Collaborative Indexing and Knowledge Exploration: A Social Learning Model

Wai-Tat Fu and Wei Dong, *University of Illinois at Urbana-Champaign*

A computational social learning model shows how meaning is constructed, shared, and learned through social tags contributed by Web users as they perform exploratory search.

The World Wide Web has evolved from a read-only information resource to a participatory environment that lets people share, explore, and learn through multiple forms of user-generated content (such as blogs and photos). A social information system, for example, is a form of social learning in which multiple users engage in knowledge exploration. Learning in such situations involves finding and evaluating relevant documents related to the topic, comprehending and extracting information from the documents, and integrating the information with existing knowledge. This form of knowledge exploration becomes social when users share found documents through systems such as Delicious (formerly del.icio.us), a site that lets users collaboratively index documents with short "tags" and share them with other users. Such collaborative indexing using social tags can not only provide structures to information on the Web but can also act as navigational cues for other users to find relevant information.

We focus on these kinds of collaborative or social tagging systems. Researchers have argued that social tagging systems can effectively improve knowledge exploration and sense-making activities,[1,2] and studies have analyzed the knowledge structures in these systems, making them an ideal testbed for social learning. We describe a social learning model that characterizes the iterative process of knowledge exploration and learning activities. We also present results from an empirical study that directly tested the social learning model.

Social Learning in Social Tagging Systems

Social tagging systems allow collaborative indexing of a massive information space based on individual users' subjective interpretation of information. Collaborative human indexing not only allows better representation of semantics that other humans can easily interpret, but also lets people with different knowledge backgrounds and information needs share their interpretations of different information.[3,4]

Researchers have found that tag dynamics tend to stabilize spontaneously as the number of users increases.[5,6] In previous research, we argued that this spontaneous stabilization might be because the inherent cognitive and semantic structures of users in an online community tend to be similar;

1541-1672/12/$31.00 © 2012 IEEE
Published by the IEEE Computer Society

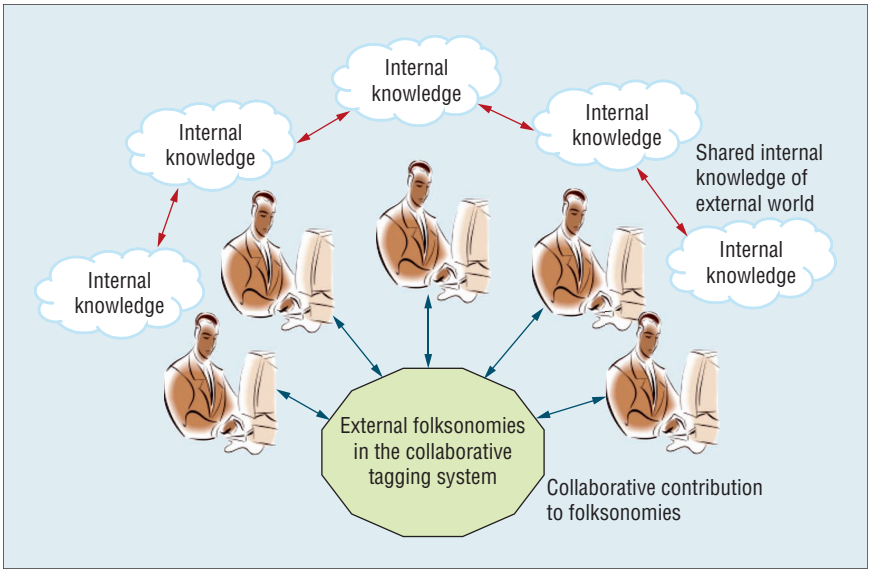

Figure 1. The semantic imitation process allows adaptation of internal concepts through interpretation and collaborative contribution to the folksonomies inherent in the collaborative tagging system.

