

# From Collaborative Indexing to Knowledge Exploration: A Computational Social Learning Model

**Wai-Tat Fu & Wei Dong**

Applied Cognitive Science Lab  
University of Illinois at Urbana-Champaign  
405 N. Mathews Ave., Urbana, IL 61801, USA

## KEYWORDS

Social Learning, Knowledge exploration, Cognitive Model, Social Tagging, Collective Intelligence, Semantic Imitation, knowledge assimilation, knowledge accommodation

## ABSTRACT

Despite the increasing popularity of social information systems, there is still a general lack of research that studies the effectiveness of these systems in facilitating social learning activities such as knowledge exploration, sharing, and exchange. A formal theory of the underlying individual cognitive processes of social learning will provide useful predictions on how changes in interface representations and interaction methods may impact effectiveness of these systems in facilitating learning and knowledge exploration by multiple users. Based on theories of human concept formation and semantic imitation of social tagging, a computational social learning model was developed and tested against longitudinal empirical data as participants performed knowledge exploration tasks across a period of eight weeks. Results showed that the quality of social tags and distributions of information contents directly impacted learning and search performance. The results and the model have important implications on how Web 2.0 technologies should be designed to optimize the match between human and machine learning processes to facilitate knowledge exploration in social media that foster collective intelligence contributed by multiple users. The model also demonstrates how computational models that keep track of cognitive changes of individuals as they interact with a social information system can complement data-mining techniques to predict how different information cues will be realistically interpreted and utilized by human users.

## INTRODUCTION

The World Wide Web (WWW) has evolved from a “read-only” information resource to a participatory environment that allows people to share, explore, and learn through multiple forms of user-generated contents (e.g., blogs, photos). One form of social learning that has attracted much attention is when multiple users engage in knowledge exploration in a social information system. Learning in such situations involve finding and evaluating relevant documents related to the topic, comprehending and extracting information from the documents, and integrating extracted information with existing knowledge. This form of knowledge exploration becomes social when users share documents they found through systems such as del.icio.us, which allows users to collaboratively index the documents with short keywords called “tags” and share them with other users. The collaborative indexing using social tags not only can provide structures to information on the Web, but they can also act as navigational cues for other users of the system to find relevant information.

In this article, we focus on a popular participatory Web technology called collaborative or social tagging systems. Researchers have argued that social tagging systems can be applied to effectively improve knowledge exploration and sense-making activities [3, 4], and there have been studies analyzing the structures of knowledge in these systems, making it an ideal “test bed” for social learning. We will first briefly discuss the history and characteristics of social tagging, followed by a description of a social learning model that characterizes the iterative process of knowledge exploration and learning activities. Finally, we will describe results from an empirical study that directly tested the social learning model.

## SOCIAL LEARNING IN SOCIAL TAGGING SYSTEMS

One important feature of social tagging systems is that they allow collaborative indexing of the massive information space based on the subjective interpretation of the information by different users. Collaborative human indexing not only allows better representation of semantics that are interpreted at the level that other humans can easily understand, it also allows multiple interpretations by people with different knowledge background and information needs to share their understanding of different information contents [5, 7].

