Semantic Imitation in Social Tagging

WAI-TAT FU, THOMAS KANNAMPALLIL, RUOGU KANG, and JIBO HE University of Illinois at Urbana-Champaign

We present a semantic imitation model of social tagging and exploratory search based on theories of cognitive science. The model assumes that social tags evoke a spontaneous tag-based topic inference process that primes the semantic interpretation of resource contents during exploratory search, and the semantic priming of existing tags in turn influences future tag choices. The model predicts that (1) users who can see tags created by others tend to create tags that are semantically similar to these existing tags, demonstrating the social influence of tag choices; and (2) users who have similar information goals tend to create tags that are semantically similar, but this effect is mediated by the semantic representation and interpretation of social tags. Results from the experiment comparing tagging behavior between a social group (where participants can see tags created by others) and a nominal group (where participants cannot see tags created by others) confirmed these predictions. The current results highlight the critical role of human semantic representations and interpretation processes in the analysis of large-scale social information systems. The model implies that analysis at both the individual and social levels are important for understanding the active, dynamic processes between human knowledge structures and external folksonomies. Implications on how social tagging systems can facilitate exploratory search, interactive information retrievals, knowledge exchange, and other higher-level cognitive and learning activities are

Categories and Subject Descriptors: H.1.2 [Models and Principles]: User/Machine Systems— Human information processing; H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces—Social tagging; J.4 [Social and Behavioral Sciences]: Psychology— Semantic representation

General Terms: Theory, Experimentation, Human Factors

Additional Key Words and Phrases: Semantic imitation, human information processing, cognitive models, social tagging, semantic representations, multilevel models

This work is supported in part by a grant from the National Science Foundation (0819840), the Office of Naval Research (N00014-07-1-0903), and the Human Factors Division and the Beckman Institute of the University of Illinois at Urbana-Champaign.

Authors' address: W.-T. Fu, Department of Computer Science, Human Factors Division, and Beckman Institute of Science and Technology, University of Illinois at Urbana-Champaign, IL; email: wfu@illinois.edu; T. Kannampallil, R. Kang, and J. He, Human Factors Division, University of Illinois at Urbana-Champaign, IL; email: {tgk2,kang57,jibohe}@illinois.edu.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org. © 2010 ACM 1073-0516/2010/07-ART12 \$10.00

DOI 10.1145/1806923.1806926 http://doi.acm.org/10.1145/1806923.1806926

ACM Reference Format:

Fu, W.-T., Kannampallil, T., Kang, R., and He, J. 2010. Semantic imitation in social tagging. ACM Trans. Comput.-Hum. Interact. 17, 3, Article 12 (July 2010), 37 pages. DOI = 10.1145/1806923.1806926 http://doi.acm.org/ 10.1145/1806923.1806926

1. INTRODUCTION

Social tagging systems allow users to annotate, categorize, and share Web content (links, papers, books, blogs, etc.) using short textual labels called tags. Tags help users in organizing, sharing, and searching for Web content in shared social systems. Some popular social information systems that support tagging include del.icio.us and Bibsonomy (for bookmarks), Flickr (for photos), and CiteULike (for research articles).³ The inherent simplicity in organizing and annotating content in these systems through "open-ended" tags satisfies a personal and social function [Ames and Naaman 2007; Thom-Santelli et al. 2008]. At a personal level, customized tags can be added to a resource based on a specific information goal (e.g., mark up for future reading or identifying books for a history library) that will help in organization of resources or future search and retrieval. At the social level, the tags facilitate sharing and collaborative indexing of information, such that social tags act as "way-finders" for other users with similar interests to search for relevant information [Millen et al. 2007; Fu 2008; Fu and Kannampallil 2009; Fu et al. 2009, 2010; Fu and Dong 2010a, 2010b; Pirolli 2009; Kang and Fu 2010].

Users exploit the open-endedness in the creation and addition of tags in social tagging systems based on their social or personal needs. However, this leads to a potential vocabulary problem in social tagging systems [Furnas et al. 1987], as the open-endedness may lead to the creation of a large number of diverse tags to describe the same resource. Furnas et al. predicted that, because of the lack of top-down control in social information systems, people could use a wide variety of words to describe the same objects, which could create significant challenges for human-system communication. In the case of social tagging systems, Furnas et al. found that users apply "different terms as tags to describe the same resource" by using synonyms, homonyms, and polysemes, leading to multiple and diverse descriptions for the same resource. The increasing number of vocabularies may imply that the connections between tags and documents will become less and less distinct, making information retrieval more difficult.

Besides different choice of words, another possible contributor for the vocabulary problem is the diversity of the information goals (i.e., the motivation for why users are looking for information using the system) under which users create tags for resources [Sen et al. 2006; Downey et al. 2008].⁴ It has been

¹http://del.icio.us; www.bibsonomy.org

 $^{^2} http://flickr.com$

³http://citeulike.org

⁴We refer to the task-specific sense of the term "information goal," which defines the set of topics or concepts that the information seeker is interested in when interacting with the system. The broad sense of the term could refer to many other related goals such as emotional or social engagement

argued that users may use different words to describe the same document based on their own interpretation of the content [Furnas et al. 2006], or to serve different information goals of the users. For example, a user may tag a book based on its content with a tag "Star Trek"; while another user may tag the same book as "to read," referring to a personal goal regarding the book; while a third user may tag it as "science fiction," based on the genre of the book. While all of these tags and the related information goals are acceptable within the context of the social tagging system, the relative difference between these tags in terms of its meanings, purposes, and goals leads to the aforementioned vocabulary problem in social information systems. In other words, the different cognitive states of the users (induced by their task-specific information goals) may map to the same document through a set of diverse tags that may or may not be coherent among themselves. Surprisingly, while researchers [Macgregor and McCulloch 2006; Ames and Naaman 2007] have emphasized the potential impact of information goals on tagging, there has been no systematic investigation of the effects of information goals imposed by the tasks on social tagging. One possible reason for the apparent lack of focus on the effect of information goals is that most previous research on social tagging was based on analysis of a snapshot of the content of an existing system. This approach, although realistic, lacks the control on the diversity of users' information goals, making it impossible to study the direct effects of information goals on tagging across time from multiple users.

In spite of the vocabulary problem we have described, there has been accumulating evidence suggesting that emergent structures do exist in social tagging systems [Golder and Huberman 2006; Cattuto et al. 2007], suggesting that the vocabulary problem may not be as detrimental to users to search for information as previously suggested. Most importantly, these emergent structures do seem to have the potential to help users to explore for information by providing meaningful organization and indexing to the information resources. For example, Golder and Huberman [2006] found that the frequency of occurrence of any particular tag tended to remain a fixed proportion of all the tags that were used. In other words, in spite of a large number of users and an often diverse set of tags, tag proportions remained relatively stable in the social tagging system (i.e., each tag's frequency is a fixed proportion of the cumulative frequency of all tags used). These stable usage patterns are important because they partially validate the usefulness of social tags in annotating information content, as it suggests that tags are at least not completely "noisy," and indeed may act as useful information cues that facilitate information search. The question that remains is: why do these emergent patterns exist?

The most generally accepted view on the emergent patterns in social tags is the "social influence" perspective. Golder and Huberman [2006] used data from del.icio.us to show that tag choices are influenced directly by tags created by previous users for the same resource (a web page in the case of del.icio.us).

needs, which we do not focus on in the current article. However, we believe the current model is in general consistent with the social nature of information goals as well.

They argued that imitation occurs as a result of the presentation of previously created tags to users. Using an evaluation of the del.icio.us and BibSonomy tag networks, Cattuto et al. [2007] found that despite the diverse backgrounds and information goals of multiple users, cooccurring tags exhibited hierarchical structures that mirrored shared structures that were "anarchically negotiated" by the users. Mostly importantly, they found that some of the patterns were sensitive to the semantics of tags. Specifically, they found that semantically general tags, such as "blogs," tended to cooccur more often with other tags and they tended to stay in the system longer. In contrast, semantically narrow tags, such as "ajax," tended to cooccur less often with other tags. Cattuto et al. presented a memory-based language model to explain the patterns they found. They argued that the emergent patterns in social tags were similar to how words were naturally used in human communications. The results suggest that the naturally shared semantic structures among users could be one reason for the social influence of tags. In other words, even though people may be using different words to describe the same document, the underlying semantic structural relations and contents of these words may be similar, as reflected by the aggregate patterns of tags in the system.

To summarize, previous research seems to suggest that stabilization in tag choices are caused by two main factors: (a) the information goals of the users (i.e., what the user is looking for), and (b) the social influence of tags (i.e., how tags created by others influence future tag choices). Although previous models did implicitly assume the social influence of tags on other users as the major reason for the formation of emergent structures in the system, what is still lacking is direct evidence demonstrating the social influence of tags, and what is the nature of this social influence. It is still not clear, for example, whether there are other variables, such as differences in information goals, that moderate the social influence. To this end, we designed an experiment to compare tagging behavior of two groups of users who can and cannot see tags created by others when using a social tagging system.

1.1 The Current Approach: From Individual to Social Behavior

In this article, we present a semantic imitation model to predict the interactive effects of information goals and social influence of tags in a social tagging system. The purpose of the model is to investigate the plausibility of semantic imitation as one of the important cognitive processes during social tagging. The model is derived from prior work, based on an empirical investigation of the role of information goals and social influence in the emergence of stable tag usage patterns. Based on predictions of the model, we conducted a controlled experiment to directly manipulate information goals and the availability of social tags to study their effects on social tagging behavior. We then analyzed the behavioral data as well as the network-based data generated in different experimental conditions to directly test the predictions of the model.

Our study is different from prior research in this domain in two main ways. First, an important contribution of the current study is to understand whether

content (i.e., semantics) of tags play a critical role in social tagging behavior. Previous models were primarily based on word-level analyses of the organization of tags (e.g., how likely the same tag was reused across time, or the informational value/entropy of word-document pairs, etc.). To our best knowledge, there is a general lack of research, analysis, or models that directly study the relations among the resource contents (meaning) and the semantics of tags in a social tagging system. Our primary hypothesis is that when a user reads a tag, a spontaneous semantic interpretation process will be invoked, which then primes (and thus constrains) the comprehension process and subsequently the tag choice process. When users share similar semantic representations, either because they have similar background knowledge structures (e.g., if they are all biologists or computer scientists) or information goals (e.g., that they are all looking for books related to a certain topic), their tag choices may tend to converge, leading to emergent behavioral patterns observed in large-scale social tagging systems. Therefore, even when the user is not reusing the exact same tag, he or she may still be influenced by the existing tags at the semantic level, and thus create tags that are semantically related to the existing tags. We will present details of this model in the next section.

