A Scalable, Collaborative Similarity Measure for Social Annotation Systems

Benjamin Markines* Filippo Menczer
School of Informatics, Indiana University, Bloomington, USA

ABSTRACT

Collaborative annotation tools are in widespread use. The metadata from these systems can be mined to induce semantic relationships among Web objects (sites, pages, tags, concepts, users), which in turn can support improved search, recommendation, and other Web applications. We build upon prior work by extracting relationships among tags and among resources from two social bookmarking systems, `Bibsonomy.org` and `GiveALink.org`. We introduce a scalable and collaborative measure that we name *maximum information path* (MIP) similarity. Our analysis shows that MIP outperforms the best scalable similarity measures in the literature. We are currently integrating MIP similarity into a number of applications under development in the GiveALink project, including search and recommendation, Web navigation maps, bookmark management, social networks, spam detection, and a tagging game to create incentives for collaborative annotations.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Information networks, Performance evaluation*

General Terms

Performance, Algorithms, Design, Experimentation

Keywords

Maximum Information Path, Folksonomy, Web 2.0, Tag, URLs

1. INTRODUCTION

As the democratization of the Web continues, social annotation systems have become prevalent. Here we expand on our recent work to extract relationships among tags and resources in a folksonomy [5, 4]. Tag relationships can lead to advanced applications in tag navigation, keyword clustering, query expansion, tag recommendation and ontology learning Resource (page/site) similarity supports result clustering, similarity search, ontology population and again page recommendation and navigation. A *scalable* similarity measure is updated to reflect new annotations at a pace that keeps up with a stream of incoming annotations. In prior work we introduced a scalable method for assembling a semantic similarity network based on folksonomies [5, 4]. Here we only consider measures that are scalable. We base our evaluation on two social bookmarking systems, `BibSonomy.org` and `GiveALink.org`. When considering only scalable measures, an analysis of the GiveALink data showed that cosine and mutual information (MI) perform well for measuring similarity among resources [5]. An analysis of BibSonomy data confirmed that cosine and MI outperform other scalable measures for mining relationships among tags and resources [4]. These datasets are available via APIs at `givealink.org/main/download` and `bibsonomy.org/help/doc/api.html`.

We expand on our previous findings in three ways. First, we introduce the *maximum information path* (MIP) measure to induce a similarity among tags and resources. MIP is a generalization of a previously defined similarity measure that was designed to induce similarities among resources [5]. Second, we expand the evaluation to include MIP as it applies to a BibSonomy dataset, finding that MIP outperforms competing measures for tags and resources. Finally, we confirm that MIP outperforms competing measures for both tags and resources using a new, more representative dataset from GiveALink.

2. TRIPARTITE SIMILARITY

Here we focus on the aggregation method that was found to best capture semantic similarity, while being scalable. Our approach is based on the *triple* representation. Each triple (u, r, t) represents user u annotating resource r with tag t. A set of triples represents a folksonomy. When computing similarities among u, t, or r, it is necessary to obtain two-mode views of the data. Here we focus on tag-tag and resource-resource similarities. Therefore we aggregate across users, and obtain dual views of resources and tags, yielding dual definitions for resource and tag similarity. More specifically, we consider only the *collaborative* aggregation method as defined in prior work because of runtime performance and accuracy [5, 4]. *Collaborative* aggregation creates a per-user binary matrix of the form $w_{u,rt} \in \{0, 1\}$, which we can use to compute a "local" similarity $\sigma_u(x, y)$ for each pair of objects (resources or tags) x and y. Finally, we sum across users to obtain the "global" similarity: $\sigma(x, y) = \sum_u \sigma_u(x, y)$. For collaborative aggregation, we assign a local similarity $\sigma_u(x, y) > 0$ to every pair of objects (x, y) present in u's annotations, irrespective of shared features. We can achieve this by adding a special "user tag" (resp. "user resource") to all resources (resp. tags) of u so that each item of u has at least

*Corresponding author. Email: `bmarkine@cs.indiana.edu`

HT'09, June 29–July 1, 2009, Torino, Italy.

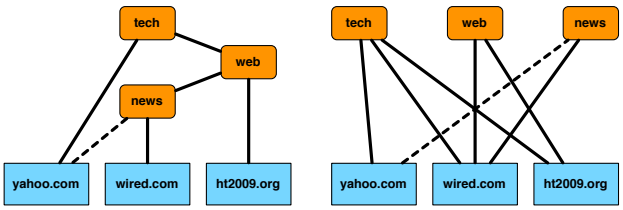

Figure 1: The annotations of user Charlie are visualized on the left in a hierarchical taxonomy and on the right as a folksonomy. If he does not tag `yahoo.com` with `news`, the annotations are hierarchical. The σ^{mip} similarity between `wired.com` and `ht2009.org` is determined by the lowest common ancestor on the left, the tag `web`. Indeed we see on the right that this is the most specific tag shared by the two resources. If he tags `yahoo.com` with `news`, the hierarchy breaks as multiple paths connect `wired.com` and `ht2009.org`. The lowest common ancestor is no longer defined, however in the folksonomy we can identify `web` as the most specific shared tag.

one annotation in common. Maximum information path uses the Shannon information (log-odds) to weigh relationships between objects. Thus we use the information of a tag (resp. resource) x defined as $-\log p(x)$ where $p(x)$ is the fraction of resources (resp. tags) annotated with x. Let us redefine user u's odds of tag (resp. resource) x as $p(x|u) = N(u,x)/N$ where $N(u,x)$ is the number of resources (resp. tags) annotated by u with x, and N is the *total* number of resources (resp. tags) in the system. This definition of N is a variation from prior work where we used the number of resources (resp. tags) in a user's profile plus a constant when computing odds. This way, $-\log p(t_u^*|u) = -\log[N(u,x)/N] > 0$ when $N > N(u,x)$. For evaluation we set $N = 10^6$.