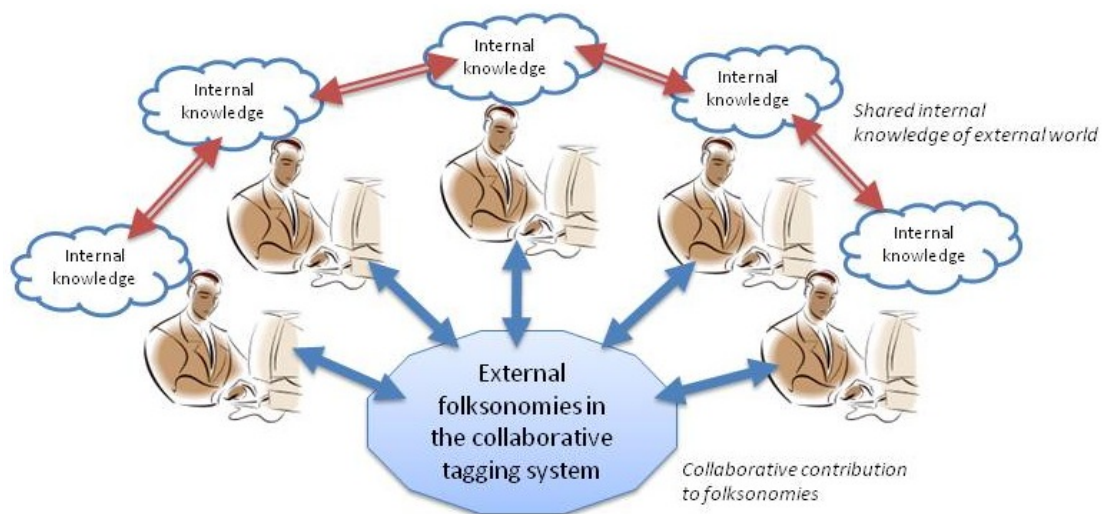
Researchers have found that the dynamics of tags tend to stabilize spontaneously as the number of users increases [2, 6]. In previous research, we argue that one reason for the spontaneous stabilization can be attributed to the fact that the inherent cognitive and semantic structures of users in an online community tend to be similar; thus, they can be treated 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 that exist in large-scale social tagging systems [3, 4]. The major assumption of the model is that, when users assign tags to Web pages, the choice of tags is not only influenced by the information contents of the Web page, but also how other users have tagged the same or similar Web pages [3, 4, 6]. In other words, as a person explores for information related to a topic, the person will learn the context (i.e., related topics) associated with the topic based on the tag-document and tag-tag structures contributed by other users of the system, and their semantic structures will become more similar to those of other users through the knowledge exploration process [4]. In this article, we will 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 internal knowledge structures of the users and the external folksonomies in the social tagging system [3].

The current model assumes that as people interact with their environment and acquire more experiences, their knowledge may be modified to make sense of the new experiences. This process of knowledge adaptation can be traced all the way back to the Piaget's [9] developmental model of equilibration of cognitive structures in children, and it has also been adopted by other prominent theories of knowledge representation and acquisition. According to Piaget, there are at least two processes through which new experiences interact with existing concepts. When new experiences are modified to fit existing concepts the process is defined as *assimilation*. In this case, existing concepts influence how we *interpret* new information extracted from documents. In contrast, *accommodation* is an adaptation process of knowledge acquisition that changes the concepts in order to fit the new experience, or the person creates an entirely new schema in order to accommodate new data that does not fit any of their existing concepts. Through the process of knowledge assimilation and accommodation, people can adapt to new experiences that they obtain from their interactions 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/or 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 different background knowledge of users, as well as differences in their information needs. The connections among the users, tags, and indexed documents define the external folksonomies in the system, from which users can explore for information and learn.



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

The model assumes that internal knowledge representations will influence how users interpret information in different web pages, the tags created by others, as well as the tags they assign to web pages that they bookmark. In previous research, we have shown that the interpretation process will influence users' own internal knowledge representations through a *semantic imitation* process [4, 5]. The major characteristics of this semantic imitation process is that: (1) both internal and external representations may influence the search and interpretation of the web document, and (2) the understanding and interpretation of web documents may influence both the internal (concepts) and external representations (tagged contents) of knowledge. Previous research [3, 4] has provided evidence supporting the idea that semantic imitation may be one of the spontaneous processes in social information systems that contribute to emergent behavioral patterns and structures in the systems [2, 6].

### ***A formal model of social learning***

To formalize the mechanisms described above, one can 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 topic  $T_j$  can be found by following a link with tags  $G$ , i.e., the user is trying to estimate this probability:  $P(T_j|S, G)$  when deciding on links. This probability can be broken down into two components. One component predicts the probability  $P(T_j|S_m)$  that a particular topic can be found in a given mental category  $S_m$ , and the second component predicts the probability  $P(S_m|G)$  that the given set of tags are associated with a given mental category in this equation:

$$p(T_j | S, G) = \sum_m P(S_m | G) P(T_j | S_m) \quad (1)$$

In other words, to predict whether topic  $T_j$  can be found in a particular document, one can first estimate  $P(S_m|G)$ : the probability that the document with tags  $G$  belongs to a particular mental category  $S_m$ . This estimate depends on how much the internal and external representations match each other: The higher the match, the better is the model able to predict to which mental category the document belongs. It also provides a measure of the “quality” of tags, as it indicates how much the tags may invoke the set of relevant mental categories of the user. The second estimate  $P(T_j|S_m)$  is the probability that topic  $T_j$  can be found in concepts  $S_m$ , which depends on the relationship between the topics and the schema. The overall probability  $P(T_j|S, G_k)$  can then be estimated by enumerating the product of these two probabilities over the concepts of the user.

