Harvesting Social Knowledge from Folksonomies

Harris Wu
Old Dominion University
Norfolk, VA 23508, USA
1-757-683-4460
hwu@odu.edu

Mohammad Zubair
Old Dominion University
Norfolk, VA 23508, USA
1-757-683-5000
zubair@cs.odu.edu

Kurt Maly
Old Dominion University
Norfolk, VA 23508, USA
1-757-683-4817
maly@cs.odu.edu

ABSTRACT

Collaborative tagging systems, or folksonomies, have the potential of becoming technological infrastructure to support knowledge management activities in an organization or a society. There are many challenges, however. This paper presents designs that enhance collaborative tagging systems to meet some key challenges: community identification, ontology generation, user and document recommendation. Design prototypes, evaluation methodology and selected preliminary results are presented.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering, clustering.*

H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia - navigation.

General Terms

Algorithms, Management, Design, Theory.

Keywords

Collaborative tagging, collaborative filtering, link analysis.

1. INTRODUCTION

Labeling objects with free-style descriptors is called tagging. Collaborative tagging systems, also called folksonomies, allow users to tag documents, share their tags, and search for documents based on these tags. While successful in popular web communities such as flickr.com, a photo sharing website, collaborative tagging systems face many challenges in supporting knowledge management activities. For example, collaborative tagging suffers from idiosyncratic tagging, also called meta noise, which burdens users and decreases the system's information retrieval utility (Wikipedia 2006).

We try to extend the knowledge management capabilities of collaborative tagging systems by improving their abilities in utilizing knowledge from tags. In this article we identify several key design challenges to harvesting social knowledge from tags: community identification, expert and document recommendation, and taxonomy generation. For these challenges we present our design solutions grounded in state-of-art advancements in hypertext analysis and other relevant research fields. The evaluation methodology and selected preliminary results are then presented.

2. BACKGROUND

A collaborative tagging system or folksonomy is defined by wikipiedia as "a collaboratively generated, open-ended labeling system that enables Internet users to categorize content such as web pages, online photographs, and Web links." Collaborative tagging generally refers to the tagging of a collection of documents commonly accessible to a large group, rather than tagging contents located all over the Web, which is instead called social bookmarking (Wikipedia 2006).

2.1 Benefits of Collaborative Tagging

Folksonomies contain individuals' structural knowledge about documents. A person's structural knowledge has been defined as the knowledge of how concepts in a domain are interrelated (Diekhoff and Diekhoff 1982). In a collaborative tagging system, tags codify the knowledge of relationships among documents and concepts represented by the tags. Harvesting individuals' knowledge through folksonomies therefore can benefit the whole society.

Folksonomies claim to have many advantages over controlled vocabularies or formal taxonomies. Tagging has dramatically lower costs because there is no complicated, hierarchically organized nomenclature to learn. Users simply create and apply tags on the fly. Folksonomies are inherently open-ended and therefore respond quickly to changes and innovations in the way users categorize content. Collaborative tagging is regarded as democratic metadata generation where metadata is generated by both the creators and consumers of the content. Folksonomy may hold the key to developing a Semantic Web, in which every Web page contains machine-readable metadata that describes its content. (Wikipedia 2006).

A collaborative tagging system allows a user to search for the content that the user has tagged using a personal vocabulary. As users with similar interests tend to have a shared vocabulary, tags created by one user are useful to others, particularly those with similar interests as the tagger's. Folksonomies cater to the "long tail" (Wikipedia 2006), which describes the mass of users who search for documents using a variety of low-frequency keywords that would have been underserved by controlled vocabularies.

HT'06, August 22-25, 2006, Odense, Denmark.
Copyright 2006 ACM 1-59593-417-0/06/0008...$5.00.

Besides search, perhaps more importantly, collaborative tagging systems assist navigation by providing dynamic hyperlinks among tags, documents and users (online profiles or homepages). Navigable structures such as hierarchies and hyperlinks help overcome searches' limitations. For example, navigation allows casual browsing and leads to serendipitous discoveries. Through tag-based navigation users can discover who created a given tag, and see the other tags that this person has created. In this way a folksonomy user can discover other users with similar interests or perspectives. As such, a collaborative tagging system helps users in not only retrieving information but also socializing with others.