Second, we believe that much could be learned by assuming that in a collaborative system such as a social tagging system, an essential part of user communications occurs at the semantic (concept) level, much like the level of representations in most other communication methods among humans (e.g., we may use seemingly arbitrary symbols/sounds to communicate, but the semantics to which these symbols refer to tend to be relatively consistent within groups/cultures). Therefore, we believe that by including the semantic level in the analysis, one could gain much insight into the complex interactions among the users and the social tagging system. In fact, we believe that a model that includes cognitively plausible representations of tags, resources, and processes will be essential for understanding individual behavior, as well as how they may provide reasonable constraints for understanding the dynamic emergent behavioral and network patterns observed in a social information system.

Although there has been much work done in the analyses of fine-grained individual human-computer interactions on the one hand and aggregate behavioral patterns of large-scale social information systems on the other, research that aims at bridging the gap between the two is scant. We believe that the current model will be useful in bridging this gap, which is critical for explaining how differences in individual behavior will lead to different emergent behavioral patterns, and is essential for providing concrete guidelines for designing interface representations and interactions methods to achieve the desirable "social" effects of these systems. One should note that although emergent patterns in a social information system reflect the aggregate behavior of a massive number of users, when individual users interact with the system, the interactions are inherently "local," in the sense that factors that influence individual users are still much influenced by the interface representations and interaction methods experienced by each individual user. Thus, most higher-level emergent

behavior in the system can be traced back to the aggregation of distributed "local" computations occurring at the individual levels [Anderson 2002; Hedstrom 2006]. In other words, a better understanding on the local interactions at the individual or small-group level will provide a solid theoretical foundation for explaining aggregate patterns at the social level.

To summarize, our approach is different from most of the prior research on social tagging systems: (a) we attempt to demonstrate not only the existence of the social influence of tags, but also the nature of the influence using a psychologically plausible mechanism that explains behavior at the individual level, and (b) we use a human information processing perspective to report on how individual semantic representations influence the choice of social tags, and how they explain aggregate behavioral patterns in social tagging systems. As opposed to prior research that uses large corpuses of tagging data to report on properties of overall tagging behavior, our approach is to conduct a laboratory experiment to study the role of semantic structures in tagging behavior, (c) we directly manipulated the information goals (which would be impossible to ascertain during the analyses of corpora tagging data) of taggers to study their effect on the social tagging process, (d) we directly study the social influence of tags by comparing effects of tag choices from a social group (where tags are visible to other users) to a nominal group (where tags are not visible to other users), and (e) we use a novel comparison between social network properties (such as connectedness and tag cooccurrences, [Albert and Barabási 2002] and statistical measures of tag semantics (Latent Semantic Analysis, [Landauer and Dumais 1997]) as a baseline for establishing the emergent semantic knowledge structures in the social tagging networks created by the participants in the study. At the outset, we want to stress that our goal is not to show that our model is exclusive of previous models or theories on social tagging. In fact, quite the contrary, we believe that alternative theories and approaches, especially from a cognitive perspective, will help us in developing deeper insights that will complement our current knowledge of social tagging systems.

2. BACKGROUND

In this section, we will first describe the role of semantic interpretation in social tagging and then we introduce the semantic imitation model. The formal and graphical representations of social tagging systems adopted in the current analyses are then described. These representations allow measures and comparisons of the structural relationships among users, tags, and resources between social tagging systems. We will then describe how these network representations can be applied to test the predictions of the semantic imitation model.

2.1 Semantic Interpretation of Social Tags

An intriguing feature of social tagging systems is that they can be considered platforms for dynamic interactions of diverse semantic structures among users [Cattuto et al. 2007; Fu et al. 2009]. It is therefore intuitive to assume that

tag choices are influenced not only by existing tags, but also by the semantic interpretation of tags and the associated content by the users. By assuming that semantic interpretation will influence tag choices, one can broaden the analysis by going beyond the statistical structures *within* the system to include the active role of individual users. Indeed, if features of social tagging systems can influence higher-level knowledge structures of users, social tags may not only provide annotation to web contents, but they may also have the potential to play an active role in facilitating exchange of knowledge structures among users [Fu 2008; Fu et al. 2009, 2010].

There has been a long history of research on semantic interpretation in the research areas of reading comprehension and information extraction. Research on reading comprehension assumes that as a person reads text, words invoke corresponding semantic representations to allow the person to extract meaningful information contained in the text [Kintsch 1998]. This kind of spontaneous semantic interpretation of words is perhaps best illustrated by the experiments on "false memories" [Roediger and McDermott 1995]. A typical false memory experiment would show that when people were asked to remember a list of semantically associated words that converged on a nonstudied word, people tended to falsely remember the nonstudied word. For example, after studying the list consisting of thread, pin, eye, sewing, sharp, point, pricked, thimble, haystack, pain, hurt, and injection, people often erroneously recalled the converging nonstudied word needle in the list. This kind of "memory illusion" is often interpreted as evidence supporting the notion that as people process a list of words (or tags, when they are browsing a social tagging system), they spontaneously activate the corresponding semantic representations for those words. When people try to recall the list of words, the converged semantic representation will again be activated to exert a top-down influence on memory recall. As the false-memory experiments showed, because the nonstudied word was representative of the converged semantic representation, it was often erroneously "recalled."

Results from these experiments demonstrate that people tend to naturally encode semantic representations of words during comprehension. In fact, the semantic representation of information is often taken as one of the major defining characteristics of human information processing, as significant information reduction can be achieved only when the most "critical" information is stored in our memory system [Anderson 1974]. This kind of semantic abstraction has a long tradition of being the most commonly accepted forms of knowledge representation in human information processing [Anderson 1974; Frederiksen 1975; Norman and Rumelhart 1975; Kintsch 1998]. The implication is that the critical level of analysis of human communication is often not at the word level, but at the *semantic* level at which meanings are communicated, interpreted, and exchanged. This is also one of the major differences in the way how humans and machines index information: Humans tend to abstract away details at the word level to spontaneously derive the semantic representations of words when indexing information in the world, while machines tend to process at the word level to derive the associations between word and documents/objects in the world.

2.2 A Semantic Imitation Model of Social Tagging

The semantic imitation model for social tagging consists of two important components: a topic inference process and a topic extraction process.⁵ As the user navigates through a social tagging system, while searching for content, tags created by other users will help them interpret whether a particular resource (e.g., a bookmark) is relevant to his or her current information goals. The set of tags assigned to the bookmark will act as retrieval cues for relevant topics (or concepts) represented by these tags. We call this the tag-based topic inference process. For example, consider a resource, a URL from WebMD on Obesity. It is likely that the associated tags for this URL could include a tag "health." On seeing this tag (i.e., "health"), the user can be semantically primed to think about related tags such as nutrition," "diet," or "exercise." This inference process based on the interpretation of tags associated with a resource is referred to as topic inference. Thus, the process assumes that the topics inferred from the tags will allow the user to predict the information content of the associated resource, as well as to provide some form of semantic priming of related concepts when the user processes (comprehends) the information in the resource [Kintsch 1998].

The second stage is the topic extraction process. It is assumed that the user extracts the concepts (topics) that describe the contents of the document. The topic extraction process is influenced (i.e., biased) by the initial tag-based topic inference. In effect, during comprehension, the user combines the topics from the document and those extracted from the tags (created by previous users). We assume that when a user processes a resource, he or she will engage in a process of topic extraction to comprehend the associated information content [Pirolli 2004; Fu 2008]. Associated information content can include abstracts of papers (in CiteULike) or overviews of Web URLs (del.icio.us), or the complete content of a Web page. Following from the discussions above, semantic priming based on existing tags may bias the user during topic extraction from the document. Specifically, the model assumes that topics activated by existing tags may prime users to allocate more attention to related topics when they process the associated resources. Thus, with all else equal, the topics extracted will be biased to topics that are semantically primed by the tags that are currently present in the resource. Assuming that users will assign tags to best represent the topics extracted from the resource, the model predicts that tag choices will be semantically similar to existing tags. This process of topic extraction has a long history in the literature on reading comprehension and human memory [Anderson 1974; Kintsch 1998]. This process is shown in Figure 1.

The initial tags (T_1, T_2, T_3) induce a process of tag-based topic inference resulting in the identification of three topics (or concepts): C_1 , C_2 , C_3 (i.e., topic inference). The user then interprets the actual contents of the resource. The process of identifying the topical contents of the resource document is influenced by both the user's understanding of the content and by the topics identified in the tag-based topic inference process. The end result of this process

 $^{^5}$ A mathematical version of the semantic imitation model can be found in Fu et al. [2009].

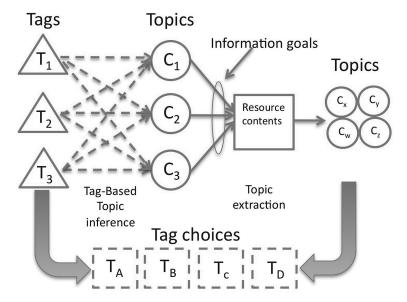


Fig. 1. A model of semantic imitation in social tagging. The model assumes that existing tags will invoke a tag-based topic inference process, which will bias the extraction of gist from the resource and semantically prime the later tag choice process. The effect of information goals on the gist extraction process is mediated by the semantic interpretation of existing tags. In the figure, existing tags (T_1, T_2, T_3) act as cues for related topics (C_1, C_2, C_3) in the topic inference process, and later lead to extraction of gist concepts C_w , C_x , C_y , and C_z .

is a set of concepts (C_w, C_x, C_y, C_z) that influences the tag choice (T_A, T_B, T_C, T_D) of the user. The current model of semantic imitation aims at extending existing models of social tagging by assuming that semantic interpretation of tags plays an important role in the users' tag choice behavior. In fact, word imitation can be considered a special case of semantic imitation, in which one can assume that the exact word is reused to represent the same semantic content (which clearly is a more restrictive model).

In addition to semantic representations of tags, another major assumption of the model is that the interaction effect between information goals and the social influence of tags in the topic extraction, and eventually, the tag choice process. Prior research has shown that information goals did show a strong influence on tag choices [Ames et al. 2007]. For example, if a user is browsing for books about retirement, he or she may be biased to extract topics that are more relevant to retirement (such as health, travel, etc.) and assign tags that are directly related his or her information goal.

An important prediction of the semantic imitation model is that the *dynamic interaction* between the influence of semantic representations of tags and information goals on tag choices. When users can see tags created by others, those users who have the same information goal will tend to be more influenced by the semantic interpretation of tags, which will more likely lead to a shared semantic representation among the users. Thus, tags created under such a

situation will more likely be semantically similar. In contrast, when users cannot see tags created by others, tag choices will be sensitive only to the mapping of the information goal of the user to the gist of the resource, and thus tags will less likely be semantically similar.