thus, they act as latent structures that provide the cohesive forces behind the otherwise undirected tagging behavior. Based on this idea, we developed a semantic imitation model of tag choices to explain many emergent structures of large-scale social tagging systems.[1,2] Our major assumption was that when users assign tags to a webpage, their choice of tags is influenced not only by the page's content but also by how other users have tagged the same or similar pages.[1,2,6] In other words, as a person searches for information related to a topic, the person learns the context (related topics) associated with the topic based on the tag-document and tag-tag structures contributed by other users, and their semantic structures become more similar to those of other users through the knowledge exploration process.[2] In this article, we show how the latent structures defined by the semantic representation of knowledge could change as users interact with a social tagging system. In other words, the model assumes that there is *mutual influence* between the users' internal knowledge structures and the external folksonomies in the social tagging system.[1]

The current model assumes that as people interact with their environment and acquire more experiences, their knowledge can change to make sense of the new experiences. This notion of knowledge adaptation can be traced back to Piaget's developmental model of equilibration of cognitive structures in children,[7] and other prominent theories of knowledge representation and acquisition have adopted it also. According to Piaget, new experiences interact with existing concepts in at least two ways. *Assimilation* is the process of modifying new experiences to fit existing concepts. In this case, existing concepts influence how we interpret new information extracted from documents. In contrast, *accommodation* is the process of changing the concepts to fit the new experience, whereby the person creates an entirely new schema (or mental catgories) to accommodate new data that does not fit any of their existing concepts. Through knowledge assimilation and accommodation, people can adapt to the new experiences they gain from interacting with others, such as when they discuss, share, or exchange information.

A Social Learning Model of Exploratory Search

Figure 1 shows a notational diagram of the theoretical framework underlying the social learning model. Multiple users have their own internal knowledge representations (concepts and mental categories), and they interact with the social tagging systems by assigning tags to multiple documents as they consume information through the system. Internal knowledge representations partially reflect the users' background knowledge as well as differences in their information needs. The connections among users, tags, and indexed documents define the external folksonomies in the system that users can explore and learn from.

The model assumes that internal knowledge representations will influence how users interpret information in webpages, the tags created by others, and the tags they assign to webpages that they bookmark. In previous research, we showed that the interpretation process will influence users' internal knowledge representations through *semantic imitation*.[2,3] In semantic imitation, both internal and external representations can influence the search and interpretation of the Web document. In addition, the understanding and interpretation of Web documents can influence both the internal (concepts) and external representations (tagged content) of knowledge. Previous research[1,2] supports the idea that semantic imitation might be one of the spontaneous processes in social information systems that contribute to emergent behavioral patterns and structures.[5,6]

To formalize the the mechanisms just described, we assume that a user has a set of mental categories S and is performing a knowledge exploration task related to a topic T_j. The information goal is to predict whether the user can

find topic T_j by following a link with tags G. In other words, the user tries to estimate the probability $P(T_j|S,G)$ when deciding on links.

We can break this probability down into two components. One component is an estimate of the probability $P(S_m|G)$ that a document with tags G belongs to a particular mental category S_m. This estimate depends on how well the internal and external representations of information match: the higher the match, the better the model can predict which mental category the document belongs to. It also provides a measure of tag quality because it indicates how much the tags can invoke the user's set of relevant mental categories. The second component predicts the probability $P(T_j|S_m)$ of finding a particular topic in a given mental category S_m, which depends on the relationship between the topics and the mental category.