2.2 Challenges

Collaborative tagging faces many challenges to support knowledge management activities in an organization or a society. Organizations need systematic mechanisms of storing and retrieving documents. Although an employee-generated folksonomy can be seen as an emergent knowledge taxonomy, the lack of a document hierarchy prevents it from being widely adopted by enterprises. Folksonomies are criticized to have flaws that formal classification systems are designed to eliminate, including polysemy, words having multiple related meanings, and synonymy, multiple words having the same or similar meanings (Golder 2005). In addition, folksonomies invite deliberately idiosyncratic tagging, also called meta noise, which burdens users and decreases the system's utility. The tags in some large tagging systems have become non-navigable and not even searchable due to the sheer volume of the tags/documents and the low quality of the tags.

In harvesting social knowledge, two key challenges are identifying communities of common interest, and identifying information leaders or domain experts (Huberman 2004). To address these challenges, tags ought to be synthesized with other knowledge sources including contents, hyperlinks, and user behavior such as click streams. Existing tagging systems lack these capabilities.

3. DESIGN COMPONENTS

We developed plug-in modules to extend existing tagging systems' capability of harvesting knowledge from tags. The modules below address three key challenges facing collaborative tagging systems: How to identify communities containing users with similar interests? How to overcome information overload by identifying high-quality documents and users? How to create scalable, navigable structures?

3.1 Community Identification

Much research in the WWW community has been devoted to identifying evolving topical communities of users and/or documents, using information such as hyperlink structure (e.g. Kleinberg 1998) and Web navigations (e.g. Wu et al. 2006). One's interests can be represented by her tagging. Existing collaborative tagging systems (e.g. flickr, del.icio.us), however, lack the ability to identify evolving communities.

Existing community identification techniques can be put into three categories: spectral, bibliometrics, and network flow based (Flake et al. 2003). Spectral methods apply singular value decompositions (or similar operations) to large matrices representing the relationships among elements in a large collection, to identify communities represented by eigenvectors. Spectral methods are "global" methods that attempt to identify all major communities in a large collection. In contrast, bibliometric methods are "local" methods that can determine the pair-wise affinity among users. Network-flow based methods are "hybrid" methods that can identify broader communities containing a known existing community.

Our design below uses a spectral method to identify global communities utilizing authorships and usage of tags and documents. All documents, tags and users are considered as nodes in a network. A link is added from each tag to every associated document. A link is also added from each user to every tag the user has created or accessed, and the documents accessed through the tag. The adjacency matrix of this network represents the associations between users and tags/documents. Applying singular value decomposition to the adjacency matrix, top singular values produce the major topics of user interests. Their associated left and right singular vectors indicate the prominent users and key documents related to these topics respectively. While there can be millions of users, tags and documents, the spectral analysis can reduce their complex associations to identify a small number of topical communities. We are currently experimenting with different ways of assigning weights to the links, to improve the results of the spectral analysis.

3.2 User and Document Recommendation

The ability to find high-quality sources, whether documents or people, is important to overcoming information overload. Collaborative filtering systems, or recommender systems, identify high quality sources utilizing individual's knowledge. It is well known that experts not only tend to use high-quality documents but also develop more elaborate schemata and can better associate documents with concepts. Existing collaborative tagging systems' identification of experts and high-quality documents, however, is limited to simply tallying the number of tags or frequency of usage associate with documents.

The HITS (Kleinberg 1999) algorithm and its various extensions are known to be effective in identifying high-quality sources in a hyperlinked environment. Below we briefly describe the HITS algorithm and our adaptation of HITS that recommends high-quality documents and experts based on tags.

HITS starts from a small root set of documents, such as results from a query to a search engine, to a larger set T by adding documents that link to/from the documents in the root set. The goal of the algorithm is to find hubs, the documents that link to many high quality documents, and authorities, the documents that are linked from many high quality documents. The hyperlink structure among the documents in T, is captured by the adjacency matrix A, where Aij indicates whether there is a link from document i to document j. Using this matrix A, a weighting algorithm repeatedly updates the hub weight and authority weight for each document, until the weights converge. In essence, the hubs and authorities are documents with largest values in the principal eigenvectors of ATA and AAT, respectively. As such, HITS is also a spectral analysis technique.

We have modified the HITS algorithm as follows to obtain experts (hubs) and high-quality documents (authorities) related to a given keyword, based on the usage and structure of tags. The root set of documents includes the documents tagged by the keyword. Then the document set is expanded to include all tags that are associated with any documents in the root set, the documents under these tags and the users who have accessed these tags. The extended set T' includes documents, keywords (tags) and users. We add a link from each keyword to every document tagged with the keyword, and a link from each user to every tag she has assigned or utilized. The hyperlink structure is captured in the matrix A', where Aij indicates whether there is a link from node (a document, tag, or user) i to node j. Because users are sources (nodes with out-going links only) and documents are sinks (nodes with in-going links only), the hubs calculated from the matrix A' are guaranteed to be users and the authorities are documents. Hubs and authorities give the experts and the authoritative documents related to the given tag. We are experimenting with different link weighting mechanisms and combinations with hyperlink analysis to improve the algorithm.