In summary, the semantic imitation model makes the following predictions that we test using our experimental study:

- 1. Users who can see tags created by others (i.e., exposure to social tags) will tend to converge in their choice of tags over time; but users who cannot see tags created by others may create tags that tend to diverge.
- 2. With all else equal, users who can see tags created by others will create tags that are more semantically similar than users who cannot see tags created by others.
- 3. Effects of social tags and information goals will interact to influence users' tag choices. Specifically, it is predicted that the effect of information goals will moderate the semantic interpretation of existing *tags*.

2.3 Testing the Semantic Imitation Model

Although most studies on social tagging adopt the approach of analyzing usage patterns in a large-scale system over a significant period of time, this approach cannot be easily applied to directly test the predictions of the semantic imitation model. One major difficulty is that we do not have direct control over the information goals of the users. We, therefore, chose to perform a laboratory study to allow for full control in the manipulation of information goals and their effects on the various properties of the social tagging network to directly test the prediction of the model.

We compared behavior between one group of users who can see other users' tags during the task (as in most social tagging systems) and another group of users who cannot see tags created by other users to study the social influence of tags on behavior. Indeed, we believe that a direct manipulation of information goals and presentation of social tags will provide more direct evidence on their effects on social tagging. To verify that the data collected are representative of those found in large-scale networks, we conducted several analyses to show that the data exhibited the similar properties. We will summarize the measures to be used in the next subsection, before we introduce the details of the experiment.

2.4 Analysis of Social Tagging Behavior

In this section, we will describe the specific data analysis techniques that were used. We will focus on how existing methods can be combined with the technique of Latent Semantic Analysis (LSA) to test the predictions of the model. We will first briefly review the widely accepted formal and graphical representations of social tagging systems and their corresponding network analysis methods, followed by a brief discussion of the LSA technique. We will then discuss how we can combine the two analysis techniques to direct test the predictions of the model.

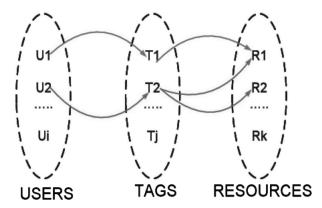


Fig. 2. Formal representation of tagging by the triples defined in the users, tags, and resources spaces. User 1 (U1) has one tag (T1) applied to one resource (R1), while User 2 (U2) has one tag (T2) applied to two resources (R1) and (U2), (R1), (R1) respectively.

2.4.1 Formal and Graphical Representation of a Social Tagging System. There are three main components for any tagging system: a set of users, a set of tags created by the users and resources (URL, books, pictures, movies etc.). Resources can be different depending on the specific purpose of the social tagging system. A social tagging system can be represented by a triple of users, tags and resources: (U, T, R) [Halpin et al. 2007]. For example, consider that a user U1 applies tag T1 to a resource R1 and user U2 applies the tag T2 to two resources R1 and R2. In this case, there are three tag applications in the system (U1, T1, R1; U2, T2, R1; U2, T2, R2). Large social tagging systems contain thousands of such tag applications resulting in interesting patterns and interactions. This conceptual model of a tagging system is shown in Figure 2.

The (U, T, R) triple can be represented in graph-theoretical terms as a tripartite graph of user, tag and resource. The nodes of the network can represent the tag, resource or users while the edges would represent their cooccurrence. The tripartite graph can be reduced to represent multiple relationships (e.g., between tags and resources; resource and users; tags and resources; tags and users). Figure 3 shows an example of this reduction. This bipartite graph consists of a set of 4 resources (e.g., books) and a set of 11 users (or tags, if tags were considered instead of users). The links between the resources and users (or tags) indicate the selection of the resource by multiple users (assignment of tags to different resources, if tags were used instead of users in the bipartite graph). This bipartite graph can be further reduced to a unipartite graph (bottom of Figure 3). For example, given that resource 1 was selected by 5 users A, B, C, D, and E (or 5 tags were assigned to resource 1), these five nodes (users or tags) will be fully connected to each other in the unipartite graph as they are all linked to the same resource. Similarly, because nodes B, D, F, and G are linked to resource 2, they are fully connected to each other in the unipartite graph.

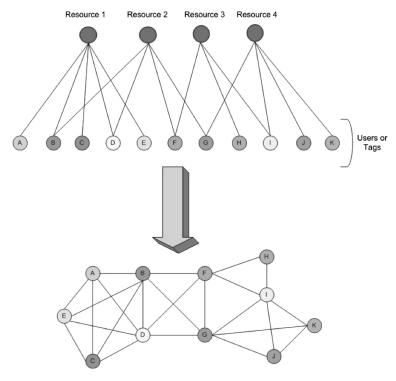


Fig. 3. The top part shows a bi-partite set of connections. The bottom half shows the corresponding network obtained by linking nodes (tags or users) that are connected to the same resources.

We use two main concepts from graph theory: node and edge. *Node* refers to a vertex in the network, while an *edge* is a link or connection between two nodes. An edge is added between two nodes when there is a relationship between the nodes. In our case, the nodes represent the individual tags and an edge is added between two tags when they cooccur on a resource. Several standardized calculations can then be performed to understand the general properties of a network.

In this paper, we computed the node degree and the clustering coefficient. The degree of a node is the number of links (edges) that are connected to that node. A higher average node degree (across all the nodes) indicates that there are many edges and nodes in the network. The distribution of these nodes across the entire network is often called the degree distribution. Networks often exhibit the property of forming cliques or clusters. Clustering coefficient is used to express the degree of connectedness of a network. It is given by the proportion of edges between the nodes within its neighborhood divided by the number of edges that could possibly exist between them. A high clustering coefficient therefore indicates that the nodes are highly clustered together by edges. Detailed descriptions of each of these parameters can be found in Newman et al. [2006] or Albert and Barabasi [2002].

While graphical analysis provides a visual perspective with easy switching between overall (global) views and microstructure (local) views, the associated network analysis methods are successful in measuring various properties of complex networks [Halpin et al. 2007]. The properties of graphical networks can be used to develop insights into some basic characteristics of tags. However, the connections between tags are typically established at the word level. Thus, this kind of network analysis typically does not take into account any semantic similarities between tags, and therefore is not sufficient to study the social influence of tags at the semantic level. We therefore decided to combine the network analysis with LSA to measure how well the network structures are related to the semantic similarities of tags. The main goal is to see whether we can derive new insights into tagging behavior by analyzing the role of semantic representations of tags in the network.

2.4.2 Intertag Cooccurrence Index (CI). As tags are created and reused by different users, we can develop a measure of their relationships based on their cooccurrence with other tags and their rank frequency. The cooccurrence index between any pair of tags can be defined using a "similarity" measure based on the number of times users choose to use these tags together in a resource. For example, let us consider two tags T_i and T_j . A simple cooccurrence index can be used to define the cooccurrence index between these two tags.

$$CI(T_i, T_j) = \frac{N(T_i, T_j)}{\sqrt{N(T_i) * N(T_j)}},$$

where $N(T_i)$ and $N(T_j)$ is the number of occurrences of tags T_i and T_j across all resources and $N(T_i, \, T_j)$ is the cooccurrence frequency of tags T_i and T_j across all resources.

The cooccurrence index, CI, represents the degree of usage similarity based on the cooccurrence and frequency of the two tags. This, in fact, can be interpreted as the degree of usage similarity between the concepts represented by the two terms. If two tags co-occur only once in the entire system (and the number of occurrences of each of those tags is 1), then the CI score for that pair would be 1. This would mean that the pair of tags are highly predictive of each other's occurrences, and in this case, they together uniquely identify a resource. Thus a higher value for the cooccurrence index shows greater usage similarity to each other.

For our analysis, we computed the cooccurrence frequencies for all pairs of tags and computed the CI for each of them in the tag networks. The CI would provide an experimental value for the degree of relationships between two tags in the tag network in terms of their usage patterns. In other words, this measure quantifies the degree of usage similarity between two tags that are created by the participants.⁶

2.4.3 Latent Semantic Analysis (LSA). LSA is one of the statistical techniques for extracting and representing the similarity of meaning of words and

⁶We have tried other similar measures and the results were similar.

passages by analysis of large bodies of text [Landauer et al. 1997]. It uses singular value decomposition, a general form of factor analysis, to condense a very large matrix of word-by-context data into a much smaller dimensional representation. The similarity between resulting vectors for words and contexts, as measured by the cosine of their contained angle (thus the value goes from -1 to 1), has been shown to closely mimic human judgments of meaning similarity and human performance based on such similarity in a variety of ways. LSA, as well as variations of similar statistical language techniques such as information scent [Pirolli and Card 1999; Fu and Pirolli 2007], had been successfully applied to explain how users interpret the relevance of link text on web pages [Blackmon et al. 2005]. We chose LSA (instead of, for example, Pointwise Mutual Information (PMI), which was shown to be a better measure in some metrics [Budiu et al. 2006]) not only because they were readily available on the Internet, and also because it was good enough for the current purpose of comparing the semantic relatedness of tags. In fact, previous studies have provided good validation for using LSA scores to reflect the general semantic relatedness of words [Landauer et al. 1997; Blackmon et al. 2005]. In the current analysis, we performed the LSA calculations through the Web site at http://lsa.colorado.edu, using the general reading topic space with 300 factors. We assumed that the LSA scores reflect the mental representation of semantics of words by the participants. Comparisons between the CI and LSA scores will therefore allow us to study the extent to which choice of tags in the system is associated with their semantic relatedness. We will elaborate on this method in the results section.

3. THE EXPERIMENT

3.1 CiteULike Social Tagging System

CiteULike (www.citeulike.com), a research literature sharing Web site with tagging and search features, was used as our research platform. CiteULike helps users add links to papers, books, and references from other digital libraries and optionally tag the available content for future reference and use. CiteULike was chosen mainly for its simplicity of use and the relative ease of creating libraries by direct import from external resources using BibTeX or RIS format. Users can add resources (e.g., links to paper or book) to CiteU-Like, create relevant tags that describe the resource, search for related resources using either the tag cloud (an occurrence frequency based list of tags) or search keywords and share resources with other users. In Figure 4, a user's library of resources is shown along with the tags associated with each resource. The article titles and tags are within the square box labeled as 4 (with tags appearing below each hyperlinked resource), while the search options (using tags or keywords) are shown within the box labeled as 3 and the tag cloud is labeled as 5.