The overall probability $P(T_j|S,G_k)$ is the product of these two probabilities over the concepts of the user:

$$p(T_j \mid S,G) = \sum_m P(S_m \mid G)P(T_j \mid S_m).$$

Assimilation: Enriching Mental Concepts

Assume we have a set of mental categories that people might have about certain topics. We can use the Bayes theorem to estimate the prior probabilities for each of these mental categories and calculate how likely a tag created by a user reflects a particular set of mental categories. Specifically, if $P(S_m)$ is the prior probability of mental category S_m, and $P(G|S_m)$ is the conditional probability that tag G belongs to S_m, we can calculate $P(S_m|G)$ by

$$P(S_m \mid G) = \frac{P(S_m)P(G \mid S_m)}{\sum_m P(S_m)P(G \mid S_m)}.$$

We can estimate the conditional probability $P(G|S_m)$ by the ratio of the number of members in S_m that contain G to the total number of members in G: $P(G|S_m) = n_m/n$. To estimate the prior probability $P(S_m)$, we assume that there exists a *coupling probability*, c, that any two documents contain the same topic in a particular informational ecology. Figure 2 shows how the coupling probability can capture the relationship between topic distributions in the folksonomies and structures of the user's internal concepts. The assumption is that in an information space, users can probabilistically categorize the contents of a Web document into multiple topics. The probability that a set of n documents can be partitioned into S concepts can then be derived in terms of the coupling probability. The user can categorize a new document into an existing concepts—in other words, $P(S_m)$—or assign it to a new schema—that is, $P(S_0)$. The diagram in Figure 2a shows the probability of any partition x for n objects in S categories. We derive $P(x)$ by

$$P(x) = \frac{\prod_{k=1}^{s}(1-c)\prod_{k=1}^{n_k-1} ci}{\prod_{i=0}^{n-1}\left[(1-c)+ci\right]},$$

where c is the coupling probability that any two objects will belong in the same category, and n_k is the number of objects in category k.

Next, we derive the probability $P(x_m)$ of a new partition x_m with a new object in category m, and the probability, $P(x_0)$, of a new partition x_0 with a new object in a new category by

$$\left.\begin{aligned} P(x_m) &= P(x)\frac{cn_m}{(1-c)+nc} \\ P(x_0) &= P(x)\frac{1-c}{(1-c)+nc} \end{aligned}\right\} \sum_{m=0}^{s} P(x_m)$$
$$= P(x).$$

Equation 1 shows how we derive the probability $P(S_m)$ that a new object belongs in category m:

$$P(S_m) = \frac{cn_m}{(1-c)+nc}. \quad (1)$$

Finally, Equation 2 shows the probability that a new object belongs in a new category:

$$P(S_0) = \frac{1-c}{(1-c)+nc}. \quad (2)$$

As Figure 2b shows, as the number of documents (n) increases, the value of $P(S_m)$ approaches n_m/n, and the value of $P(S_0)$ approaches zero. This implies that without any information cue, users tend to favor popular topics (those with larger n_m) over unpopular topics. In addition, the likelihood that users perceive a document to contain a new topic decreases, creating the "rich get richer" effect observed in many social information systems. Figure 2b also shows that as c increases from 0 to 1, the value of $P(S_m)$ increases and the value of $P(S_0)$ decreases. This implies that as the information ecology changes from low overlap (where documents contain few common topics) to high overlap, the probability of forming a new mental category will decrease. In other words, the model predicts that knowledge exploration in a high-overlap environment will lead to formation of fewer internal mental categories than in a low-overlap environment. For example, if someone is looking for topics related to the independence of Kosovo, found articles will likely have many overlapping facts related to this event. However, if someone is looking for topics related to anti-aging, articles may have less overlap because there are more disjoint topics related to the subject—skin care, genetics, nutrition, and so on.