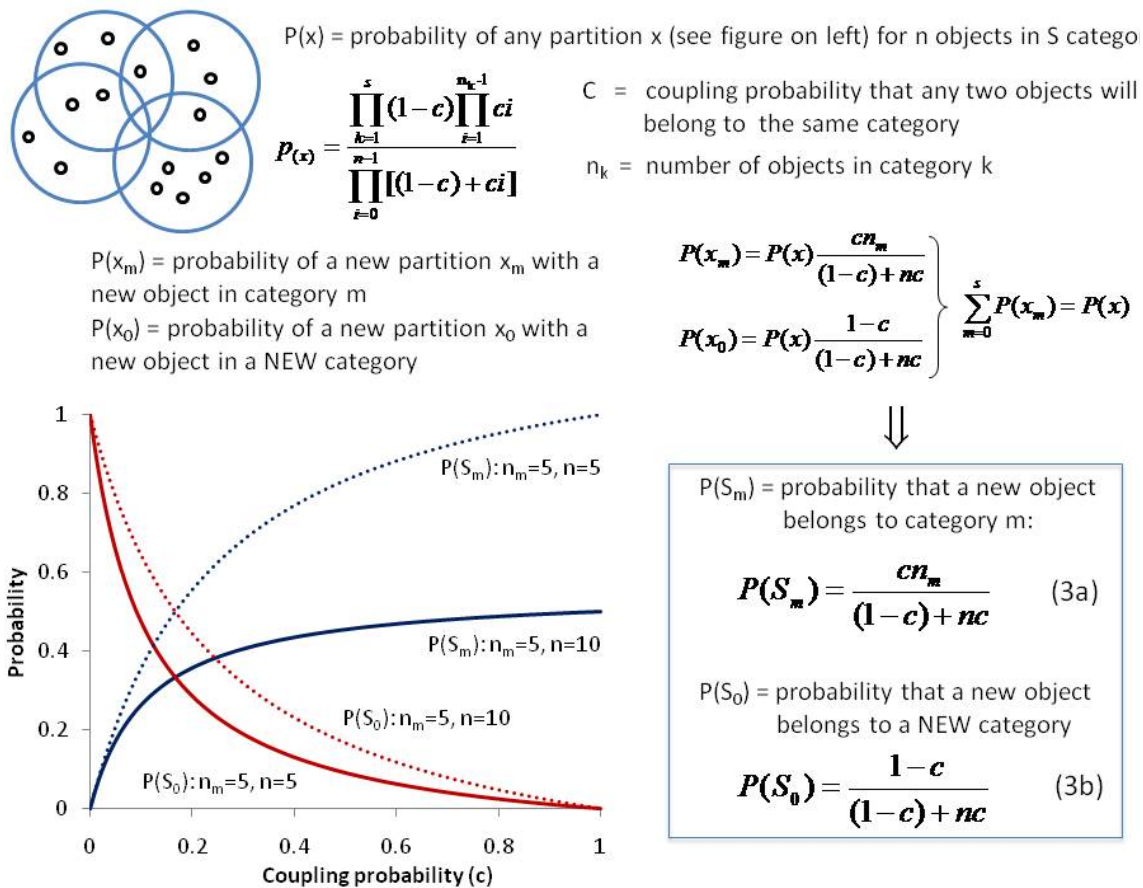
### **Assimilation: Enrichment of mental concepts**

If we assume a set of mental categories that people may have about certain topics, one can first estimate the prior probabilities for each of these mental categories, and calculate how likely a tag created by a user is created based on a particular mental categories by the Bayes theorem. 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  $S_m$ , then we can calculate  $P(S_m|G)$  by:

$$P(S_m | G) = \frac{P(S_m)P(G | S_m)}{\sum_m P(S_m)P(G | S_m)} \quad (2)$$

The conditional probability  $P(G|S_m)$  can simply be estimated by the ratio of the number of members in  $S_m$  that contains  $G$  and the total number of members in  $G$ , i.e.,  $P(G|S_m)=n_m/n$ . To

estimate the prior probability  $P(S_m)$ , one can assume that there exists a *coupling probability*,  $c$ , for any two documents to contain the same topic in a particular informational ecology. Figure 2 shows how the coupling probability can be used to capture the relationship between topic distributions in the folksonomies and structures of internal concepts of the user. The assumption is that, in an information space, each element (e.g., Web document) may contain contents that allow users to probabilistically categorize it 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 (see top of Figure 4). When a new document is encountered, the user may categorize it into one of the existing concepts (i.e.,  $P(S_m)$ ), or it may be assigned to a new schema (i.e.,  $P(S_0)$ ). The derivation for these prior probabilities (see equation (3)) is shown in Figure 2.