Users can also use CiteULike to find resources. There are three possible search methods to explore and navigate through resources: browsing through available resource titles in a general library, using keywords for search queries,

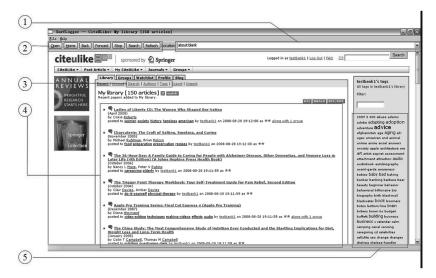


Fig. 4. CiteUlike with SurfLogger. 1: URL that is recorded, 2: List of function buttons (Open, Back, etc.), 3: Navigation tabs (e.g., Search), 4: Contents of the books showing titles and tags, and 5: Tag cloud showing all tags currently in the library.

selecting relevant tags from the tag cloud or from the tags associated with one or more of the available resources in the library. There was no tag recommendation or "suggestions" provided in the search boxes as participants searched. In short, users could use one or more of these options to navigate to their required resources. CiteULike supports personalization of resources using libraries and tags. All available resources in CiteULike are grouped into a general public library. A user can add a particular resource to his/her personal library for future reference and use. Since CiteULike supports all the basic features common to other social tagging systems, we believe that our results are generalizable to other similar systems.

User activities including mouse events, URLs, time stamp, and contents of each web page were recorded using a browser-plugin called Surflogger. The tool allows automatic saving of all mouse clicks and Web contents, which were further processed such that user behavior related to tagging, navigation and browsing was extracted. Figure 4 shows the CiteULike interface integrated with SurfLogger. Parts 1 and 2 were the URL bar and controls buttons provided by Surflogger, and the rest were from the original CiteULike interface.

3.2 Search Tasks

Our information resource consisted of 150 books that were imported directly from Books@Amazon.com. This "library" of books covered eight categories with approximately equal number of books in each category: Arts & Photography, Business & Investing, Children, Computers & Internet, Cooking, Food & Wine, Health, Mind & Body, Medical, and Self-Help. Each participant was instructed to search for books related with one of the 8 different topics. We assume that the

Table I.

The Exploratory Information Search Topics Given to the Participants (detailed descriptions of the topics can be found in Appendix 1)

Software Company	Books for technical staff (programmers and non-programmers) in a			
	software company.			
Local Arts Center	Books that would be appropriate for a library at the new art center.			
Traveler's Books	Books that a vacationer would enjoy reading by the pool or on the			
	beach, possibly promoting socialization in small groups.			
Career Center	Books for job seekers ranging from college graduates, recently			
	unemployed, and people looking to switch careers.			
Rehabilitation Center	Books that could be of use for patients, visitors and care-givers at			
	the rehabilitation center.			
Daycare Center	Books used on a daily basis by children as well as books that can be			
	checked out by their parents and caregivers.			
Retirement Community	Books for older adults interested in investing their money and			
	providing health tips for old age.			
Wellness Center	Books on nutrition, balanced diets; for chronically ill having dietary			
	restrictions, and young parents who want to learn new recipes.			

eight different topics will bias participants to pay attention to different aspects of the books (such as whether a book may be suitable for children or older adults, or whether the book has technical or artistic content), thus influencing the topic extraction process. Given that our focus is on investigating how information goals may influence choice of tags, the eight information search topics would allow us to understand whether an initial bias to the topic extraction process will lead to differences in tag choices, and how the initial bias may interact with the social influence of tags created by others.

Participants were asked to look for books relevant to the particular topic assigned to them. For example, if a participant was asked to select and recommend books that would be ideal for members of a rehabilitation center or a daycare center, they would need to search for books related with health care, medication, and other concepts related with this topic. Each topic search task was designed to provide a relatively unconstrained, yet specific information goal for participants, such that they have to perform exploratory search in the entire library of books.

Short descriptions of the eight topics are given in Table I. The complete descriptions of the eight tasks are provided in Appendix 1. Assigned with one of the eight search tasks, each participant was instructed to find relevant books from the common library with 150 books, save useful books to an assigned personal library account, and assign tags to each of the books they select. The tags assigned to each book are displayed below the book and the global list of all tags in the library will be displayed as a tag cloud (see area marked as 5 in Figure 4).

3.3 Participants

Sixty-four college-aged adults from the University of Illinois community participated in this study (average age =21.3 years, s.d. =3.57 years). All participants were skilled computer users with more than 10 years of experience in using

computers on average (average experience = 11.5 years, s.d. = 2.8 years). Each participant was paid \$8 for participation in the experiment. The experiment lasted for about an hour.

3.4 Experimental Design

The 64 participants were randomly divided into 2 groups: the social group and the nominal group (with 32 participants in each group). The main difference between the two conditions was the visibility of tags created by users in the earlier sessions for the users of future sessions (details explained later). Within each group, participants were further blocked into 4 sessions, with 8 participants in each session. Within each session, each participant was randomly assigned to one of the 8 different search tasks. For each session, each of the eight participants was given exactly one unique topic to search. The order of the search tasks within each session was randomized. Thus, in each group (social or nominal), each search topic would be assigned to exactly 4 participants (one in each session).

- 3.4.1 *Social Group.* The social group was designed to examine the social nature of the tagging process. In the social group, the tags created by a user are visible to the future users of the system. In other words, the tags created by the first participant in session 1 would be visible to every other participant in session 1 and future sessions.
- 3.4.2 *Nominal Group*. In the nominal group, the tags created by a participant are not visible to other participants in the same session or any other ensuing sessions. In other words, a participant added tags to a resource based on their own understanding about the resource. Thus, in the nominal condition, the participants worked as though they were tagging individually. This condition was designed to work as a control group to compare with the *social* group described earlier.

3.5 Procedure

An initial set of "neutral" tags (tags directly from the title/subtitle/abstract of the book) was added to a book prior to the start of the experiment. This initial set of tags (on average 3 tags) was added to avoid complete randomness in the tagging process (e.g., using sentence/phrase as tags, meaningless tags like "article 1," etc.), and to make the library resembles realistic ones. The same initial library of tags was used in both the social and nominal groups and thus their influence are equalized. After being debriefed about the tasks and the goal of the research, participants signed the consent form. Then they filled a survey on demographic information, experience with computers and social tagging systems. The researcher then described the search task and demonstrated how to perform information search and add tags on the CiteULike interface. Participants were allowed to familiarize themselves with the interface before experiment began and were encouraged to ask any questions regarding their task.

Participants could in general use three methods (or combinations of them) to explore for books that might fit their search task as mentioned above: Browsing, searching by keywords, or searching by tags. In all these methods, tags directly associated with the books would be visible (i.e., initial tags and tags created/revised by previous participants in social group; in the nominal group, only initial tags would be visible). If a book was considered appropriate for a search task, participants could create new tags or edit/reuse existing tags for the selected book, and then add the book to an assigned personal CiteULike library. At the end of the experiment, the page contents of the users' personal library and group library (including the selected books by each subject) and all the log files were saved for further data analysis.

4. RESULTS

In this section, we describe the results from analysis of tags created in the social and nominal groups. First, we provide the descriptive statistics of the tag assignments and properties of the networks generated by each of the groups, and how the network properties correlate with the semantic relatedness of tags. We will then report on the influence of information goals on tagging behavior. We will then present the interaction effects of information goals and social influence of tags, as predicted by the semantic imitation model.

4.1 Descriptive Statistics and Network Properties

Participants in both groups selected approximately the same number of books (social, M=22.39, s.d. = 7.1.0; nominal, M=24.60, s.d. = 4.38; t (31) = 1.49, p > 0.05). There were a total of 703 unique tags created in the social group and 852 unique tags in the nominal group. At the end of the experiment, the tags assigned to the library in the nominal group were merged to form a final library of tagged books that has the same number of books as in the library in the social group for direct comparison. Duplicate tags (tags that used the exact same word) were removed for both groups (e.g., if users X and Y assign tags (A, B) and (B, C) to the same book, the merged set of tags for that book will be (A,B,C)) within each session (but not across sessions). This is to ensure a better comparison between the two groups when we calculated the latent semantic scores. However, the number of duplicate tags constituted less than 5% of the total number of tags for both groups, and thus did not significantly influence the general patterns of results.

The number of tags in the merged library from the nominal group was significantly higher than that in the social group (social, M=6.66, s.d. = 3.75; nominal, M=10.08, s.d. = 4.47; t (31) = 7.69, p < 0.01), suggesting the more diverse tag choices in the nominal group. However, the mean number of *unique tags* assigned to each book per participant was not significantly different (t (31) = 0.81, p > 0.1, see Figure 5). There was no significant difference in the number of unique tags per participant across sessions in the nominal group. However, the linear downward trend for number of unique tags across sessions for the social group was significant (F (1, 30) = 3.92, p < 0.05), confirming that as

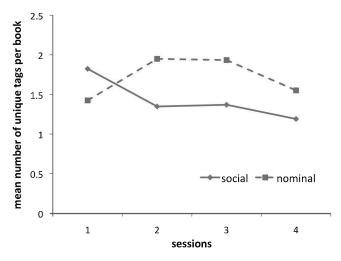
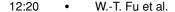


Fig. 5. Mean number of unique tags assigned to each book by participants in the social and nominal group.

more tags were added to the library, the number of unique tags decreased. No other effect was significant. We also did not find any significant event of search topics on any of the variables (i.e., all search topics seemed to induce the same pattern of behavior).

The pattern of results provided at least partial support for the social influence of tags. Consistent with previous research on small group performance [Steiner 1972; Hill 1982], the number of unique tags contributed to the library in the social group was smaller than that in the nominal group. Another way to interpret the result is that participants slowly converged on tag assignment in the social group but not in the nominal group, as shown by the decreasing number of unique tags across sessions in the social but not the nominal group. Apparently, when participants could not see tags created by other participants, the general choice of words to describe the books did not seem to converge at least in the 4 sessions that we observed, but we did observe a clear trend of convergence in the social group. We believe this is an intriguing result. Although researchers [Golder et al. 2006] have showed the general trend of tag convergence over time in a large social tagging network, to our knowledge, no empirical evidence has yet been collected to directly demonstrate the social influence of tags by finding the difference between the social and nominal group as in the current study.

To further test the predictions of the semantic imitation model discussed earlier, additional analyses on this convergence process based on the semantic relatedness of these tags were performed. We will first present results comparing the general network properties of the systems from the two experimental groups and how they correlate with the semantic relatedness of tags, before we present results on the effect of information goals and how it interacted with the social influence of tags in the social group.