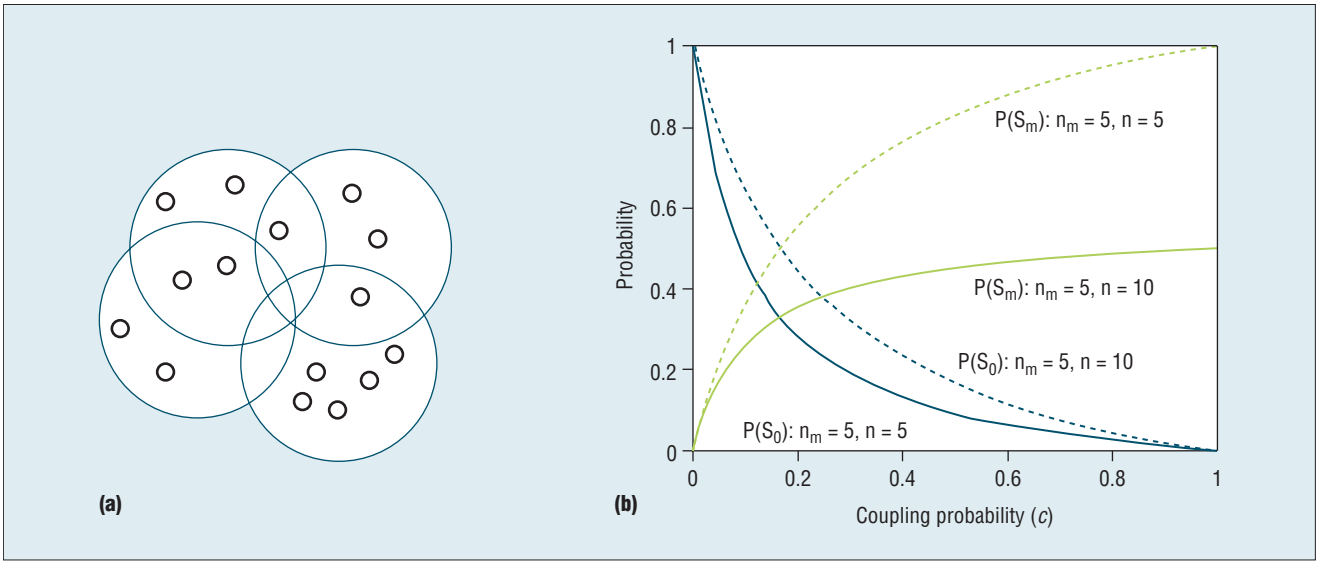

Figure 2. The relationships between topic distributions in an information ecology and prior probabilities of concepts. (a) An example of a partition with 15 objects (*n*) in four categories (*S*). (b) The change in prior probabilities $P(S_m)$ and $P(S_0)$ as a function of the coupling probability (*c*).

Accommodation: Formation of New Mental Categories

When users encounter a document that does not fit into any existing concepts, they add a new mental category to accommodate the new discoveries. This decision reflects the value of $\max[P(S|G)]$, where G represents the document's contents and tags, S represents the schema, and the max operation is performed in the set S. Specifically, a new mental category will be created only if

$$P(S_0|G) > \max[P(S|G)],$$

or when the probability that the document belongs to a new mental category (see Equations 1 and 2) is larger than the probability that it belongs to any existing mental category.

Assigning Tags

Given an existing tag G_k, the model will calculate the value of $P(S_m|G_k)$, where S_m is the mental category to which the current document belongs. The model will assign this tag G_k to this document only if

$$P(S_m|G_k) > \tau_{threshold}, \quad (3)$$

where $\tau_{threshold}$ is a parameter that represents the general willingness to create a new tag. The model creates a new tag G_0 only if

$$P(S_m|G_0) > \max[P(S_m|G)],$$

that is, if a new tag (from the user's previous vocabulary) can predict category m better than any of the document's existing tags. In other words, the model assumes that users will assign tags to best represent the topics in the document. By extracting all tags used and created by participants, we can match the assignment and creation of tags by the model to those of the participants in the exploratory task.

Empirical Study

We designed a set of exploratory learning tasks to test the model. In all tasks, we gave participants a rough description of the topic, and they gradually acquired knowledge about the topic through an iterative search-and-learn exploration cycle. We instructed them to imagine that they wanted to understand the given topic and to write a paper and give a talk to a diverse audience. We chose two general topics:

- Find relevant facts about the independence of Kosovo.
- Find relevant facts about anti-aging.

