A Semantic Imitation Model of Social Tag Choices

Wai-Tat Fu, Thomas G. Kannampallil, and Ruogu Kang
Applied Cognitive Science Lab, Human Factors Division and Beckman Institute
University of Illinois at Urbana-Champaign
Urbana, IL 61801
wfu/tgk2@illinois.edu

Abstract—We describe a semantic imitation model of social tagging that integrates formal representations of semantics and a stochastic tag choice process to explain and predict emergent behavioral patterns. The model adopts a probabilistic topic model to separately represent external word-topic and internal word-concept relations. These representations are coupled with (1) a tag-based topic inference process that predicts how existing tags may influence the semantic interpretation of a document, and (2) a random utility model of tag choices based on the inferred topics and concepts extracted from the document. We show that the model is successful in explaining the stability in tag proportions over time, and different power-law frequency-rank distributions of tag co-occurrences for semantically general and narrow tags. The model also generates novel, testable predictions on how emergent behavioral patterns may change when users with different domain expertise interact with a social tagging system.

Keywords—*Social Tagging; Semantic Imitation; Random Utility Model; Social Behavior Modeling; Social Information Foraging; Cognitive-Social Modeling*

I. INTRODUCTION

Social tagging systems, such as del.icio.us (<http://del.icio.us>) and CiteUlike (<http://citeulike.org>), allow users to annotate, categorize and share their web contents using short textual labels called tags. The popularity of tagging arises from its benefits for supporting personal online search, ability to browse content using tags, and organization and sharing of tagged contents. The success of these social tagging systems has attracted much attention from researchers to study how and why these systems are useful to different populations of users for knowledge sharing and information search [4,5,7,9,10,15]. In contrast to the traditional keyword annotation system, a social tagging system provides users an unstructured mechanism for organizing and managing content. In fact, the major characteristics of tags are an open-vocabulary and non-hierarchical nature, and are usually created by users of the information documents rather than by professional annotators. The relative unstructuredness of tagged contents is also suggested as its potential weakness. Researchers have argued that the relative openness of the tagging systems may present a “vocabulary problem” for indexing [5], resulting in a large number of tags that are not meaningful to other users. However, research has shown that social tagging systems do tend to exhibit stable overall patterns over time, which has prompted researchers to investigate how these structural patterns emerge from multiple users in the system.

Many attempts have been made to formally characterize the mechanisms behind various emergent behavioral patterns in social tagging systems (e.g., [4,7,10]). In particular, many researchers are interested in the underlying semantic structures, often known as *folksonomies*, which emerge over time. Folksonomies are, however, formed very differently than traditional knowledge structures that one can find in dictionaries or encyclopedia. In essence, in contrast to global coordination by experts, folksonomies are formed by the collaborative effort of small-scale annotations by multiple users, who have very diverse knowledge backgrounds and information needs when they interact with a social tagging system. One major focus of the current paper is to simulate how different knowledge structures of multiple users may lead to different emergent structures in a social tagging system.

More generally, another purpose of the current paper is to show that much insight can be gained about the structures of social behavior from cognitive theories of individuals, such as those that aim at deriving formal representations of semantic knowledge structures and those that aim at characterizing human choice behavior. Our goal is to show that by imposing theory-based constraints on the cognitive processing of information at the individual level, one can develop strong predictions on emergent behavioral patterns at the social level. In fact, many have argued that this kind of cross-level modeling and analysis are essential in understanding large-scale multi-user system [23], as these levels often interact with each other and may not be easily analyzed separately, and that their functional characteristics are often intermingled without clear-cut boundaries.

In the rest of the paper, we will first review existing models of social tagging, focusing on those that aim at explaining user behavior at both the individual and social levels. We will then present the details of the semantic imitation model. We will then present simulation results of the model, and show how they could explain emergent behavioral patterns found by others, as well as how the model can lead to novel predictions. Implications of the model as well as future directions are then discussed.