**Figure 2. The relations between topic distributions in an information ecology and prior probabilities of concepts.**

As shown in the bottom of Figure 2, 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, popular topics (those with larger  $n_m$ ) tend to be favored over unpopular topics, and the likelihood that a document is perceived to contain a new topic will decrease, creating the common "the rich gets richer" effect as observed in many social information systems. Figure 2 also shows that as  $c$  increases from 0 to 1, the value of  $P(S_m)$  ( $P(S_0)$ ) increases (decreases). This implies that as the information ecology changes from a low-overlap (documents contain few common topics) to a high-overlap environment, the probability for the formation of a new



schema 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 one is looking for topics related to the “independence of Kosovo”, articles on this topic will likely have many overlapping facts related to this particular event; but if one is looking for topics related to “anti-aging”, articles may have less overlap, as there are more disjoint topics related to anti-aging, such as skin care, genetics, nutrition, etc.

#### Accommodation: Formation of new mental categories

When a user encounters a document that does not fit into any of his or her existing concepts, a new mental category will be added to accommodate for the new "discoveries". This decision was based on the value of  $\max[P(S|G)]$ , where  $G$  represents the contents and tags of the document and  $S$  represents the schema, and the max operation is performed in the set  $S$ . Specifically, a new mental category will be created only if

$$P(S_0|G) > \max[P(S|G)] \quad (4)$$

i.e., the probability that the document belongs to a new mental category (see Eqn (3)) is larger than that for it to belong to any of the existing mental categories. If this condition is met, a new mental category will be created.

#### Assigning tags to a bookmark of a web document

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

$$P(S_m|G_k) > \tau_{threshold} \quad (5)$$

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

$$P(S_m|G_0) > \max[P(S_m|G)] \quad (6)$$

i.e., the probability that a new tag (from the user's previous vocabulary) can predict schema  $m$  better than any of the existing tags assigned to the document. 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, the assignment and creation of tags made by the model could be matched to those made by the participants in the exploratory task.

### **EMPIRICAL STUDY**

A set of exploratory learning tasks was designed to test the model. In all tasks, participants were given a rough description of the topic and gradually acquired knowledge about the topic through an iterative search-and-learn knowledge exploration cycle. Participants were told to imagine that they wanted to understand the given topic and to write a paper and give a talk on the given topic to a diverse audience. Two general topics were chosen: (1) “Find out relevant facts about the Independence of Kosovo” and (2) “Find out relevant facts about Anti-aging”. These two tasks were chosen after a series of pilot studies that showed that they were representative of the general exploratory search tasks [8]. In addition, the two tasks were chosen to represent the two very different distributions of the information ecology. Specifically, because the first task (independence of Kosovo) referred to a specific event, information related to it tended to be more

specific, and there were more Web sites containing multiple pieces of well-organized information relevant to the topic. The second task (anti-aging), on the other hand, was more ambiguous and was related to many disjoint areas such as cosmetics, nutrition, or genetic engineering. Because Web sites relevant to the first task have more overlapping concepts than those relevant to the second task, they will be called high-overlap and low-overlap tasks respectively. The other characteristic is that because the low-overlap task was more general, the tags tended to be more generic (such as “beauty”, “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.

Following the tradition of representative design [1], we chose to follow a small number of subjects over a period of eight weeks to closely keep track of their interactions with the system. Furthermore, previous research has shown that the impact of the interfaces on knowledge acquisition might depend much on each subject's idiosyncratic learning patterns and background knowledge. For these reasons, we analyzed the results for each individual subject separately and compared them to the model rather than matching model results to group averages. We recruited eight participants from the University of Illinois at Urbana-Champaign. Participants were randomly split and assigned to one of the tasks. From their self-report it was obvious that they were unfamiliar with the given topics. Participants were told that they should explore all relevant information to learn about the topic using either the search function in del.icio.us or any other Web search engines, and should create tags for Web pages they found relevant to the topic and store them in their own del.icio.us accounts. Participants were told that these tags should be created for two major purposes. First, to allow them to re-find the information quickly in the future; second, to help their colleagues to utilize the relevant information easily in the future for their search purposes.