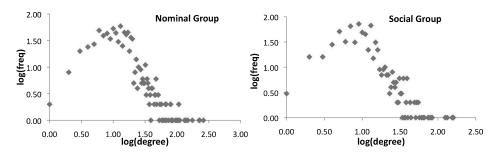


Fig. 6. The degree distributions in the social and nominal groups. The patterns exhibited scale-free power-law distribution as found in small-world networks.

Table II. Network Properties of the Social and Nominal Group Tag Networks

	Nominal	Social
Average node degree	18.9	14.1
Clustering Coefficient (C)	0.76	0.86

4.2 General Network Characteristics

Two separate networks were created for tag libraries created by participants in the social and nominal groups. Both tag networks from the social and nominal groups exhibited properties similar to that of known large-scale social tagging and other popular networks (e.g., del.icio.us, coauthorship networks, WWW). Both networks exhibited "small-world" properties, in which most nodes can be reached from every other node by a small number of transitions (average path length < 3 for both networks). Another typical characteristic of a small-world network is that there are a large number of highly connected nodes (acting as hubs), exhibiting a fat-tailed distribution in the node degree distribution that can be characterized by a power function. Indeed, the node degree distributions of the two networks were very similar, and exhibited the scale-free power-law distributions (see Figure 6).

These highly connected nodes were likely common (semantically general) words that were used to tag multiple books by both groups, which we will further investigate in the analyses presented later. Nevertheless, the scale-free property has at least partially validated our design: data generated from our experiment seemed representative of the large-scale social tagging networks (e.g., del.icio.us) analyzed by other researchers, even though we had relatively small number of participants and considerably fewer number of tags in the networks (Social =703 tags, Nominal = 852 tags).

Although the node degree distributions were similar, other basic network measures did show some differences between the two groups (Table II). The nominal group had a higher average node degree (the average number of nodes connected to any particular node) but had a lower clustering coefficient, indicating that the network generated by the nominal group was more distributed and less connected than the social group. Consistent with the pattern shown

Table III.

 $\label{eq:co-occurrence} Co-occurrence\ Measures\ for\ the\ Tag\ "Health"\ (Abridged\ List)\ for\ the\ Social\ and\ the\ Nominal\ Group.\ The\ Number\ of\ Tags,\ Distance\ Measure\ and\ the\ LSA\ Score\ for\ a\ Snapshot\ of\ Tags\ Cooccurring\ with\ the\ Tag\ "Health"\ (Note\ that\ for\ the\ tag\ "health,"\ N\ (T_i)(social)=22\ and\ N\ (T_i)(nominal)=14)$

		Social Group				Nominal Group			
T_i	T_j	$N(T_i, T_j)$	$N(T_j)$	CI Score	LSA Score	$N(T_i, T_j)$	$N(T_j)$	CI Score	LSA Score
health	nutrition	6	9	0.43	0.51	8	13	0.59	0.51
health	diet	5	6	0.44	0.34	11	13	0.82	0.34
health	exercise	5	6	0.44	0.24	3	14	0.27	0.24
health	self help	3	9	0.21	0.19	2	15	0.14	0.19
health	age	1	8	0.12	0.07	4	10	0.04	0.07
health	advice	2	22	0.09	0.26	6	28	0.30	0.26
health	cook	2	4	0.21	0.01	1	4	0.13	0.01

in Figure 5, the higher node degree in the nominal group suggested that when participants could not see tags created by others, there was higher variability in tag choices than when participants could see tags created by others. Similarly, the higher clustering coefficient in the social group suggested that there were more clusters of tags in the network, indicating that tags in the social group had a higher tendency to converge than those in the nominal group.

To investigate the semantic relatedness among tags, we computed the CI (co-occurrence index) and LSA scores for all pairs of cooccurring tags in the social and nominal groups. For the sake of simplicity and explanatory purposes, we used one of the most commonly used tags as an example to report our results from this analysis. The tag "health" had 22 occurrences across all the resources in the social group and 14 occurrences in the nominal group. In other words, there were 22 (14 in the nominal group) unique occurrences of the tag "health" in the resources (books) in the social group. In the social group, the tag "health" was linked to 154 other tags, while in the nominal group, the tag "health" was linked to 132 other tags. The CI score and the LSA score was computed for all cooccurring pairs of tags (with health). A snapshot of a few tags cooccurring with "health" is shown in Table III, together with the CI and LSA scores.

As shown in Table III, some of the co-occurring tags (in both the social and nominal group) had fairly similar CI and LSA scores (e.g., health-nutrition had CI and LSA scores of 0.43 and 0.51 respectively for the social group), while others were quite different (e.g., health-cook had CI and LSA scores of 0.21 and 0.01 respectively for the social group). However, in general we see a reasonably high correlation between the two scores. Indeed, the correlation between the CI and LSA scores for "health" and its cooccurring tags was 0.67 and 0.38 in the social and nominal group respectively.

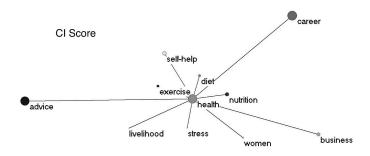
We obtained similar correlation values for the top 20 cooccurring tags. These top 20 tags (out of 703 and 850 total number of tags in the social and nominal groups, respectively) were highly connected nodes in the network and represented more than 50% of all links in both the networks. Table IV shows the list of correlations of the two networks of the top 20 tags. The mean of the

Table IV.
Correlation of the Top 20 Tags in the Social and Nominal
Group; the Average Correlation for the Social Group was 0.65
and 0.35 for the Nominal Group

Tag	Social Group	Nominal Group
How-to	0.76	0.55
Advice	0.69	0.39
Guide	0.58	0.45
Parent	0.55	0.28
Self-help	0.76	0.22
Success	0.53	0.32
Food	0.61	0.28
Health	0.67	0.37
Children	0.54	0.32
Tutorial	0.45	0.39
Recipe	0.77	0.41
Design	0.68	0.61
Art	0.65	0.22
Life	0.69	0.21
Manage	0.77	0.51
Tip	0.69	0.26
Nutrition	0.70	0.55
Cook-book	0.72	0.25
Adult	0.63	0.21
Love	0.52	0.24

correlations from the social group was significantly higher than that from the nominal group (social group: mean = 0.65; nominal group; mean = 0.35, t(19) = 11.8, p<0.01). The results show that cooccurring tags in the network created by the social group had significantly higher semantic relatedness than those in the nominal group. In other words, tags that were assigned to the same books tended to have a higher semantic relatedness in the social group than those in the nominal group. The results were therefore consistent with the first prediction of the semantic imitation model: When participants could see tags created by others, tags would be more semantically similar than when they could not see tags created by others, confirming the social influence of tags at the semantic level in the social group.

One point that is worth mentioning is that CI scores were computed from the experimental data (i.e., based on the tags created by the users), and they provided a metric for measuring the likelihood of cooccurrence of tags in the network. On the other hand, the LSA score is a theoretical measure that is believed to reflect the semantic relatedness between two words calculated based on their long-term cooccurrence statistics in a large text corpus. Our assumption was that a higher correlation between the CI and LSA scores in the social group would imply that tags that were more likely to cooccur in the network created by the participants were also semantically similar in general. In contrast, the lower correlation in the nominal group implied that tags that tended to cooccur in the network were less semantically similar. This pattern of results were again consistent with the prediction of the semantic imitation model: The



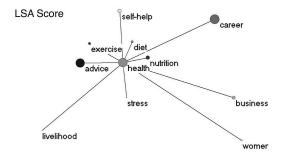


Fig. 7. Graphical representation of a set of cooccurring tags for "health" in the social group condition. The figure on the top shows the representation using the distance measures (CI) score and the bottom figure shows the representation using the LSA score.

social influence of existing tags leads to the emergent patterns that tags that are closer to each other in the network are also semantically similar. However, we do not see the same pattern in the nominal group, suggesting that when individual choose tags based solely on their own information goals and background knowledge, tags created tended to be more diverse and do not naturally become semantically similar, creating a situation similar to the "vocabulary" problem predicted by others [Furnas et al. 1987]. In other words, our results suggested that when participants were allowed to indirectly communicate through the interface (in this case, by introducing the simple feature that makes tags created by other users visible to all), this vocabulary problem apparently was naturally mitigated by the general tendency for users to "semantically imitate" each others.

To better illustrate the relations between the CI measure and the LSA measure, we created graphical representations for a selected set of co-occurring tags for the tag "health" in the social group (see Figure 7). The top panel of Figure 7 shows the graphical representation based on the cooccurrence index (CI). The central tag was "health". The tags that were closer to health (e.g., nutrition, diet, advice) had higher CI score than tags that were further away (e.g., women, livelihood). The same graph was also generated using the LSA scores (bottom panel in Figure 7). In both figures, the sizes of the nodes were proportional

to the frequencies of the tags. The figures show that although the structures of the connections were in general similar, there were apparent differences between the two measures. For example, the tags "Advice" and "Health" had a relatively high likelihood of cooccurrence (top panel of Figure 7), however, their LSA score was relatively low (bottom panel of Figure 7), indicating that their cooccurrences were likely not attributable to their general semantic relatedness. Similar representations can be used in large networks to evaluate and develop ontologies of tags based on cooccurrence and theoretical semantic similarity. Indeed, in this particular example, we found that the high cooccurrence in the network had much to do with the information goals given to the participants. We will present analysis on the effects of information goals next.

4.3 Effects of Information Goals on Tag Choices

To study the effect of information goals on creation of new tags, we extracted all book selections and analyzed when and how tags were created and reused. Although participants were told that they could assign a tag to any book in the library during experiment, we found that almost all tags were assigned only to books selected by the participants (i.e., only those books that were selected and added to their personal library). Therefore, we only focused on books selected by the participants and extracted all tags that were assigned to them. A standard stemming algorithm was then performed on the tags.

In the social group, a tag was coded as "new" if the tag did not overlap with any of the existing tags associated with that particular book; otherwise it was coded as "reused." In the nominal group, tags were first sequentially merged as if they were in the social group before the coding, such that tags assigned by participant 2 were compared to the tag library after participant 1 finished the task (which included the initial tag library plus the tag assignment by participant 1), and tags assigned by participant 3 were compared to the tag library formed by merging the tag libraries from participants 1 and 2, and so on. Similar to the social group, each tag assigned to a book was coded as "new" or "reused" according to whether the tag overlapped with existing tags of the same book in the sequentially merged tag library. In other words, tags in the nominal group were merged as if they were collaboratively created in the social group. The new and reused tags coded in each group were then used for further analysis and evaluation.