We chose these tasks after a series of pilot studies showed they were representative of general exploratory search tasks.[8] In addition, they represent two different distributions of the information ecology. Specifically, because the first task (independence of Kosovo) refers to a specific event, information related to it tends to be more specific, and there were more websites containing multiple pieces of well-organized information relevant to the topic. The second task (anti-aging), on the other hand, is more ambiguous and relates to many disjoint areas such as cosmetics, nutrition, and genetic engineering. Because websites relevant to the first task have more overlapping concepts than those relevant to the second task, we call them high-overlap and low-overlap tasks, respectively. Because the low-overlap task is more general, the tags tended to be more generic (such as "beauty" and "health"). In contrast, for the high-overlap task, tags tended to be more semantically narrow (such as "Kosovo"), and thus had higher cue validity than generic tags.

1. Initialize number of category (k) to 0.
2. For documents $D_i = D_1$ to D_n:
   a. Extract words and tags G from D_i : $D_i = \{G_1, G_2, ...\}$
   b. Find m such that $P(S_m/D_i) > P(S_j/D_i)$ for all $j = 0$ to k
   c. Assign D_i to category m
   d. Update $P(S_m/D_i)$, $P(S_m)$, $P(S_0)$, and k
   e. Find G_x such that $P(S_m/G_x) > P(S_m/G)$ for all G from D_1 to D_i
   f. If $P(S_m/G_x) > \tau$, assign G_x as a new tag to D_i

Figure 3. Pseudocode for the model simulations.

In the representative design tradition,[9] we chose to follow a small number of subjects over a period of eight weeks to keep close track of their interactions with the system. We recruited eight participants from the University of Illinois at Urbana-Champaign, divided them randomly into two groups, and assigned each to one of the tasks. The participants' self-reports made it obvious that they were unfamiliar with the given topics. We told them to explore all relevant information on the topic using the search function in Delicious or any other Web search engine, create tags for webpages they found relevant, and store the tags in their Delicious accounts. We asked participants to create the tags for two major purposes:

- to allow them to refind the information quickly in the future, and
- to help their colleagues use the relevant information easily in the future.

Because previous research has shown that the impact of the interfaces on knowledge acquisition might depend greatly on each subject's idiosyncratic learning patterns and background knowledge, we analyzed the results for each individual separately and compared them to the model rather than matching model results to group averages.

Procedure

Each student performed the task for eight 30-minute sessions over a period of eight weeks, with sessions approximately one week apart. We instructed participants to think aloud during the task and to provide a verbal summary of every webpage they read before creating tags for the page, after which they could bookmark the webpage and create tags. After they finished reading a document, they could either search for new documents by initiating a new query or selecting an existing tag to browse documents tagged by others. This exploratory search-and-tag cycle continued until the end of the session.

We captured screen interactions using the Camtasia screen recording software, and we manually recorded all bookmarks and tags from participants' Delicious accounts after each session. We extracted all tags used and aligned them with the verbal protocols in each session to track changes in the external and internal representations and during the exploratory search process.

After the last session, we asked the participants to perform a categorization task. We gave them printouts of all the webpages they had read and bookmarked during the task as well as the tags associated with the pages (both their own and other participants'). We then asked them to "put together the webpages that go together on the basis of their information content into as many different groups as you'd like." Finally, we matched the concepts formed by the participants to those predicted by the assimilation and accommodation processes in the social learning model.

Results

Participants on average created 90.2 bookmarks and 425.4 tags for the high-overlap task, and 42.2 bookmarks and 212.3 tags for the low-overlap task. Participants in the high-overlap task (Kosovo) created more bookmarks and assigned more tags than those in the low-overlap task (anti-aging), but the average number of tags per bookmark is about the same (5.2 tags per bookmark) for the two tasks. As expected, finding relevant information for the low-overlap task was more difficult, as reflected by the fewer bookmarks created. Given that distribution of information was more disjoint in the low-overlap task, the results were consistent with the assumption that the average rate of return of relevant information was lower for the low-overlap task than for the high-overlap task.