II. PREVIOUS MODELS OF SOCIAL TAGGING

A. The stochastic Urn Model

In spite of the inherent unstructuredness in social tagging systems, researchers [4, 10] have found long-term stability in tag usage patterns. These stable usage patterns are important

because they provide partial validation and support to the usefulness of social tags in annotating information content and facilitating information search. For example, Golder and Huberman [10] used data from del.icio.us to argue that a users' tag choice was directly influenced by tags created by other users for the same document (in this case, a web page). They found that the proportions of tags assigned to a particular document tended to converge over time. They showed that the stochastic urn model by Eggenberger and Polya [6] was useful in explaining how a simple imitation behavior at the individual level could explain the aggregate converging usage patterns of tags. In particular, they showed that the convergence of tag choices could be simulated by a process in which a colored ball was randomly selected from an urn and was replaced in the urn along with an additional ball of the same color, simulating the probabilistic nature of tag reuse. This simple, probabilistic word-imitation model was shown to be able to produce the same patterns of convergence of tag proportions. The simple model, however, does not explain why certain tags would be "imitated" more often than others, and therefore does not provide a good account of tag choices at the individual level.

B. The Memory-Based Yule-Simon Model

The memory-based Yule-Simon (YS) model of Cattuto [4] attempted to explain tag choices by a stochastic process that assumed that the temporal order of tag assignments would have different impact on tag choices. Similar to the stochastic urn model, the YS model assumed that at each time step a tag would be randomly sampled: with probability p the sampled tag was new, and with probability $1-p$ the sampled tag was copied from existing tags. When copying, the probability of selecting a tag was assumed to decay with time, and this decay function was found to follow a power-law distribution. Thus, tags that were recently used had a higher probability of being reused than those used in the past.

One major finding by Cattuto et al. was that semantically general tags (e.g., "blog") tended to co-occur more frequently with other tags than semantically narrower tags (e.g., "ajax"), and this difference could be captured by the decay function of tag reuse in their YS model. Specifically, they found a slower decay parameter (such as the tag will be reused more often) explains why semantically general tags tended to co-occur with a larger set of tags. In other words, they argued that "semantic breadth" of a tag could be modeled by a memory decay function, which could lead to different emergent behavioral patterns in a tagging system. However, there was no explicit semantic representation of words in their model, and the probability of tag reuse was not directly related to any semantic representation of tags.

C. Semantic models

Results from previous models were based on analyses of word-word relations as revealed by the various statistical structures in the organization of tags (e.g., how likely one tag would co-occur with other tags or how likely each tag was reused over time). These models therefore have little to offer about the word-concept or concept-concept relations that exist in most knowledge systems. Indeed, one intriguing feature of social tagging systems is that they can be

considered platforms for dynamic interactions of diverse knowledge structures among users (e.g., [4,7,9,22]). It is therefore reasonable to assume that tag choices are influenced not only by tags at the word level, but also by the semantic interpretation of tags and their related information source. By assuming that semantic interpretation of tags will influence tag choices, one can broaden the analysis by going beyond the statistical structures at the word to include how semantic structures of the words from different users may contribute to the folksonomies in social tagging systems.

Although there are many existing semantic models [3, 11, 13], we focus on the probabilistic topic model [3] that has properties that allow us to separately model the external and internal knowledge structures, and to integrate it with a psychologically plausible choice mechanism. The topic model was originally used in areas of information retrieval, which assumes a hierarchical structure of probabilistic topical structures among words in documents. The topic model has demonstrated its successes with extracting latent topics in collections of documents [3] and social information systems [19], and seems to match the general characteristics of human semantic memory well [11]. We will discuss how it can be used to represent semantic structures in social tags next.

III. A SEMANTIC IMITATION MODEL OF TAG CHOICES

One major assumption of the model is that, when a user is navigating in a social tagging system, existing tags associated with Web documents will invoke a *tag-based topic inference*, such that the user can infer the topics contained in the Web documents based on the semantic interpretation of these tags. Based on the inferred topics, the user *chooses tags* to represent these topics. In other words, background knowledge structures of multiple users are dynamically "connected" through the interpretation and creation of tags in the social tagging system. We will describe the major components of the model next.

A. Knowledge representation

In the model, knowledge is represented by a set of concepts, each of them are instantiated by a set of words over a probability distribution. Specifically, if c represents the set of concepts, and w represents the set of words, then $p(w|c)$ will represent the probability distribution of words given a set of concepts. For any given set of words $\bar{w}$, we can then calculate the probability that these words represent a particular concept c_k using the Bayes's theorem:

$$p(c_k | \bar{w}) = \frac{p(\bar{w} | c_k) p(c_k)}{\sum_i p(\bar{w} | c_i) p(c_i)} \quad (1)$$

In (1), the summation is over all concepts. In the model, $p(w|c)$ will be represented by multinomial distributions, and $p(c)$ will be represented by Dirichlet distributions. Because multinomial and Dirichlet distributions form a conjugate pair, $p(c|w)$ will also follow a Dirichlet distribution.