To understand how different information goals influenced tag choices, we listed the search topics assigned to the participants for each book selected. The topics were then rank ordered according to the actual order the search topics were assigned in the experiment for both groups. The topics for each book selection were then coded according to whether the task appeared earlier in the list or not. If the task appeared earlier in the list (indicating that same book had been selected before under the same search topic), it was coded as "same"; otherwise, it was coded as "different." For example, if book A was selected by participants 1, 5, 9, and 12, who were given search topic X, Y, X, and W respectively, then the codes for the four book selections would be "different," "same," and "different" respectively.

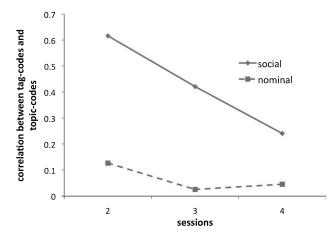


Fig. 8. Correlation between tag-codes and topic-codes. A high correlation indicates that most new tags were created under different exploratory search topics.

We created the corresponding tag-codes (new or reused) and topic-codes (different or same) as discussed earlier for each group, and then calculated the correlation between the tag-codes and topic-codes by converting "new" and "different" to 1, and "reused" and "same" to 0. A high correlation would therefore indicate that new tags are associated with a different search topic (new tags are created in different search topic) or reuse of tags occurs when the search topic is the same. Figure 8 shows the correlation between the tag-codes and topic-codes in the social and nominal group across the four sessions. Note that because in the first session all participants were given new search topics, the correlation could not be calculated (there was no variance). All correlations shown in Figure 8 were significant (p < 0.05), as was the obvious difference between the social and nominal groups (t (2) = 5.36, p < 0.05). There was no significant difference across sessions in the nominal group, but the linear downward trend of the correlations across sessions in the social group was significant (F (1, 22) = 4.21, p < 0.05), indicating that the correlation between new tags and new topics tended to decreases across sessions for the social group.

The pattern of results shown in Figure 8 in general supports the second prediction of the semantic imitation model: Tag choices were dependent on the interaction between the social influence of tags and the information goals of the user. In the social group, the creation of new tags was strongly associated with differences in information goals (or reuse of existing tags was strongly associated with whether the information goals were the same). However, this association was much stronger early on than in later sessions when the number of tags increased. In other words, different information goals led to more creation of new tags, and same information goals led to more reuse of existing tags in the early sessions, but this effect dissipated over time.

In the semantic imitation model, it is assumed that information goals moderate the social influence of existing tags on the topic extraction process, which eventually influence the tag choice process. Thus, the influence of information

goals on tag choices is assumed to be stronger when there are fewer existing tags (thus less bias from the tag-based topic inference process), such that the topic extraction process will be relatively more influenced by the information goals, which eventually influence their choice of tags to describe the topics extracted from the document. But as more tags are added, the tag-based topic inference process will become stronger, making the influence from information goals relatively weaker.

The current results showed that in the nominal group, without the influence from the semantic interpretation of social tags, creation of new tags showed much weaker association with differences in information goals. In other words, new tags were just as likely created under the same or different information goal (hence, the low correlation values). This was possibly due to the fact that participants in the nominal group could not see tags created by others, and therefore even when participants had the same information goal, their choice of tags tended to be more variable than participants in the social group, whose choice of tags were semantically primed by the social tags. The results were also consistent with the model's prediction that as the number of tags increased across sessions, the influence by the semantic representations of existing tags on tag choices became stronger, such that reuse of existing tags became less associated with whether the participants had the same information goal as previous participants or not. This explanation was supported by the decrease in the correlation in the social group across sessions. To further verify this possibility, we will present further analysis on the semantic relatedness of tags in each group next.

4.4 Semantic Relatedness of Tags

We calculated the LSA scores between every set of new tags created and the existing tags in each book selection, broken down by whether the books were selected (and tags assigned) under the same or different search topics as previous participants. This measure would inform the extent to which the creation of new tags was influenced by the presence of existing tags assigned to the same book, and whether the new tags created were semantically similar to the existing set of tags. In addition, by comparing the same measure for new tags created under the same and different information goals, we could further test the moderating effect of information goals on tag choices as predicted by the model, and how this effect would changes across time.

Figure 9 shows the mean LSA scores between the new and existing tags averaged across participants in each session, broken down by whether the search task was the same or different (note that for tag reuse, the LSA score would be 1.0). The main effect of groups was significant (F (1, 31) = 3.41, p < 0.05), confirming the obviously higher scores in the social than the nominal group. In the social group, the main effect of same/different information goals was not significant, but the interaction between same/different information goals and sessions was significant (F (2, 31) = 3.41, p < 0.05). There was no significant difference between sessions or between same/different information goals in the nominal group.

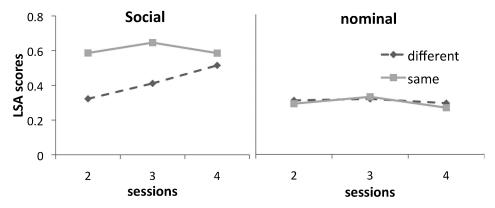


Fig. 9. LSA scores for new tags created across sessions in the social and nominal groups, broken down by whether the books were selected under the same or different search tasks.

In the social group, the LSA scores for tags created under the same information goal were not significantly different across sessions, but the LSA scores for tags created under different information goals in sessions 3 and 4 were significantly higher than that in sessions 2 and 3 respectively. The LSA scores for the same information goal were significantly lower than those for the different tasks in sessions 2 and 3 (p < 0.05), but there was no significant difference between the scores in session 4. In other words, in the social group, the LSA scores stayed approximately at the same level for tags created under the same information goal across sessions, but the LSA scores increased significantly across sessions for tags created under different information goals. In addition, new tags created under different information goals were semantically farther apart than those created under the same information goal in early sessions; but as more tags were created, new tags created under different information goals were at the same level of semantic relatedness as those created under the same information goal.

The results shown in Figure 9, together with the results shown in Figure 8, provided further support for the semantic imitation model. Specifically, the results were consistent with the notion that semantic interpretation of existing tags constrained tag choices by semantic priming of the topics. The results also supported the prediction that semantic priming mediated the effect of information goals given to the participants: In the nominal group, there was no difference between the two groups. On the other hand, in the social group, as the number of tags increased, semantic priming from existing tags increased its influence on tag choices over information goals, confirming the interaction effects between information goals and social influence of tags in tag choices.

4.5 Summary of Results

The sets of results we have presented provided support to the two predictions of the semantic imitation model. Consistent with previous results, we found that participants in the social group converged on the choice of tags, demonstrating the social influence of tags. However, the most important finding from the current study was that this convergence seemed to be strongly associated with the semantics of tags. Specifically, compared to the nominal group, we found that in the social group (1) tags that cooccurred more frequently in the tag networks tended to have high semantic relatedness, (2) the semantic relatedness of tags were stronger, and (3) the semantic relatedness of tags tended to increase even when participants were tagging under different information goals. These set of findings provided strong support for the main assumption of the semantic imitation model: Semantic representation and interpretation plays a critical role in the social influence of tag choices.

The results from the experiment provided support for the second prediction of the semantic imitation model: there were significant interaction effects between the social influence of tags and information goals. Specifically, we found that in the social group, the decision to create new tags or to reuse existing tags was strongly associated with whether participants had the same or different information goals, but we did not find this association in the nominal group. In addition, we found that tags created under the same information goals had significantly much higher semantic similarity in the social group than those in the nominal group throughout the sessions. However, tags created under different information goals had the same low level of semantic similarity in both the social and nominal group in early sessions, but the level of semantic similarity increased significantly across sessions only in the social group, but not in the nominal group. This pattern of results provided further support for the hypothesis that semantic priming of tags constrained the tag choice process by increasing the likelihood of using semantically related tags, and the effect of information goals on topic extraction and tag choices was mediated by the semantic interpretation of social tags.

5. GENERAL DISCUSSION

We presented a semantic imitation model of social tagging, which provided concrete predictions that guided the design of our experiment. Specifically, the model assumes that when a user interprets existing tags associated with a document, a tag-based topic inference process will be invoked, which biases the user to pay attention to certain topics semantically represented by the existing tags. This bias in attention will later influence the topic extraction process, as the user comprehends the document, and subsequently influence the tags assigned to the same document. In other words, as a user interprets a tag, information passes from the word level to the semantic level (meaning of the tag) and influences the comprehension of the document, and subsequently influences the tag choices back at the word level. Based on this assumption, the model predicts that tags created by multiple users will be more similar at the semantic level (i.e., they tend to have similar meanings), even though they may or may not be more similar at the word level (i.e., they may not be the exact same words). In addition, the model assumes that information goals play a moderating role in the topic extraction process. Specifically, the model predicts that the effect of semantic relatedness of tags will be magnified by the effect of information goals: Semantic relatedness of tags created by different users tend to increase even further when the information goals of the users are the same, but the semantic relatedness of tags will decrease when the information goals are different.

The results from our experiment confirmed these predictions derived from the semantic imitation model. We found that tags generated by the social group tended to converge over time, but those in the nominal group did not. Based on the measures of the tag cooccurrence index and the LSA scores, we also found that in the social group, tags cooccurred in the network were also semantically similar (and those that did not cooccur tend to be less semantically similar), providing support for the model's prediction about the social influence of tag choices at the semantic level. This pattern was not found in the nominal group. Finally, we found that tag choices were moderated by the information goals. In the social group, participants who had the same information goals tended to create tags that were more semantically similar than when they had different information goals. However, as the number of tags associated with the resources increased, the moderating effect of information goals tended to diminish, such that tags became more semantically similar even when participants had different information goals. The results suggested that the social influence of tags became stronger as the number of tags increased, and eventually outweighed the effect of information goals. In the nominal group, tags were much less semantically similar than those in the social group, regardless of whether the information goals were the same or not. This result suggested that the visibility of social tags was a precondition for the effect of information goals. In other words, only when participants could see other people's tags, they would interpret and assign tags that were semantically similar to others who had similar information goals.

We believe that our results are significant in several ways. First, although previous studies have demonstrated the social influence of existing tags, our study was the first that directly demonstrate not only the existence of social influence, but also the *nature* of this social influence. Specifically, our results provide strong support for our assumption that the social influence appears to occur at the semantic level, not at the word level as implicitly assumed by previous studies. Second, our results demonstrated the value of the humaninformation processing approach from cognitive sciences: It provides detailed process data that cannot be easily obtained from statistical analyses of a snapshot of an existing system. Indeed, we showed that by carefully controlling the presentation of tags and the information goals, we could tease apart the relative contribution of multiple processes that contribute to the tag choice processes. Third, to the best of our knowledge, our study was the first that directly manipulated the information goals given to the users as they performed the exploratory search and assign tags, and not only that we found significant moderating effect of information goals on tag choices, but also how this effect changed across time. Our results therefore complemented previous research that estimated the information goals of the users based on self-reports. Fourth, our study was also the first that directly compares tagging behavior between a social and a nominal group, and the difference provided much stronger scientific support for the social influence than data collected from self-reports [Sen et al. 2006]. Finally, we provided a novel analysis approach that combines social network analysis at the aggregate level with information processing analysis at the individual level, which allows us to draw conclusions that span across multiple levels of analysis as predicted by our semantic imitation model.

