σ as using the full body and had a very strong (0.77) correlation against the original list, yet it is on average one-fifth of the page source. We find that this is a good shortcut in practice.

4. APPROXIMATE SIMILARITY

We performed a one-to-many comparison by taking a random synopsis, comparing against the others and ranking the results. From 13 values (or combinations of values) obtained from enriched synopses, we trialled 156 different similarity measures for producing a ranked list by using the 13 as primary sort criteria and the other 12 to break any ties. Of these, 126 used the enriched tag contribution variable in some way; the others used data Del.icio.us currently has.

Fig 1(b) shows a summary of the results. The curve N2 is the naive approach. The curve N1 plots the best *instance* among *any* of the 30 non-enriched measures (averaged over 500 runs). It was similar to E2, the *worst* instance among the 126 enriched measures. The *average* of the enriched measures, E1, was much better.

The result is clear: even the worst of the enriched measures was a closer approximation to cosine. We also see that τ increases with percentile cut-off as low-scoring pages are prune. This suggests that under TagScore, low-scoring pages are indeed poorly represented by their tags.

Fig 1(c) shows a many-to-many comparison where we find the most similar pairs in a large pool of synopses. We trialled one of the better enriched measures against the best non-enriched measure. The result is clear: the ranking produced by the enriched measure is closer to cosine. Again, we see that τ increases with percentile.

5. CONCLUSION

Social tagging systems have large repositories that require new methods for search. We created a scoring function to rate the goodness of tags and summarized how it was used to fill a gap in currently missing functionality: to 'find similar' pages in a tagged repository. We contribute:

• TagScore, a scoring function to rate the goodness of tags and give the page an overall confidence score.

• Experiments on real Del.icio.us data to show TagScore's distribution and behavior. TagScore allows us to quantify various comparisons, such as the author's description of a page (title, meta) compared to users' tags. We also documented the effects of matching related terms.

• Two experiments on finding similar pages in a tagged repository. Our approximation has good correlation with cosine but is 490 times faster. Our implementation compared about 70,000 synopsis pairs per second.

6. REFERENCES

[1] Brooks, C., and Montanez, N. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *WWW* (2006).

[2] Chirita, P. A., Costache, S., Nejdl, W., and Handschuh, S. P-tag: large scale automatic generation of personalized annotation tags for the web. In *WWW* (2007).

[3] Noll, M., and Meinel, C. Authors vs. Readers - A Comparative Study of Document Metadata and Content in the WWW. *ACM DocEng* (2007).