To simulate differences in background knowledge, we assume that for each concept c , the prior probability for each word w in the multinomial distribution $p(w|c)$ will be

Figure 1. The semantic imitation model. Users interpret tags based on their background knowledge to infer topics contained in a document (tag-based topic inference) and based on these inferred topics, users choose tags to represent the latent semantics in the documents. It is assumed that there is a shared (but not identical) semantic knowledge backgrounds by the users in a system. The models shows that the match between the internal knowledge structures of users and the external knowledge structures contained in documents will determine stability of the folksonomies formed in a social tagging system.

normally distributed. In other words, for each concept, there will be words that are a priori more *central* to a concept than others. The normal distribution also implies that concepts overlap with each other, so there exists ambiguous words that belong to multiple concepts. As we will show later, the standard deviations of the prior probability distributions for $p(w|c)$ will be manipulated to reflect different background knowledge structures of the users. The assumption is that people may differ in terms of their knowledge in different domains, therefore they may differ in terms of their ability to predict different word-concept relations (inferring concepts based on words, or using words to represent a concept), as reflected by the prior distributions of concepts for each word in the knowledge representation of the user.

B. Tag-Based Topic Inference

As shown in Figure 1, the model assumes two major processes when the user tags a Web document: (1) tag-based topic inference and (2) tag choices. We assume that as the user is navigating on a social tagging system, tags created by others will help them interpret whether a bookmark is relevant to his or her information goals [7]. The set of tags assigned to a bookmark will act as retrieval cues for relevant concepts represented by these tags. This topic inference process assumes that tags will allow the user to predict the information content in the document ([7, 9]), as well as to provide some forms of semantic priming of related concepts when the user comprehends contents in the document [12]. For example, given the tag “health”, one may predict that the document may contain health-related information, which also acts as a semantic prime that facilitates the retrieval of related concepts such as “nutrition”, “diet”, or “exercise” when the user comprehends the document.

This role of semantic priming in the topic inference process is perhaps best illustrated by the experiments on “false memories” (e.g., [17, 18, 20, 21]). These studies show

that when people were asked to remember a list of semantically associated words that converged on a non-studied word, people tended to falsely remember the non-studied 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 non-studied word *needle* in the list. This kind of “memory illusion” is interpreted as evidence supporting the notion that as people process a list of words (similar to when they browse tags in a social tagging system), semantic representations for those words will be spontaneously activated. These semantic representations will then exert a top-down influence on future recalls.

Based on the knowledge representation of the model as shown in (1), the tag-based topic inference process can be nicely cast as the estimation of probabilities $p(c|t)$, where t represents the set of tags assigned to a Web document, and c is the set of concepts that are inferred based on the set of tags. In other words, the topic inference process (see Figure 1) can be simulated by (1), substituting $\bar{w}$ by t and calculating the probabilities $p(c_k|t)$ for all k (see (2) below).

To simulate the comprehension process, the model recalculates the probabilities of all concepts c_k using (1), substituting the prior distributions of all concepts $p(c_i)$ by the set of $p(c_k|t)$ calculated from the topic inference process, and the set of words $\bar{w}$ by the actual words in the document. The final set of concepts extracted from the document can then be represented by the set of posterior probabilities $p(c_k|d)$ extracted from the document d :

$$p(c_k|t) = \frac{p(t|c_k)p(c_k)}{\sum_i p(t|c_i)p(c_i)}, p(c_k|d) = \frac{p(\bar{w}|c_k)p(c_k|t)}{\sum_i p(\bar{w}|c_i)p(c_i|t)} \quad (2)$$

C. Tag Choices

Assuming that users will assign tags that best represent the concepts extracted from the document, the model predicts that tag choices will be semantically similar to existing tags. Note that in contrast to simple imitation models (e.g., [10]), which assume that users will probabilistically decide whether to create new tags or reuse existing tags, the current model assumes that existing tags will first activate semantic representations of topics (concepts) inferred from these tags, and the activated semantic representations will in turn influence tag choices. Semantic representations of tags and tag choices are therefore integrated in the current model, and a tag choice will involve both the word-concept and concept-word relations, rather than the word-word relations in previous model.