5.1 The Nature of the Social Influence of Tags

The semantic imitation model assumes that semantic representations play a pivotal role in social tagging behavior, and our experiment was designed to investigate the role of semantic representations in tag choices. Results from the experiment provided strong support to our assumption that semantic representations played a pivotal role in tag choices, as confirmed by the wide set of analyses we conducted. These results also provided important information about the nature of the social influence of tags: Social influence of tags occurred at the semantic level. We believe this finding is a significant step forward in the understanding of social tagging behavior, as most analyses aimed at revealing patterns at the word level, which we believe had inflated the variances in tag choices. Thus, the conclusion that tags tend to have lower information value in predicting documents across time is premature. In fact, when the semantic interpretation process is taken into consideration, people could be just as good at predicting resource contents even when the tags assigned to the resources become more diverse, so long as the latent semantic structures behind the diverse set of tags remain stable during exploratory search. This is similar to the case in natural communication and knowledge growth: Even though the vocabulary is continuously growing, our ability to communicate do not necessarily diminish, so long as the shared latent semantic structures behind the vocabulary remain stable and consistent within a cultural groups or domain.

We believe that our model is more general than many existing models developed at the word level [Golder et al. 2006; Cattuto et al. 2007], as one can consider reuse of the exact tag as an extreme form of semantic imitation. What we showed was that in cases when the exact tags were not reused, there was a graded influence of semantic priming from existing tags on the choice of words, demonstrating the social influence of tags. However, we believe that the current model provides a more useful mechanism to explain the growing stability in tag proportions, as the analysis will take into account not only the statistical patterns of words (tags) in the system, but also the mapping between distribution of words and meaning in human knowledge [Anderson 1974; Kintsch 1998; Griffiths et al. 2007], and the processes that allow humans to interpret words based on their background knowledge, as well as to utilize their background knowledge to generate tags or keywords for search. In other words, the current model includes the human knowledge system as part of the dynamic system that word-level models fail to do. Including humans as part of the system can have significant implications on predicting performance.

While word-level models will predict that tagging behavior is purely driven by tags in the systems, the semantic imitation model predicts that if two persons have different background knowledge structures, their interpretation of tags will be different, and their choice of tags or keywords to search will also be different. Indeed, results from a recent study have confirmed this finding [Kang et al. 2010]. Specifically, Kang et al. found that experts could perform exploratory search better because they are better at interpreting tags as well as generating tags and keywords to search than novices. In general, we believe that including the human knowledge representations in the model can greatly enhance the model's predictive power on human performance, especially in applications in which human knowledge representations play an important role, such as those in education, scientific discovery, and other domain-specific knowledge sharing systems.

5.2 Implications of the Semantic Imitation Process on Social Tagging Systems

Many argued that one of the great potential of social tagging systems is that social tags can act as navigational cues that facilitate exploratory search of information [Marchionini 2006; Qu and Furnas 2008; Fu and Dong 2010a, 2010b; Fu et al. 2010; Kang et al. 2010]. One important distinction between traditional fact retrievals and exploratory search is that in exploratory search, the user may not have a clear target page in mind, but only wants to look for information related to a general topic that he or she is interested in. For example, one may be interested in finding some information about good restaurants around a particular area (an exploratory search), as opposed to finding the phone number and address of a particular restaurant in order to make a reservation (a fact retrieval). Therefore, in exploratory search, an important criterion is to allow users to explore many documents related to a particular topic, rather than to retrieve a single document based on an information cue (e.g., by entering a search term into a search engine). Users can then navigate and read the topically related documents to make sense of the information, and to refine their search goals for the next round of exploratory search until they have obtain enough information relevant to the topic that they are interested in.

The semantic imitation model and results from our experiment suggest that social tagging systems can indeed facilitate exploratory search. Consistent with prior research, we found that the number of unique tags created by users for a resource decreased with time in the social group. In other words, resources were progressively less likely to be tagged with new tags over time. Additionally, we also found that over time the semantic relatedness of tags tended to increase irrespective of whether the tag was created under the same or different information goals. Based on the semantic imitation model, semantically related tags can help the topic inference process. In the model, it is assumed that part of the reason why tags become semantically related is that users shared similar semantic structures that are socially constructed in daily lives through natural communications [Clark and Brennan 1991]. Topics inferred from the tags are therefore similar in the shared semantic space, and thus, tags created based on these topics will likely be semantically similar. The shared semantic contents thus act as the latent structures that induce convergence of tags in the semantic space.

In fact, in the psychological literature, researchers have established that redundant cues actually help, rather than hurt, retrieval of relevant concepts in long-term memory [Anderson et al. 2004]. Activation of relevant concepts is indeed a critical component in the process of comprehension, or sensemaking of information [Russell et al. 1993; Qu et al. 2008]. In other words, contrary to expectations, although the number of tags associated with any document tend to increase over time, the inherent social influence of tag choices may lead users to create semantically related tags that help people to infer the semantic contents of the resources, thereby facilitating exploratory search of information. Although the above conclusion may seem counter-intuitive from an information-theory perspective, if one assumes that the goal of exploratory search is to search for semantic contents, rather than direct retrieval of information content defined by traditional information theory, higher number of semantically related tags should increase the likelihood of topic inference, thus facilitating exploratory search. While this is subject to more rigorous testing in future research, the current results did show the added value of incorporating human semantic representations in the analysis of social tagging behavior on top of traditional information theories.

5.3 Implications of the Model on Higher Level Learning

One major reason why social tagging becomes popular is that people are becoming less satisfied with the Internet being used as a large information database, from which users can retrieve facts easily through powerful search engines. Instead, people are increasingly relying on the Internet to explore and comprehend information, and to share experiences and socialize among other users. The assumption of the semantic imitation model is that as people interact with the social tagging system, they tend to process the meaning of text (i.e., at the semantic level) rather than the text itself. This implies that the interaction is much broader than merely seeking relevant information, as it potentially also involves active interpretation of the information, as well as integration of external information with their internal knowledge structures. In other words, social tagging systems may also have the potential to encourage higher level learning that allows users to acquire general knowledge about a topic by studying the context of information (e.g., cues, other relevant documents, etc) during their interaction.

Researchers has shown that one way our knowledge grows is by developing and refining our internal representations (sometimes called schemas, see Rumelhart and Ortony [1976]) of the external world. As people interact with their environment and acquire more experiences their schemas may be modified to make sense, or used to make sense of the new experiences. Consider the case when a user is browsing for information related to a broad topic of interest, such as when one is interested in knowing more about facts or events related to the independence of Kosovo. During the search, social tags created by others can be utilized as navigational cues to select and navigate to the documents pertaining to the topic of interest, and information from the documents can lead them to explore for other relevant information to expand their understanding

of the topic. Through this process of exploratory search-and-learn, the user gains a better understanding of a set of related topics through the enrichment of internal representations of concepts (and their relations). In other words, through the iterative exploratory search-and-learn cycles, not only that the tags created in the system may become semantically similar, but the internal concepts of users may also be influenced by how others are tagging the same document as they interpret how others interpret the same set of information. Thus, the interactions between internal concepts and external tags gradually lead to sharing and assimilation of conceptual structures as more and more people assign social tags to represent ideas or concepts that they extract from the massive amount of Web documents. Indeed, a preliminary study shows that after exploring for information for a period of eight weeks, students do seem to have acquired the external knowledge structures from the social tagging system and assimilated those information into their own internal knowledge [Fu 2008]. A fruitful future research direction is therefore on understanding how social information systems in general can serve the general purpose of education, assisting knowledge exchange among people with different knowledge backgrounds, or facilitating multi-disciplinary collaboration.

5.4 Effects of Information Goals on Social Tagging

Results from our experiment show that information goals moderate the social influence of tags, such that participants with the same information goals tended to assign tags that were more semantically similar than those who had different information goals. Our manipulation of information goals was, however, relatively limited compared to the possible broad range of goals that people may have when they use a social tagging system [Sen et al. 2006]. In fact, our information goals were all task-specific, differing only on the topics that participants need to search for. More research is therefore needed to understand how semantic imitation plays a role (and how it interacts with other factors) when users are not merely searching for information, but using the tagging systems for other reasons. On the other hand, we believe that the semantic imitation model is developed at a fairly general assumption about semantic representations. Therefore, at least in applications in which exploratory search is concerned, we believe the model at least will provide some general guidelines on how semantic representations will influence tagging behavior.

5.5 Relations between Social Tagging and Human Language

Lastly, the current results may be relevant to recent research relating social tagging to human language networks. For example, Cancho and Sole [2003] found that human language exhibits properties of small world networks; i.e., two words can be reached in very few network steps. They also found frequent words are more likely to be more in sentences. Cancho and Sole [2001] describe language as consisting of word-object associations that are either referential (e.g., a camera, referring to something that is used to take photographs), non-referential (e.g., sleep as an act of sleeping) or both. There is a great deal

of similarity between language networks and tagging networks: (a) both impose cognitive constraints on comprehension and conceptualization; (b) both have a word-object representation which could be referential, non-referential or both and (c) both exhibit similar scale free and small world network properties.

An interesting parallel between word networks and tagging networks is that one can use a basic utility model to describe both. Cancho and Sole represent the utility model for language for a speaker and listener: speakers use semantically general terms (which are ambiguous) during a conversation, while listeners want clarity and explicitness. For speakers using general terms would mean selecting the some frequently occurring words, which is consistent with the principle of least effort [Zipf 1949]. However, the choice of semantically general words reduces the ability of the listeners to infer the true semantic contents of the message. Listeners are thus forced to semantically interpret the conversation by "filling in" using their own knowledge structures during the conversation. The exchange of messages between the speaker and listener eventually allows the listener to comprehend the "true" semantic contents of what the speaker is trying to convey in the first place.