We performed separate model simulations for each participant based on the documents and tags that person interacted with. Figure 3 shows the pseudocode for the simulations.

Figure 4 shows the proportion of new tags assigned by each participant and the corresponding model simulations, and Table 1 shows the match between the model and each participant. Interestingly, even though participants assigned fewer tags in the low-overlap task, the proportions of new tag assignments to total numbers of tag assignments were higher in the low-overlap task than in the high-overlap task. This was consistent with the lower rate of return of relevant information in the low-overlap task, possibly because the existing tags on Delicious were less informative for that task. Indeed, concepts extracted by the participants in the low-overlap task differed from the existing tags more frequently than in the high-overlap task, suggesting that the existing tags did not serve as good cues to information contained in the documents.

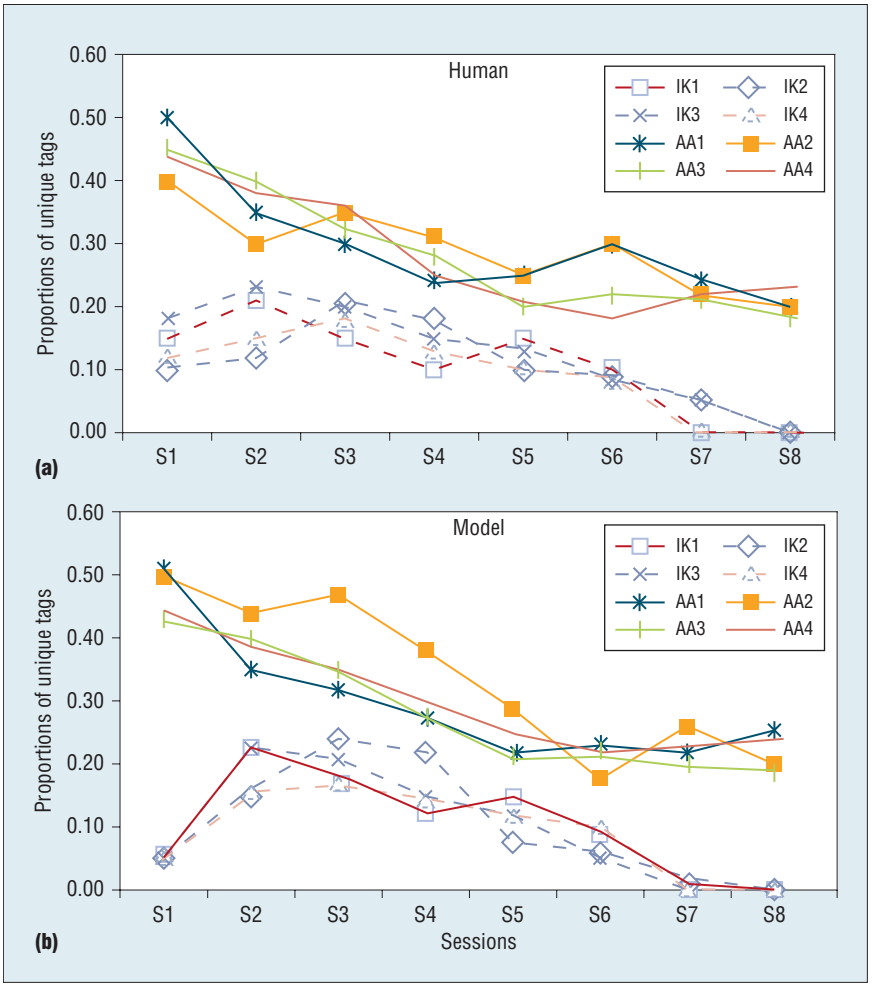

Figure 4. The mean proportions of unique tag assignment for the high-overlap (IK) and low-overlap (AA) tasks by (a) participants and (b) the model across the eight sessions. IK1 represents participant 1 in the IK task, AA1 represents participant 1 in the AA task, and so on.

