December 2006

Supporting a Knowledge Society through Social Tagging

Harris Wu
Old Dominion University

Supporting a Knowledge Society through Social Tagging

Harris Wu
Old Dominion University
hwu@odu.edu

ABSTRACT

Social tagging systems have the potential to become public infrastructure to enhance knowledge workers' productivity. There are many critical challenges, however. This paper presents an ongoing research that attempts to enhance social tagging systems to meet some of these challenges. The paper develops a conceptual framework to illustrate how a social tagging system may help knowledge creation in a society. Key design components of such a system are then presented.

Keywords

Social tagging, social intelligence, design science, social knowledge creation, structural knowledge

INTRODUCTION

Labeling content with descriptive terms, also called tags, helps a person understand and retrieve the content in the future. Many social tagging systems now allow users to share their tags on Web documents, and search for documents based on these tags. There are many challenges for social tagging systems to be effective productivity tools, however. For example keyword searches using tags often return many irrelevant results, as the same tag has been used by different users in various ways.

Using a design science approach, we try to design techniques that extend the capabilities of social tagging systems to meet these challenges. After summarizing the critical challenges faced by social tagging systems, this paper will develop a conceptual framework of how a social tagging system may help knowledge creation on a society level. Several key design components of such a system are then presented. The paper is concluded with next steps of this research.

BACKGROUND

Structure Knowledge

Tags contain individuals' structural knowledge, the knowledge of relationships among documents and concepts. While individuals' structural knowledge is personal and tacit, it can be elicited and codified into tangible structures such as networks, hierarchies, or keyword-labeled categories. Once codified, individuals' knowledge can be shared with the whole community.

One's structural knowledge is limited to the collection of documents that she has visited. Effective knowledge sharing in a large community requires aggregating individuals' local knowledge into a higher form of global knowledge. While the existing social tagging systems have made the individuals' tags publicly available, they have not been able to effectively filter, distill and synthesize the tags. The challenge of aggregating tags into global knowledge relates to ontology management, a field that deals with mapping, merging, evolving, sharing and querying knowledge representations (Madeche et al. 2003).

Information Retrieval

Social tagging provides new structural and semantic information for Web information retrieval. Tags enable keyword search of non-textual documents such as photos. Utilizing human assigned keywords can improve precision and recall. Current tagging systems, however, lack the capability to identify high-quality documents when returning search results.

Besides search, tagging systems assist navigation by providing dynamic hyperlinks to tags (document categories), users and the documents tagged by these users. Navigable structures, particular hierarchies and hyperlinks, help overcome searches' limitations (Labrie 2003). However, the growing volume of tags along with their characteristics such as polysemy and synonymy make tag-based navigation increasingly difficult. The tags in many popular tagging systems have become non-navigable due to the sheer volume of tags and the low quality of them. How to create high-quality navigable structures out of tags is a challenge.

Community and Expert Identification

"Who you know" is often more important than "what you know." In harvesting organizational or social knowledge, two key issues are identifying communities of common interest, and identifying information leaders or domain experts (Huberman 2004). Much research in the WWW community has been devoted to identifying evolving communities based on the hyperlink structure (e.g. Kleinberg 1998) or Web navigations (e.g. Wu et al. 2006). One's interests can be represented by her tagging. However, existing social tagging systems lack the ability to identify evolving communities.

It is well known that experts develop more elaborate schemata and can better associate documents with concepts. However existing social tagging systems' identification of experts is limited to simply tallying the number of tags owned by the user or the number of positive feedbacks on tags. This approach often mistakes hyperactive users as experts.

Challenges

From above we have seen that social tagging faces critical challenges such as: how to merge individuals' structural knowledge? How to generate navigable structures using tags? How to identify evolving communities? How to identify experts? The conceptual framework below guides the design for a social tagging system as a productivity tool.