Besides finding quality documents related to a keyword, a user often needs to find documents similar to a given document. To recommend documents similar to a given document, the system first identifies all tags assigned to the given document. Any document related to any of these tags can be represented by a vector using tags as attributes. Pair-wise similarities are then computed between the given document and the rest of the documents. A similar approach is used to find users with interests similar to a given user's.

3.3 Ontology Generation

An ontology, or hierarchy, is one of the most efficient structures for navigation as any document can be reached with an effort of o(log(n)). Ontologies can also assist keyword-based search. An ontology on a document collection not only allows for systematic retrieval of documents but also assists social interactions by providing a common reference.

Tags can be used to generate a common hierarchy for a large set of documents, such as documents from tagging portfolios of a group of users. While a person's tags represent her structural knowledge about the documents that she has visited, a common hierarchy represents a higher form of global knowledge about a larger document collection.

Ontology generation is a hierarchical clustering problem. There are many hierarchical clustering algorithms, most of which are agglomerative (bottom-up) methods. In essence, an agglomerative hierarchical clustering algorithm computes pair-wise document similarities, merges most similar documents into groups, computes group-wise similarities and then merges groups until all documents are in the same group. A hierarchy containing all the documents then results from reversing the merging steps. Our hierarchy generation algorithm is as follows. First, the algorithm identifies the set of documents for which the hierarchy needs to be generated, and identifies all tags associated with these documents. The algorithm then constructs a document-tag matrix, denoted by A. Aij = 1 if and only if document i is tagged by tag j. Based on a thesaurus database, the algorithm constructs a tag-tag matrix to store the semantic similarities between tags. The original document-tag matrix A is multiplied by the tag-tag semantic matrix. Each document is now represented by a row vector Ai. Hierarchical clustering techniques can then be applied to this matrix to generate a hierarchy. Each new category in the hierarchy is labeled by extracting keywords from the tags of all documents in the new category. Different clustering techniques use different pair-wise similarity measures such as cosine and Euclidean similarities, and different group-wise similarity methods such as average linkage and centroid. Wu et al. (2003, 2004) have shown that Jaccard similarity measure and the average linkage method perform better than other alternatives in producing hierarchies from categorical data. Through evaluation we have confirmed that the combination of Jaccard similarity measure and average linkage method performs better, i.e. generate better hierarchies from tags, than other alternatives. Different labeling methods are still being evaluated.

3.4 Discussion

Collectively the above modules consist of a framework that harvests social knowledge from folksonomies. The framework harvests social knowledge from tags (more accurately, associations between tags and documents) as well as hyperlinks and user behavior (such as document usage). All these knowledge sources are associations, and to some extent, functionally equivalent to each other. For example, the association between a tag and a document can be considered as a hyperlink from the tag to the document, and vice versa. The source data matrix in each design is a weighted sum of matrices embedding knowledge from different sources. We are improving the weighting mechanism using machine learning techniques.

Note that synonymy and polysemy in tags do not have a significant impact in any of these design solutions. Let us take the taxonomy generation component for an example. Synonymous tags tend to include the same documents. For a polysemous tag, the documents associated with the tag disambiguate its different meanings. Since the end result of the taxonomy generation is a taxonomy of documents instead of a taxonomy of tags, polysemy and synonymy do not present a significant problem. Similarly are the cases of community identification and expert/document recommendation. As tags are intermediate objects associating users and documents, polysemy and synonymy change the associative routes but not the end results of spectral analyses. It is also worth noting that since all these design solutions are statistical techniques, idiosyncratic tagging, or meta noise, tends to have a negligible impact as the number of users get large. In fact these design solutions are more robust with a larger amount of underlying data.

4. EVALUATION

Are system-identified communities, system-recommended documents and experts, and system-generated taxonomies indeed able to assist users? How effective are our design solutions compared to alternative design choices? To answer these questions, we utilized three types of evaluation: offline studies, test websites, and pilot systems. Below we describe a few examples of the evaluation techniques used in these environments, along with some preliminary results.