Table 1. The match between the model and each participant.

Participant	R^2
IK1	0.76
IK2	0.92
IK3	0.74
IK4	0.82
AA1	0.84
AA2	0.62
AA3	0.98
AA4	0.97
Mean	0.83

The model closely matched the general trends and differences between the two tasks (average $R^2 = 0.83$, min = 0.62, max = 0.98). The only free parameter, $\tau_{threshold}$, controlled the model's willingness to assign tags (see Equation 3) and was set to 0.2 for all participants to best fit the data. Although we could set a different τ for each participant (the average would be $R^2 = 0.96$, min = 0.94, max = 0.99), we chose to balance the number of free parameters and the fit to the data. Nevertheless, the current results demonstrated the ability of the model to track the tag assignments for each participant based on that individual's history.

As Figure 4 shows, the major mismatches occurred in the first sessions, when the model tended to underpredict the creation of new tags, especially for the high-overlap task. That's because the model assumed that the participants had no background knowledge about the topic—in other words, the model has no common sense knowledge, a well-known problem in AI. However, the model quickly picked up new concepts beginning in the second session.

We also created a model that randomly assigned tags. Chi-square tests show that both human and model performance was significantly higher than the chance model ($p < 0.01$).

Formation of Mental Categories

One core assumption of the social learning model was that the assignment of tags and the selection of links were dynamically related to the set of mental categories formed during the knowledge exploration cycles. It is therefore critical to verify that the set of mental categories formed by the model matched those formed by the participants. To do this, we constructed match tables for each participant and model. We gave items in the same mental categories the value 1, and items in different mental categories the value 0. For example, two possible partitions (categorizations) for the set (*a*, *b*, *c*, *d*, *e*) are (*a*, *b*), (*c*, *d*), (*e*) and (*a*, *b*, *c*), (*d*, *e*). In this example, the partitions' correlation is $r = 0.102$ based on the match table in Table 2.

Table 3 shows the number of mental categories formed by each participant and the model, as well as the correlations between their partitions. As predicted, participants formed more mental categories in the low-overlap task, reflecting the structures of the information sources. However, as discussed earlier, participants in the low-overlap task also had a lower rate of return in their information search and thus fewer tags overall

(but more unique tags, as shown in Figure 1). Apparently, mental categories in the low-overlap group tended to be more general than those in the high-overlap group, presumably because the saved documents had less shared content and were therefore grouped under more general mental categories. In contrast, documents in the high-overlap group tended to be more specifically related to the independence of Kosovo, and thus the mental categories were more specific. The correlations between the participants and the models were high in both tasks, suggesting that the model not only created a similar number of mental categories as participants, but the partitions of the mental categories between the two tasks were also similar, even though the inherent information structures differed.

Table 2. An example match table that calculates the correlation between two partitions of objects.

Partition									
	(a, b), (c, d), (e)					(a, b, c), (d, e)			
	a	*b*	*c*	*d*		*a*	*b*	*c*	*d*
b	1				*b*	1			
c	0	0			*c*	1	1		
d	0	0	1		*d*	0	0	0	
e	0	0	0	0	*e*	0	0	0	1

Table 3. Number of categories formed by participants and the model, and the correlations of the category partitions of the models and the students calculated using the match tables.

	No. of categories (human)	No. of categories (model)	Correlations of partitions
Hi-S1	6	6	0.71
Hi-S2	5	6	0.68
Hi-S3	7	7	0.81
Hi-S4	5	6	0.86
Lo-S5	12	13	0.59
Lo-S6	10	11	0.67
Lo-S7	11	12	0.79
Lo-S8	10	10	0.87

Notes: Hi = High-overlap task, Lo = Low-overlap task, S1 = participant 1, S2 = participant 2, and so on.