The same comparison can be applied to tagging. Some taggers add tags of a general nature based on the principle of least collaborative effort. Other taggers of that resource may face with the task of interpreting the existing tags, and to add their own tags based on their understanding of the tags and personal knowledge about the resource. The iterative process of tag assignments is therefore similar to conversation, in which the tags are converging on the "true" latent semantic contents of the resource. Our semantic imitation model implies that the shared underlying semantic representation of multiple users (based on general conformity of knowledge structures through regular social exchange) is critical for multiple users to iteratively converge to the semantic contents of the resource. Future research would include how differences in the shared semantic representation may lead to differences in convergence in tagging behavior, such as when people who have different domain expertise may have different semantic interpretation of tags, and they may also create tags that are different from those generated by novices. In that case, novices may have trouble correctly interpreting tags created by experts (and vice versa), and the tags may not be as useful to different user population. On the other hand, if we understand better the relations between internal knowledge structures and tag choices, one may be able to design algorithms that identify clusters of "expert taggers" of different domains, such that people who are interested in those topics will be able to follow these experts more easily with some interface guidance.

APPENDIX

APPENDIX 1: DESCRIPTION OF INFORMATION TASKS

1. Software Company. A mid-sized software company is looking to update their library to be used by all its employees, which include technical staff such as programmers and engineers, as well as nontechnical staff such as

- accountants, managerial staff, and employees in the marketing, purchasing and personnel departments. Please recommend a well-balanced set of books that you think would be useful and helpful in this library.
- 2. Local Arts Center. A local community is trying to promote art, design, and architecture and is building a new art center. They are looking for books that would be appropriate and interesting in the library of the new art center. Please select a well-balanced set of books that you would recommend the art center add to its library.
- 3. Career Center. A career center is creating a library for its patrons and is requesting your help in picking out a set of books. The career center sees a variety of people, from college graduates to the recently unemployed to people looking to switch careers. Please select a list of books that you would recommend the career center add to its library.
- 4. Rehabilitation Center. A rehabilitation center is planning on updating its library and is searching for books that could be of use for their patients, visitors and caregivers. Please recommend a well-balanced set of books that you think would be useful and helpful in this library.
- 5. Daycare Center. A daycare center is looking to update their library that will be used on a daily basis by children as well as have books that can be checked out by their parents and caregivers. Please recommend a well-balanced set of books that you think will be useful and helpful in this library.
- 6. Retirement Community. A retirement community in Silicon Valley, California, is building a library for its members. The members are particularly interested in how to invest their money and also about ways to stay healthy in old age. Please select a well-balanced set of books that you would recommend to the retirement community.
- 7. Wellness Center. A wellness center is searching for books to add to its library. The center has a wide range of customers including people who want to learn about nutrition and have a balanced diet, people who are chronically diseased and may have dietary restrictions, and young parents who want to learn about new recipes for their family. Please select a well-balanced set of books that you believe would be useful and helpful for the wellness center.
- 9. Traveler's Books. A cruise line is starting a new reading program aboard one of its ships and is asking for your help in deciding what books to add to their library. The company has requested books that it believes vacationers would enjoy reading by the pool or on the beach and perhaps promote socialization in small groups. Please recommend a well-balanced set of books that you feel would be useful for the cruise line.

REFERENCES

Albert, R. and Barabasi, A.-L. 2002. Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74, 47–97.

Ames, M. and Naaman, M. 2007. Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI'07).

ACM Transactions on Computer-Human Interaction, Vol. 17, No. 3, Article 12, Publication date; July 2010.

- Anderson, J. R. 1974. Verbatim and propositional representation of sentences in immediate and long-term memory. J. Verb. Learn. Verb. Behav. 13, 149–162.
- Anderson, J. R. 2002. Spanning seven orders of magnitude: a challenge for cognitive modeling. *Cognit. Sci. 26*, 1, 85–112.
- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., and Qin, Y. 2004. An integrated theory of the mind. *Psych. Rev.* 111, 1036–1060.
- Blackmon, M. H., Kitajima, M., and Polson, P. G. 2005. Tool for accurately predicting website navigation problems, nonproblems, problem severity, and effectiveness of repairs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'05)*.
- Budiu, R., Royer, C., and Pirolli, P. 2006. Modeling information scent: A comparison of LSA, PMI-IR, and GLSA similarity measures on common test and corpora. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'06)*.
- Cancho, R. F. and Sole, R. V. 2001. The small world of human language. *Proc. Royal Soc., London B* 268, 2261–2266.
- Cancho, R. F. and Sole, R. V. 2003. Least effort and the origins of scaling in human language. *Proc. Nat. Acad. Sci. 100*, 788–791.
- Cattuto, C., Loreto, V., and Pietronero, L. 2007. Semiotic dynamics and collaborative tagging. *Proc. Nat. Acad. Sci. 104*, 1461–1464.
- CLARK, H. H. AND BRENNAN, S. A. 1991. Grounding in communication. In Perspectives on Socially Shared Cognition, L. B. Resnick, 1. M. Levine and S. D. Teasley Eds., APA Books, Washington.
- Downey, D., Dumais, S., and Horvitz, E. 2008. Understanding the relationship between searchers' queries and information goals. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. ACM Press.
- Frederiksen, C. H. 1975. Representing logical and semantic structure of knowledge acquired from discourse. *Cognit. Psych.* 7, 317–458.
- Fu, W.- T. 2008. The microstructures of social tagging: A rational model. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. 229–238.
- Fu, W.-T. and Dong, W. 2010a. From collaborative indexing to knowledge exploration: A social learning model. *IEEE Intell. Syst.* 25, 4, 15–23.
- Fu, W.-T. and Dong, W. 2010b. Faciliating knowledge exploration in folksonomies: Expertise ranking by link and semantic structures. In *Proceedings of the International Conference on Computational Science and Engineering*. 66–73.
- Fu, W.-T. and Kannampallil, T. G. 2009. Harnessing Web 2.0 for context-aware learning: The impact of social tagging system on knowledge adaption. In *Educational Social Software for Context-Aware Learning: Collaborative Methods and Human Interaction*, N. Lambropoulos and R. Margarida Eds., IGI Global, Hershey, PA, 98–113.
- Fu, W.-T., Kannampallil, T. G., and Kang, R. 2009. A semantic imitation model of social tag choices. In *Proceedings of the IEEE Conference on Social Computing*. 66–72.
- Fu, W.-T., Kannampallil, T. G., and Kang, R. 2010. Facilitating Exploratory Search by Model-Based Navigational Cues. In Proceedings of the International Conference on Intelligent User Interfaces.
- Fu, W.-T. and Pirolli, P. 2007. SNIF-ACT: A cognitive model of user navigation on the World Wide Web. *Hum.-Comput. Interact.* 22, 355–412.
- Furnas, G. W., Fake, C., von Ahn, L., Schachter, J., Golder, S., Fox, K., Davis, M., Marlow, C., and Naaman, M. 2006. Why do tagging systems work? In *Extended Abstracts of the SIGCHI Conference on Human Factors in Computing Systems (CHI'06)*.
- Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. 1987. The vocabulary problem in human-system communication. *Comm. ACM* 30, 1: 964–971.
- Golder, S. A. and Huberman, B. A. 2006. Usage patterns of collaborative tagging systems. *J. Inform. Sci.* 322, 198–208.
- Griffiths, T. L., Steyvers, M., and Tenenbaum, J. B. T. 2007. Topics in semantic representation. *Psych. Rev.* 114, 2, 211–244.
- Halpin, H., Robu, V., and Shepherd, H. 2007. The complex dynamics of collaborative tagging. In *Proceedings of the 16th International Conference on World Wide Web*. ACM.
- HEDSTROM, P. 2006. Sociology: Experimental macro sociology: Predicting the next best seller. Science 311, 5762, 786–787.

- Hill, G. W. 1982. Group versus individual performance: Are N C 1 heads better than one? Psych. Bull. 91, 517–39.
- KANG, R. AND FU, W.-T. 2010. Exploratory information search by domain experts and novices. In Proceedings of the Conference on Intelligent User Interfaces.
- Kang, R., Fu, W.-T., and Kannampallil, T. G. 2010. Exploiting knowledge-in-the-head and knowledge-in-the-social-web: Effects of domain expertise on exploratory search in individual and social search environments. In *Proceedings of the ACM Conference on Computer-Human Interaction*.
- Kintsch, W. 1998. Comprehension: A Paradigm for Cognition. Cambridge University Press, Cambridge, UK.
- Landauer, T. K. and Dumais, S. T. 1997. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psych. Rev.* 104, 211–240.
- Macgregor, G. and McCulloch, E. 2006. Collaborative tagging as a knowledge organization and resource discovery tool. *Library Rev.* 55, 5, 291–300.
- Marchionini, G. 2006. Exploratory search: from finding to understanding. Comm. ACM 49, 4, 41–46.
- MILLEN, D. R., Yang, M., Whittaker, S., and Feinberg, J. 2007. Social bookmarking and exploratory search. In Proceedings of the 11th European Conference on Computer-Supported Cooperative Work.
- Newman, M., Barabasi, A-L. and Watts, D., Eds. 2006. The Structure and Dynamics of Networks. Princeton University Press, Princeton, NJ.
- Norman, D. A. and Rumelhart, D. E. 1975. Explorations in Cognition. Freeman, San Francisco, CA.
- PIROLLI, P. 2004. The infoclass model: Conceptual richness and inter-person conceptual consensus about information collections. *Cognit. Studies: Bull. Japanese Cognit. Sci. Soc.* 2, 197–213.
- PIROLLI, P. 2009. An elementary social information foraging model. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI'09)*.
- Pirolli, P. and Card, S. K. 1999. Information foraging. Psych. Rev. 106, 643-675.
- Qu, Y. AND FURNAS, G. 2008. Model-driven formative evaluation of exploratory search: A study under a sensemaking framework. *Inform. Process. Manage.* 442, 534–555.
- ROEDIGER, H. L. AND McDermott, K. B. 1995. Creating false memories: Remembering words not presented in lists. J. Exper. Psych.: Lear., Mem. and Cognit. 21, 803-814.
- Rumelhart, D. E. and Ortony, A. 1976. The representation of knowledge in memory. Tech. rep., Center for Human Information Processing, Department of Psychology, University of California, San Diego.
- Russell, D. M., Stefik, M. J., Pirolli, P., and Card, S. K. 1993. The cost structure of sense making. In *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*. ACM.
- Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, M. F., and Riedl, J. 2006. Tagging, communities, vocabulary, evolution. In *Proceedings of the 20th anniversary conference on Computer Supported Cooperative Work*. ACM.
- Steiner, I. D. 1972. Group Process and Productivity. Academic Press, New York.
- Thom-Santelli, J., Muller, M. J., and Millen, D. R. 2008. Social tagging roles: Publishers, evangelists, leaders. In *Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems*. ACM.
- ZIPF, G. 1949. Human Behavior and the Principle of Least Effort. Addison-Wesley.

Received February 2010; accepted March 2010 by Scott Robertson