We use offline studies as pre-tests of our design concepts. We collect data through paper-based questionnaires and face-to-face interviews. Here we briefly describe one experiment that compares the usefulness of hierarchies generated from tags by different clustering techniques. In this experiment, subjects were asked to tag the same set of documents. After the tags were collected, a common taxonomy was generated using different techniques. Then each subject was presented with four hierarchies for the set of documents, without knowing how these hierarchies were generated. Three of the four hierarchies were generated using different clustering techniques; another one was manually created. The subjects were then asked to evaluate the usefulness of these hierarchies on a 1-5 scale, with 5 as most useful. Further, the subjects were asked to improve the category labels in each hierarchy and then re-evaluate them. The hierarchy generated by our technique performed better than other hierarchies. The results confirmed the effectiveness of our taxonomy generation technique. Offline studies have provided us many good ideas for improving the design without going through implementation.

We use test websites to evaluate selective modules of our design solutions. These test websites may not be designed as collaborative tagging systems, but they have existing users and documents that can be leveraged. We try to avoid negative impacts to users, such as a drastic change of screen layout, when adapting these websites to support collaborative tagging. For instance, one test website was a class website used by graduate students. The class website contains course-related documents contributed by current and past students. We added collaborative tagging functionality prior to the start of the semester. We also added simple online feedback functionality that allows students to rate a document or a tag as "useful" or "not useful". The goal was to evaluate the effectiveness of our expert/document recommendation algorithms. Based on students' tags and click-streams, the system identified subject experts and high quality documents/tags for different keyword searches. The results showed that the high-quality documents and tags identified by the system had higher-than-average user ratings. The experts identified by the system also had higher-than-average scores in objective exams. It appeared that knowledgeable users tended to create high quality tags.

We use pilot systems to evaluate our design in large knowledge creation environments. We are attempting to apply our techniques to the ARCHON digital library (Maly et al. 2002) via a pilot study. ARCHON is widely used by researchers in Physics society.

Our design solutions are exploratory techniques and difficult to evaluate. How to prove that our design solutions are better than alternative design choices? Above we have briefly mentioned comparative evaluation. Overall we have developed a comprehensive approach for evaluating exploratory analyses. Our evaluation methodology consists of internal validation, external validation, scalability and robustness check. For internal validation, we evaluate algorithms and results against the original data. For example, we evaluate our system-generated taxonomy against the underlying tags using the Cophenetic correlation coefficient (Everitt and Dunn 1992), which compare the fitness of clusters generated by our technique against those generated by other methods. For external validation, we ask human subjects to evaluate the results from our design solutions through online feedbacks, questionnaires or interviews. To test the scalability, we simulate large amounts of user input data. For robustness check, we study the impact of statistical sampling and perturbations of the input data.

5. CONCLUSION

Collaborative tagging systems have the potential of becoming a technological infrastructure for harvesting social knowledge. There are many challenges, however. We have designed prototypes that enhance social tagging systems to meet some of the key challenges. We have developed a comprehensive evaluation methodology. Preliminary results show promise.

6. REFERENCES

[1] Diekhoff, G.M., & Diekhoff, K.B (1982). "Cognitive maps as a tool in communicating structural knowledge." Educational Technology, 22(4), 28-30.

[2] G. W. Flake, K. Tsioutsiouliklis, and L. Zhukov. (2003) "Methods for Mining Web Communities" In A. Poulovassilis and M. Levene, editors, Web Dynamics, Springer Verlag.

[3] Huberman, B.A. New ways of identifying and using organizational information. IST News, July 2004.

[4] Kleinberg, J. Authoritative sources in a hyperlinked environment. ACM-SIAM Symposium on Discrete Algorithms, 1998.

[5] Maly, K., et al. (2002). "Archon – a digital library that federates physics collections." DC-2002: Metadata for e-Comunities, October 13-17, 2002.

[6] Walsh, Tony (2004). http://www.secretlair.com/index.php?/clickableculture/entry/what_is_delicious/

[7] Wikipedia, the free encyclopedia. "Folksonomy," "collaborative tagging," "long tail". Retrieved May 3, 2006.

[8] Wu, H., M. Gordon, K. DeMaagd, N. Bos, "Link Analysis for Collaborative Knowledge Building," Proc. of the ACM Hyptertext'03, Nottingham, Aug 26-30, 2003, pp 216-217.

[9] Wu, H. and M. Gordon. "Collaborative Filing in a Document Repository," Proceedings of the ACM SIGIR 2004, Sheffield, UK, July 24-29, 2004, pages 518-519.

[10] Wu, H., M. Gordon, K. DeMaagd and W. Fan, (2006). "Mining Web Navigations for Intelligence," Decision Support Systems, 41(3), pp. 574-591.