### ***Procedure***

Each student performed the task for eight 30-minute sessions over a period of 8 weeks, with each session approximately one week apart. Students were told to think aloud during the task in each session. All verbal protocols and screen interactions were captured using the screen recording software *Camtasia*. All tags created were recorded manually from their del.icio.us accounts after each session. Students were instructed to provide a verbal summary of every Web page they read before they created any tags for the page. They could then bookmark the web page and create tags for the page. 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 continues until a session ended. All tags used and created during each session were extracted to keep track of changes in the shared external representations, and all verbal description on the Web pages were also extracted to keep track of changes in the internal representations during the exploratory search process.

After the last session, participants were asked to perform a categorization task. Participants were given printouts of all web pages that they read and bookmarked during the task, and were given the tags associated with the pages (either by themselves or other members in del.icio.us). They were then asked to “put together the web pages that go together on the basis of their information content into as many different groups as you’d like”. The concepts formed by the participants were then matched 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 created more bookmarks and assigned more tags than those in the low-overlap task, 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 is more difficult, as reflected by the fewer number of 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.

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  $D_i$ :  $D_i = \{G_1, G_2, \dots\}$
  - 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. Pseudo code for the model simulations.**

Separate model simulations were performed for each participants based on the documents and tags that they interacted with (see Figure 3 for the pseudo code). Figure 4 shows the proportion of new tags assigned by each participant and the corresponding model simulations. Interestingly, even though participants assigned fewer tags in the low-overlap task, but the *proportions* of new tag assignment over 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 lower-overlap task, and this lower rate could be caused by fact that the existing tags on del.icio.us was less informative for the lower-overlap task. Indeed, concepts extracted from the documents by the participants in the lower-overlap task were more often different from the existing tags than in the high-overlap task, suggesting that the existing tags did not serve as good cues to information contained in the documents. The general trends and differences between the two tasks were closely matched by the model (average  $R^2=0.83$ , min=0.62, max=0.98). The only free parameter,  $\tau_{\text{threshold}}$ , controlled the willingness of the model to assign tags (see Eqn 5), and was set to 0.2 for all participants to best fit the data. While we could set a different  $\tau$  for each participant to fit individual participants (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 good match of the model in keeping track of the tag assignments for *each* participant based on their histories of processing of Web pages across a period of 8 weeks.



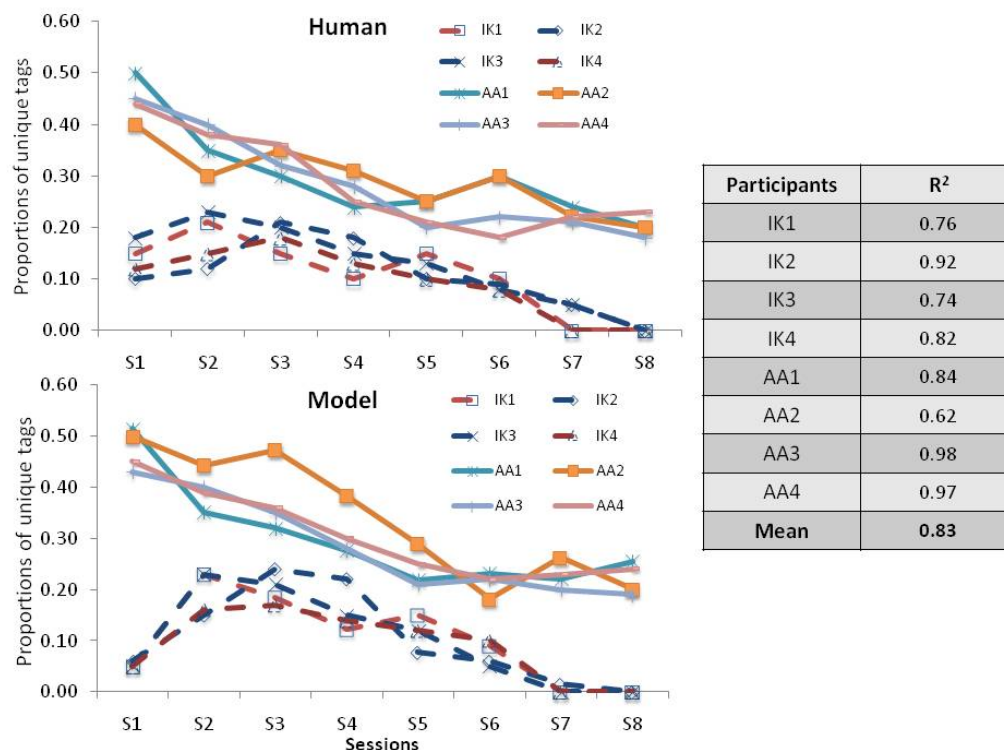


Figure 4. The mean proportions of unique tag assignment for the high-overlap (IK) and low-overlap (AA) tasks by participants (top) and the model (bottom) across the eight sessions. IK1 represents participant 1 in the IK task, AA1 represents participant 1 in the AA task, etc. The table (right) shows the match between the model and the each of the participants.