We first calculated the probabilities that a word will be chosen to best represent the document by the following equation:

$$p(w_{new}|d) = \sum_k p(w_{new}|c_k)p(c_k|d) \quad (3)$$

In (3), w_{new} represents a potential new tag drawn from the vocabulary of the user, and d represents the set of words and tags associated with the document. The probability $p(w_{new}|d)$ therefore represents how likely it is that the words and tags

associated with the document will predict the word w_{new} . This probability is calculated by summing over the products of the probabilities that a concept c_k will be inferred from the document ($p(c_k|d)$), and the probability that the new word w_{new} will be selected to represent the concept c_k ($p(w_{new}|c_k)$).

To directly couple the semantic representations with the tag choice process, we adopted the *random utility model* (RUM) in behavioral decision making [16] to simulate tag choice behavior. RUM (or some variations of it) has been used to simulate human choice process as a function of utilities of the existing alternatives [2, 8, 14], and has shown to be robust in characterizing the stochastic nature of human choices. The current model assumes that the utility of a tag is reflected by the *degree of representativeness* of the tag to the semantic contents of the document. To also take into account the uncertainty involved in the estimation of the utilities, we define the utility U_w of a word w as

$$U_w = p(w|d) + \sigma \quad (4)$$

in which σ is a random variable that follows the extreme values distribution, such that

$$p(\sigma < t) = \exp(-\exp(\frac{-t}{b})) \quad (5)$$

It can be shown that the probability that U_w is the maximum value among all words j in the vocabulary and can be expressed as

$$p(U_w > U_j \text{ for all } j) = \frac{\exp(U_w/2b)}{\sum_j \exp(U_j/2b)} \quad (6)$$

We assumed that a new tag would be assigned only when the the maximum utility tag was more representative than existing tags. Specifically, we used a threshold value h such that a new tag would be added only when

$$\frac{\max[U_w]}{\max[U_i]} > h \quad (7)$$

In (7), $\max[U_w]$ represents the maximum utility among all words, and $\max[U_i]$ represents the maximum utility among existing tags. *The model therefore assumes that people tag a document to help them to reconstruct the information content of the document in the future.* Note that although we only focus on the information value of tags, representativeness has also been shown to reflect the association between a cue (tag) and an item to be retrieved in memory [1,2]. This is perhaps more relevant when people assign personal tags to remind themselves what and where to re-find certain information in the future. The tag choice process should therefore be applicable in these “personal use” scenarios as well.

To summarize, the major difference between the current and previous models of social tagging is that the current model takes into account the dynamic couplings among the semantic contents of the web documents, background knowledge of the users, as well as the stochastic choice process involved in tag assignment. The model can therefore provide a richer explanation on the emergent structures of a

social tagging system based on *not only* the word-word relations as in previous models, but also the folksonomies formed by the diverse topic-word-concept relations as users interpret and select tags to annotate web documents. Indeed, by showing that features of social tagging systems can influence higher-level knowledge structures of users, one may argue that social tags not only can provide annotation to web contents, but they also have the potential to play an active role in facilitating exchange of knowledge structures among users [7, 9].

IV. MODEL SIMULATIONS

A. Overview

To test the basic properties of the model, we first generated a set of document documents with a random set of topics and words based on some assumptions of their distributions. We then show how the model can produce some of the signature emergent patterns identified by previous researchers, such as the convergence of the tag proportions across time [10], the power curves identified in the frequency-rank plots of tags [4]. We will then show how the model can predict different emergent patterns when users have different background knowledge structures interact with the same social tagging system over time.

B. Generation of Resource Documents

Following the generative topic model [3], we generated a set of 100 resource documents for the simulation. In each document, a set of topics were randomly sampled from a uniform distribution of 100 topics, and for each topic, a set of words were randomly sampled from a multinomial distribution of 5000 words. The prior probabilities for the multinomial distribution of words in each topic were normally distributed with a standard deviation of 1. The prior probabilities were set such that for each topic, the mean of the normal distribution of words at each topic would anchor at a word. As the probabilities decreased towards the tails of the normal distributions, words would more likely belong to multiple topics, representing the inherent ambiguities of word-concept relations in natural language. The sampling of topics and words would continue until there were 500 words in each of the 100 documents.