CONCEPTUAL FRAMEWORK

Different social tagging systems have different goals. For example, Flickr.com tries to help users store, share and search for photos. The research in this paper aims to improve social tagging systems' capability as a productivity tool to support knowledge workers on the Web. From a public utility perspective, the ultimate goal of social tagging and other public information infrastructure is to facilitate knowledge creation in a society. Other knowledge processes including storage, retrieval, transfer and application of existing knowledge can all be considered as part of the continuous cycle of knowledge creation.

Nonaka (1994) developed a rich conceptual framework for organizational knowledge creation. In his spiral model (Figure 1), knowledge is created through a cycle of four intertwining modes of conversion between tacit and explicit knowledge: externalization, internalization, socialization and combination. Externalization refers to explication of tacit knowledge into explicit knowledge, which corresponds to the traditional notion of codification. Internalization refers to conversion of explicit knowledge into tacit knowledge, which corresponds to the traditional notions of learning, understanding or sense-making. Socialization refers to creating tacit knowledge through social interactions and shared experience. Combination refers to creating explicit knowledge from explicit knowledge, through merging, categorizing, sorting, and re-contextualizing.

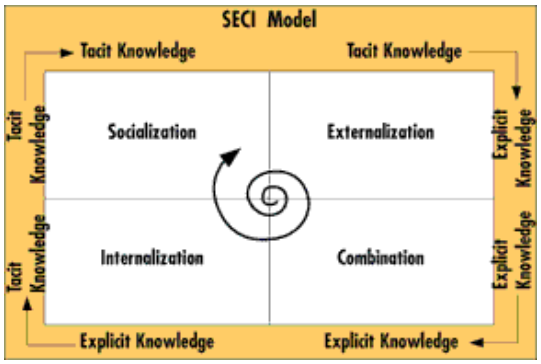

Figure 1. Nonaka's Knowledge Creation Model.

Utilizing Nonaka's framework, we have developed a model of how a social tagging system may assist in all four modes of knowledge creation:

Externalization: A user's tacit knowledge about the relationship of documents can be externalized into her tags.

Internalization: Assigning tags to documents can help users learn from the documents and understand the domain. A user's tags can help other users locate information and transfer structural knowledge to others. Tags can be utilized to identify high-quality documents, which make learning more effective.

Socialization: Sharing tags can help users get to know each other, identify experts and find like-interested peers. Tags can be used to identify and support communities.

Combination: Personal tags can be merged or connected, or otherwise combined to form a higher level of explicit knowledge about the relationship of documents. Personal tags can be utilized to generate hierarchies or hyperlink networks.

The above framework leads to various design requirements and research hypotheses for a social tagging system, some of which will be discussed in the next section.

KEY DESIGN COMPONENTS

We are developing a social tagging system, with requirements derived from the above social knowledge creation framework. Figure 2 illustrates the key design components and how they address the four modes of knowledge creation in Nonaka's framework. As design science is inherently iterative (Hevner et al. 2004), these solutions are still being refined and evaluated. Our goal is not to replace existing social tagging systems, but rather, to enhance them with support for a knowledge society. Below we focus on three key server components of the system.

Externalization	User Interface (allowing users to create tags)
Socialization	Community Identification
Internalization	Expert and Document Identification
Combination	Ontology Generation

Figure 2. Four modes of knowledge creation and supporting design components

Community Identification

The system needs to detect major topics of user interests and connect users with similar interests. Existing community identification techniques can be put into three categories: spectral, bibliometrics, and network flow based (Flake et al. 2003). Spectral methods are "global" methods that can identify communities from a global perspective. Bibliometric methods are "local" methods that can determine the pair-wise affinity. Network-flow based methods are "hybrid" methods that can identify broader communities containing a known existing community.

Our system uses a spectral method to identify global communities. 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 that the user has accessed or assigned. We are experimenting with different ways of assigning weights to these links. Applying singular value decomposition to the weighted adjacency matrix of this directed network, top singular values produce the major topics of user interests. Their associated singular vectors indicate the prominent users and documents related to these topics. The system also utilizes a network-flow based method to identify broader communities containing an existing user community.