As shown in Figure 4, the major mismatches were found in the first sessions, where the model tended to under-predict the creation of new tags, especially for the high-overlap task. This was caused by the fact that the model assumed that the participants had no background knowledge about the topic at all, thus the model tended to under-predict the use of unique tags in the first session (i.e., the model has no common sense knowledge, a well-known problem in AI). However, the model quickly picked up new concepts starting from the second session. A model that randomly assigned tags was created and compared to performance by humans and model. Chi-square tests show that both human and model performance was significantly different from the chance model ( $p < 0.01$ ), showing that they were significantly above the chance level.

### 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, correlations between the mental categories formed by the model and the participants were calculated by constructing “match” tables for each participant and model. Items that are in the same mental categories will be given a value 1, otherwise a 0. For example, two possible partitions (categorization) for the set (a,b,c,d,e) are (ab),(c,d),(e) and (a,b,c), (d,e). In this example, their correlation can be calculated as  $r=0.102$  based on the match table as shown at the top of Figure 5.

(a)								
{a,b},{c,d},{e}					{a,b,c},{d,e}			
	a	b	c	d	a	b	c	d
b	1				1			
c	0	0			1	1		
d	0	0	1		0	0	0	
e	0	0	0	0	0	0	0	1

(b)			
	#categories (human)	#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

**Figure 5. (a) An example of the match table that calculates that correlation between two partitions of objects. (b) Number of categories formed by each participant and model, and the correlations of the partitions of the categories of the models and the students calculated using the match tables. Hi = High-overlap task, Lo = Low-overlap task, S1 = participant 1, S2 = participant 2, and so on.**

The bottom table in Figure 5 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 lower rate of return in their information search, and thus had fewer over tags (but more unique tags, see Figure 4). Apparently, mental categories in the low-overlap group tended to be more general than those in the high-overlap group, presumably because documents saved by participants in the low-overlap group had less overlap in the contents 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 similar number of mental categories as participants, but the partitions of the mental categories were also similar as participants, even though the inherent information structures were different between the two tasks.

## DISCUSSION

From our knowledge, the current study is the first that shows collaborative tagging systems not only can facilitate indexing of information, but can also facilitate 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. Indeed, many innovative ideas were generated by the sudden realization that knowledge structures in disjoint domains are

relevant. 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. Brunswik, E., *Perception and the representative design of psychological experiments*. 1956, berkeley: University of California Press.
2. Cattuto, C., V. Loreto, and L. Pietronero, *Semiotic dynamics and collaborative tagging*. Proceedings of National Academy of Sciences, 2007. **104**: p. 1461-1464.
3. Fu, W.-T., *The microstructures of social tagging: A rational model*, in *Proceedings of the ACM 2008 conference on Computer supported cooperative work*. 2008: San Diego, CA. p. 229-238.
4. Fu, W.-T., T.G. Kannampallil, and R. Kang, *A Semantic Imitation Model of Social Tag Choices*. , in *Proceedings of the IEEE conference on Social Computing*. 2009: Vancouver, BC. p. 66-72.
5. Fu, W.-T., et al., *Semantic Imitation in Social Tagging*. ACM Transactions on Computer-Human Interaction, 2010. **17**(3).
6. Golder, S.A. and B.A. Huberman, *Usage patterns of collaborative tagging systems*. J. Inf. Sci., 2006. **32**(2): p. 198-208.
7. Halpin, H. and H.S. Thompson, *Social Meaning on the Web: From Wittgenstein to Search Engines*. Intelligent Systems, IEEE, 2009. **24**(6): p. 27-31.
8. Marchionini, G., *Exploratory search: from finding to understanding*. Commun. ACM, 2006. **49**(4): p. 41-46.
9. Piaget, J., *The equilibration of cognitive structures*. 1975, Chicago, IL: The University of Chicago Press.