C. Simulating Tag Choices

When the simulation started, a document would be randomly selected, and because there was no tag assigned to the document initially, an “unbiased” topic inference process would be performed by calculating the probabilities $p(c_k|w)$ (for $k=1$ to 100 topics) for the set of words w in the document (see (1)). This set of $p(c_k|w)$ will then be used to calculate $p(w_i|d)$ for all words i in the vocabulary (for $i=1$ to 5000 words) by (3). This set of $p(w_i|d)$ will then be used to calculate the utilities of all possible words in the vocabulary by (4-5), and the one that has the highest utility would be selected as the tag to be assigned to the document (see (7)).

After a tag was assigned, in the next iteration, the assigned tag would invoke the topic inference process, which would semantically prime the later comprehension process. Specifically, in (1), each $p(c_k)$ would be substituted by

Figure 2. Top: Scatter-plots of tag proportions against model cycles by the model when there was high (left) and low (right) level of match between the word-topic distributions in the document and word-concept distributions in the background knowledge of the simulated user. Bottom: Scatter-plots of choice proportions of each tag assigned to a single-topic document simulated by the model when there was high (left) and low (right) level of match between the word-topic distributions in the document and the word-concept distributions in the background knowledge of the simulated user.

$p(c_k|w)$ obtained in the last iteration, and $\bar{w}$ would be substituted by the tag t . The values for $p(c_k|t)$ could then be updated for all k . This set of $p(c_k|t)$ would then be used as the prior distributions of concepts during comprehension (see (2)), such that $p(w_i|d)$ for all words i in the vocabulary could then be re-calculated. Similarly, the values of $p(w_i|d)$ would be used to calculate the utilities of all words, and the word that had the maximum utility would be selected and compared to the maximum utility of existing tags. If the ratio of these maxima exceeded the threshold parameter h (see (7)), the new word would be added as a tag to the document, otherwise no tag would be added. This sequence of processes would then repeat for the next iterations of tag assignments. We fixed h to be 1.0 and b to be 0.01 in all simulations.

D. Stable Patterns in Tag Proportions

One of the earliest emergent behavioral patterns in social tagging was identified by [10], who showed that for a given document, the proportions of tags assigned to it tended to converge over time. In other words, as the number of tags increased, each tag's frequency reached a fixed proportion of the total frequency of all assigned tags. The convergence was taken as evidence supporting the social nature of tags, in the sense that even though individual users have different personal preferences on the choice of words, consensus among users can be formed spontaneously without any top-down influence from the system.

The top panels of Figure 2 show the simulation results of the model based on a randomly generated set of 100 documents. Each data point in the figures represents the tag proportion (y-axis) of a particular tag at a particular cycle (x-axis) of the simulation. The figure shows that over time, the tag proportions flattened out. Even when new tags were added in each cycle, the overall tag proportions remained relatively fixed over time, leading to stable patterns of tags.

The major reason for the semantic imitation model to reach stability in tag proportions was the shared common semantic representations among users. The model assumes that tag choices are directly influenced by the extent to which new tags are representative of the concepts extracted from the document, and these extracted concepts are influenced by the semantic interpretation of existing tags. The common semantic representations of words and concepts among users will therefore naturally lead to coherence in semantic interpretation of tags, and thus, their tag choices over time.

To illustrate the importance of the role of semantic representations in the overall stability of the system, we created two sets of simulated users who differed in their background knowledge structures. First, we simulated users who had word-concept distributions that matched perfectly with the word-topic distributions in the documents. These simulated users could represent users who have strong domain expertise and therefore have developed highly

structured knowledge that are well adapted to the external knowledge structures represented across documents in that domain. On the other hand, novices in a domain are unlikely to have knowledge as well structured as the experts. We simulated novices by changing the spread of the prior distributions of words over concepts in their background knowledge. The wider spread in prior distributions implied that words were less accurate in predicting any particular concept (thus less effective topic inference), and, given a particular concept, there was a higher variance in the choice of words (tags) to represent the concept.

Results shown in the top left panel of Figure 2 were obtained from the simulated experts and those in the top right panel were obtained from the simulated novices¹. The results showed clearly that experts reached stability much faster than novices. The faster convergence could be explained by the fact that tags assigned to each document were more predictive of the topics contained in the document, and that experts were much better at extracting the correct concepts based on the “high quality” tags created by other experts. On the other hand, novices tended to create tags that were less representative of the concepts. Novices were also less effective in extracting the *optimal* set of topics (in the Bayesian sense) from the documents, and their choice of words to represent those concepts were more diverse. Tags were therefore less similar and thus convergence was slower than experts.