As far as we know, ours is the first study to show that collaborative tagging systems can facilitate not only the indexing of information but also social learning through the processes of knowledge assimilation and accommodation that are typically found in traditional social learning situations. The current results also show that it has the potential to promote formal or informal learning of diverse topics and the development of common concepts or understanding within or across different communities. In addition, given the direct impact on the development and refinement of mental categories, it is not hard to imagine that social tagging systems can also facilitate collaborative activities that involve higher-level cognitive processing, such as problem solving, decision making, or creative designs. It seems that we have only started to harness the potential of socio-technological systems, especially for areas such as education and scientific knowledge sharing and understanding.

The current social learning model provides design guidelines for future social tagging systems. For example, the model can facilitate development of data-mining tools that extract external folksonomies and infer the underlying mental categories of people who have different domain expertise by analyzing their selection and creation of tags. The current model can also be combined with knowledge engineering tools to facilitate knowledge adaptation by users in different domains. More generally, the model demonstrates how research on human learning processes can be combined with machine learning techniques to allow better human-system integration. It also highlights the importance of studying how information cues generated by machine-learning methods are *actually* used by human users to better understand whether they can help users to achieve their social goals.

References

1. W.-T. Fu, "The Microstructures of Social Tagging: A Rational Model," *Proc. 2008 ACM Conf. Computer Supported Cooperative Work* (CSCW 2008), ACM, 2008, pp. 229–238.
2. W.-T. Fu, T.G. Kannampallil, and R. Kang, "A Semantic Imitation Model of Social Tag Choices," *Proc. 2009 IEEE Int'l Conf. Social Computing,* IEEE CS, 2009, pp. 66–72.
3. W.-T. Fu et al., "Semantic Imitation in Social Tagging," *ACM Trans. Computer-Human Interaction,* vol. 1, no. 3, 2012, pp. 1–37.
4. H. Halpin and H.S. Thompson, "Social Meaning on the Web: From Wittgenstein to Search Engines," *IEEE Intelligent*

THE AUTHORS

Wai-Tat Fu is an assistant professor in the Department of Computer Science at the University of Illinois at Urbana-Champaign and a research faculty member at the Beckman Institute of Science and Technology and the Department of Psychology. His research interests include cognitive science, socio-computer interaction, human-computer interaction, and social computing. Fu has a PhD in psychology and cognitive science from George Mason University and completed his postdoctoral training at Carnegie Mellon University. He is a member of the ACM and IEEE, an associate editor of *Topics in Cognitive Science*, and a member of the editorial board of the *Journal of Experimental Psychology: Applied*.

Wei Dong received her master's degree from the Human Factors Division at the University of Illinois Urbana-Champaign and the Beckman Institute of Science and Technology. Her research interests include socio-computer interaction, human-computer interaction, and computer-mediated collaborative work. Dong has a master's degree in Human Factors and a master's degree in psychology from the University of Illinois at Urbana-Champaign. She is currently pursuing her PhD in information science at Cornell University. She is a member of the ACM, IEEE, and Human Factors and Ergonomics Society.

Systems, vol. 24, no. 6, 2009, pp. 27–31.
5. C. Cattuto, V. Loreto, and L. Pietronero, "Semiotic Dynamics and Collaborative Tagging," *Proc. Nat'l Academy of Sciences*, vol. 104, no. 5, 2007, pp. 1461–1464.
6. S.A. Golder and B.A. Huberman, "Usage Patterns of Collaborative Tagging Systems," *J. Information Science*, vol. 32, no. 2, 2006, pp. 198–208.
7. J. Piaget, *The Equilibration of Cognitive Structures*, Univ. of Chicago Press, 1975.
8. G. Marchionini, "Exploratory Search: From Finding to Understanding," *Comm. ACM*, vol. 49, no. 4, 2006, pp. 41–46.
9. E. Brunswik, *Perception and the Representative Design of Psychological Experiments*, University of California Press, 1956.

Selected CS articles and columns are also available for free at http://ComputingNow.computer.org.