Maximum Information Path. Maximum information path (MIP) is symmetric with respect to resources and tags, therefore we simplify the notation as follows: x represents a tag or a resource and X is its vector representation. For example, if x is a resource, X is a vector with tag elements w_{xy}. For a single user u, $y \in X^u$ means $w_{u,xy} = 1$ and $|X^u| = \sum_y w_{u,xy}$.

We define MIP as:

$$\sigma_u^{mip}(x_1, x_2) = \frac{2\log(\min_{y \in X_1^u \cap X_2^u}[p(y|u)])}{\log(\min_{y \in X_1^u}[p(y|u)]) + \log(\min_{y \in X_2^u}[p(y|u)])} \quad (1)$$

where probabilities use the above construction for $p(x|u)$. This measure has linear runtime and space complexity.

MIP is an extension of traditional shortest-path based similarity measures [6] and Lin's similarity measure [2]. MIP differs from traditional shortest-path similarity measures by taking into account Shannon's information content of shared tags (resp. resources). Lin's measure only applies to hierarchical taxonomies, such as bookmarks organized in folders. If the folksonomy is derived from such a hierarchy, the two measures are equivalent. However when the folksonomy includes non-hierarchical annotations, Lin's measure is undefined while MIP is well defined and captures the same intuition. Namely, that the semantic association between two objects is determined by the ratio between the maximum information they have in common and the information they do not share. In the hierarchical case the maximum shared information coincides with a unique lowest common ancestor; however in folksonomy, there may be many paths between two objects, and the maximum information path passes through the most specific shared tag. Fig. 1 illustrates this idea.

Table 1: Object-object similarity accuracy, according to Kendall's τ correlations between the similarity vectors generated by the various measures and those from the reference dataset. Higher τ means more accurate similarity measures. Errors are obtained by a random shuffle of the reference similarities.

BibSonomy	tag	resource
Cosine	$(4115 \pm 2) \times 10^{-5}$	$(594 \pm 3) \times 10^{-5}$
MI	$(6088 \pm 2) \times 10^{-5}$	$(593 \pm 3) \times 10^{-5}$
MIP	$(6118 \pm 2) \times 10^{-5}$	$(663 \pm 3) \times 10^{-5}$
GiveALink	tag	resource
Cosine	$(201 \pm 1) \times 10^{-4}$	$(3061 \pm 2) \times 10^{-5}$
MI	$(299 \pm 1) \times 10^{-4}$	$(3569 \pm 2) \times 10^{-5}$
MIP	$(316 \pm 1) \times 10^{-4}$	$(3983 \pm 2) \times 10^{-5}$

3. EVALUATION

In prior work, cosine and mutual information (MI) outperformed all other measures for extracting relationships among tags and resources [5, 4]. We expand on those findings by applying σ^{mip} to the BibSonomy and GiveALink datasets, while comparing its performance to cosine and MI.

We use an evaluation framework that gauges how a similarity measure predicts a reference dataset. For tag similarity, we use the WordNet (`wordnet.princeton.edu`) term collection for the semantic grounding. The tag pairs are ranked by their Jiang-Conrath distance [1]. For resource similarity, we use the URL collection of the Open Directory Project (`dmoz.org`) for the semantic grounding. The resource pairs are ranked by Maguitman *et al.*'s graph-based similarity measure [3]. This indirect method of evaluation is justified and explained further in prior work [5, 4].

Table 1 has the Kendall's τ correlation between the reference dataset and the proposed similarity measures. MIP performs better than cosine and MI for measuring similarity among tags and resources across the BibSonomy and GiveALink folksonomies. Some of the differences are small, but all differences are statistically significant.

4. CONCLUSION

We have improved on the state of the art in scalable collaborative similarity measures from social annotation systems. Our extended evaluation shows that among scalable measures, the new maximum information path similarity outperforms mutual information and cosine in accuracy. These results are consistent across tag and resource relationships, and across datasets from two diverse social tagging sites.

5. REFERENCES

[1] J. J. Jiang and D. W. Conrath. Semantic Similarity based on Corpus Statistics and Lexical Taxonomy. In *Proc. Intl. Conf. on Research in Comput. Linguistics (ROCLING)*, 1997.

[2] D. Lin. An information-theoretic definition of similarity. In *Proc. 15th Intl. Conf. on Machine Learning (ICML)*, pages 296–304, 1998.

[3] A. G. Maguitman, F. Menczer, F. Erdinc, H. Roinestad, and A. Vespignani. Algorithmic computation and approximation of semantic similarity. *World Wide Web*, 9(4):431–456, 2006.

[4] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *Proc. WWW*, 2009.

[5] B. Markines, H. Roinestad, and F. Menczer. Efficient assembly of social semantic networks. In *Proc. Hypertext*, 2008.

[6] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Trans. on Systems, Man and Cybernetics*, 19(1):17–30, 1989.