Expert and Document Recommendation

Information overload is a formidable challenge. The HITS (Kleinberg 1999) algorithm and its various extensions are known to be effective in identifying authoritative sources in a hyperlinked environment. Below we briefly describe the HITS algorithm and how our system applies HITS to social tagging.

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 and are linked from the documents in the root set. The hyperlink structure among the documents in T, is captured by the 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, documents that are linked from many other documents are authoritative documents, and documents that link to many authoritative documents are hubs. The hubs and authorities are documents with largest values in the principal eigenvectors of A^TA and AA^T, respectively.

The HITS algorithm is modified 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 used or assigned. The hyperlink structure is captured in the matrix A', where A_{ij} 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.

There are also needs for finding documents similar to a given document. To identify the documents similar to a given document, we first identify all tags assigned to the given document. Each document related to any of these tags can be represented by a vector using tags as attributes. Pair-wise similarities can be 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.

Ontology Generation

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 with overlapping interests. While a person's tags represent her local knowledge about the documents that she has visited, a common hierarchy represents a higher form of global knowledge about a large document collection useful to a community. Navigating a hierarchy is highly efficient, as any document can be reached with an effort of o(log(n)).

Hierarchy 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 system generates the hierarchy 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. $A_{ij} = 1$ if and only if document i is tagged by tag j. Each document is represented by a row vector A_i. Hierarchical clustering techniques can now 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) has shown that Jaccard similarity measure and the average linkage method can produce high quality hierarchies from categorical data. Different similarity measures, clustering techniques and labeling methods are being evaluated in this ongoing research.

CONCLUSION

Social tagging systems have the potential to become public infrastructure to enhance knowledge workers' productivity. There are many challenges, however. Using a design science approach, we try to design techniques that enhance social tagging systems to meet these challenges. We have developed a conceptual framework of how social tagging may help knowledge creation in a society. Utilizing this framework, we have identified several key design requirements and technical solutions. We are refining and evaluating the solutions by implementing a social tagging system for researchers in an academic setting. We plan to evaluate the solutions on a larger scale by partnering with social tagging systems in the public domain. A prototype may be demonstrated at the conference.

REFERENCES

1. G. W. Flake, K. Tsioutsiouliklis, and L. Zhukov. (2003) "Methods for Mining Web Communities: Bibliometric, Spectral, and Flow." In A. Poulovassilis and M. Levene, editors, Web Dynamics, Springer Verlag.
2. Huberman, B.A. New ways of identifying and using organizational information. IST News, July 2004.
3. Hevner A.R., March S.T. and Park, J. (2004) "Design Science in Information Systems Research," MIS Quarterly vol.28 No. 1, pp. 75-105.
4. Kleinberg, J. Authoritative sources in a hyperlinked environment. ACM-SIAM Symposium on Discrete Algorithms, 1998
5. LaBrie, R. C. (2003). Investigation of information retrieval accuracy from knowledge management systems: a multi-method approach. Americas Conference on Information Systems, New Orleans, AIS.
6. Maedche, Alexander, et al. "Ontologies for Enterprise Knowledge Management," IEEE Intelligent Systems 18(2) (2002), pp 26 – 33.
7. Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. Organization Science (5), pp 14 - 37.
8. H. Wu, M. Gordon, K. DeMaagd, N. Bos, "Link Analysis for Collaborative Knowledge Building," Proceedings of the ACM 14th Conference on Hypertext and Hypermedia (Hyptertext'03), Nottingham, UK, August 26-30, 2003, Pages 216-217.
9. H. Wu 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. H. Wu, M. Gordon, K. DeMaagd and W. Fan, (2006). "Mining Web Navigations for Intelligence," Decision Support Systems, 41(3), pp. 574-591.