The bottom panels show the scatter-plots of the relative tag frequencies of one special document that we created to illustrate the difference. This special document contained a single topic, with the mean of the prior distribution of words over this single topic at word 300. As expected, for both experts and novices, tag proportions were highest around the most representative words. However, experts clearly had a much more focused vocabulary than novices, as shown by the wider spread of tag choices. In addition, novices seemed to have “misinterpreted” the topic and chose tags around word 800 (the initial choice of this tag was due to random noise) to represent the wrong topic, which led others to follow (the initial choice led to a second cluster of words around the “wrong” topic). Both the wrong interpretation of topics and the higher variance in word choices had contributed to the slower convergence for novices in the system. The model therefore predicts that systems that are often used by domain expertise (e.g., by academic researchers, as in CiteUlike) will likely converge faster and have more high quality tags than those that are designed for general users (e.g., Del.icio.us). This prediction is obviously subject to future verification.

The simulation results in Fig. 2 showed that the semantic imitation model was successful in explaining the same stability of tag proportions as found by others. The major contribution of the current model is that the prediction was based on a cognitively plausible tag choice mechanism that was coupled to the formal representations of semantic

knowledge that exist in both external documents and internal knowledge structures of the users. The results not only showed that the model was capable of offering a sophisticated explanation of the stabilization of tag proportions based on a cognitive model of individual users, but also showed that it can generate testable predictions of different emergent behavioral patterns in systems used by different user populations.

E. Semiotic Dynamics of Tag Choices

Cattuto et al [4] showed that by plotting the frequency of co-occurrences of tags against their frequency ranks, the obtained relations could be characterized by a power function. In addition, they found that the power functions differed for semantically narrow and general tags. Specifically, they found that for semantically general tags (e.g., common words such as “blog”), the lower-rank (more frequent) portion of the frequency-rank curve tended to be flatter than that found in semantically narrow tags (e.g., specific terms such as “ajax”).

Cattuto et al. used the memory-based Yule-Simon model to simulate the observed power curves in the frequency-rank plots, and found that by changing the decay parameter in their model (which controlled how likely the same tag will be reused over time), they could simulate the differential effects observed between the semantically general and narrow tags. They argued that semantically general tags tended to “stick” in the system longer, therefore the decay rate of these tags tended to be lower, thus explaining the effect. However, this explanation was obviously not directly based on the factor of semantic narrowness as hypothesized, and thus did not provide a direct explanation on the effects of different semantic representations of words on tag dynamics in social tagging systems.

To simulate semantically general words, we created a subset of words that had a wider spread (standard deviation equals 2 for semantically general words, twice as large as that for semantically narrow words) in their prior probabilities of belonging to different concepts (see Figure 3). In other words, we assume that the wider spread in the prior distribution defines the “semantic spread” of a word, or how semantically general the word is. Based on this definition, a word that is likely to be used to represent a wider range of concepts (e.g., “blog”) will be semantically more general than a word that is specific to a concept (e.g., “ajax”).

Figure 3. Representation of semantically narrow and general words by the prior distributions of words in each concepts. The wider the spread, the more semantically general is the word.

¹ We also simulated a mix of experts and novices and the results were similar to the current results. As expected, convergence rate was slower than pure experts but faster than pure novices.

Before the simulation, semantically general words were added to the topic distributions to create a set of 100 documents. Because the focus of the simulation was on the differences in tag dynamics between semantically general and narrow words, the same distributions were used to represent user knowledge (thus the simulated user's background knowledge had the same prior distributions of semantically general and narrow words as in the documents). We then performed the same simulation of tag choices for each of these 100 documents. Fig. 4 shows the frequency-rank log-log plots for tags that co-occur with a semantically general and a semantically narrow tag aggregated across the 100 documents. For both curves, the slopes of the low-rank tags were clearly flatter than those of the high-rank tags. As suggested by [4], the difference between the low-rank and high-rank tags is an important feature, as it clearly *deviates* from the typical emergent behavioral pattern that can be characterized by the Zipf's law [25].

Figure 4. Frequency-rank log-log plots for tags that co-occur with a semantically general and a semantically narrow tag obtained from the simulation of the semantic imitation model. For the sake of clarity, the curve for the semantically narrow tag is shifted down by 1. The best fitting straight lines for the first 20 most frequently co-occurring tags

The semantic imitation model provides a straightforward explanation for this difference between the low-rank and high-rank tags in the frequency-rank plots: The shared semantic representations of concepts by multiple users imply that the internal representation of the concepts contained in the document (the gist) will likely be similar, therefore tags generated based on the shared semantic representations tend to co-occur more often than those that are less semantically related to the shared semantic representations. Under this assumption, the flatter low-rank curves in Fig. 4 represent tags that were semantically similar and related to the gist of the documents, while those that are not semantically related to the gist (the high-rank tags) tend to follow the generalized Zipf's law (with the exponent in the power law between -1 and -2). As shown in Fig. 4, these emergent behavioral patterns were well captured by the semantic imitation model.

Another interesting pattern in Fig. 4 is that the curve of the semantically general tag has a flatter slope than those of the semantically narrow tag in both the low-rank and high-rank portions. The explanation provided by the model is again straightforward: Generic (semantically general) tags tend to co-occur more with other tags than specific (semantically narrow) tags because generic tags have a wider

spread in their prior distributions over different concepts. Because the semantic imitation model assumes that the tag choice process is sensitive to both the semantic interpretation of existing tags and the representativeness of the tag to the underlying concepts extracted from the document, a semantically general tag will likely invoke a wider range of concepts, and the tag will also be more likely to be selected to represent a wider set of concepts that are extracted out from different documents. Figure 5 shows an example of the tags that co-occurred with a semantically general (top) and a semantically narrow (bottom) tag from the simulation. One can clearly see that there are more tags that co-occur with the general tag than the specific tag.

Figure 5. Tags that co-occur with a semantically general (top) and semantically narrow tag (bottom). Columns represent tagging events in a cycle of tag choices to a document by the model. Rows represent the top ten more frequent tags selected by the model. A filled cell represents the presence of the tag in the corresponding row.

The current model explained the emergent behavioral patterns observed from the frequency-rank plots directly through properties in the semantic representations of words and concepts. In contrast, the memory-based Yule-Simon model explained the flatter low-rank curve of the semantically general tags by a lower decay parameter that controlled how likely a tag will be reused over time. Although semantic representations tend to exhibit the same power-law decay function in memory [1,2], we believe that the use of a memory decay function to represent the semantic breadth of a tag is less direct than the current formal representations of topic-word-concept relations in our model. In addition, we believe that the integration of the semantic representations and the stochastic tag choice process can lead to a wider set of testable predictions of emergent tagging behavior in systems that have different combinations of user profiles and information contents.

V. GENERAL DISCUSSION

Although a significant amount of work has been done to develop models of social tagging, the link between cognitively plausible mechanisms and emergent social tagging behavior has seldom been the focus of research. Instead, the focus has primarily been on using tag occurrence patterns to infer potential user behavior during tagging. As a result, the underlying cognitive mechanisms behind social tagging are still not well understood. We believe that cognitive models at the individual level can provide a more realistic basis for understanding emergent behavioral patterns by imposing theory-based constraints, representations, and processes of individual cognitive agents in their interactions with the social tagging system. Indeed, this kind of cross-level modeling has shown much success in multiple domains [23, 24], and it seems useful to researchers in the domain of social computing.

In the current semantic imitation model, we showed that by integrating theory-based cognitive representations of semantics and a stochastic human choice mechanism, we are able to generate a range of emergent behavioral patterns observed in social tagging systems and provide testable predictions of behavior when there are differences in user knowledge structures. In our simulation, we assumed that users shared a common set of probabilistic concept-word relations, and simulate how different users (in each cycle) interpret existing tags and assign their own tags when they collaboratively annotate the documents. Because the model is developed at the individual level, simulating different mixes of individual representations and mechanisms will be relatively straightforward: Different models could be constructed and to interact with the environment (documents) in each cycle, and the aggregate behavioral patterns could then be observed. In terms of model prediction, the current approach is therefore much more flexible than a single model developed at the social level.

It is worth mentioning that although our model provides good match to the data, our main purpose is to illustrate how an integrated model based on separate external and internal semantic representations and an individual stochastic choice process can lead to novel predictions at the social level, we do not intend to argue that the model is exclusive of other generative models developed at separate levels. In fact, we believe that there is much to learn from these single-level models to provide a coherent picture of the dynamics that cut across multiple levels of human activities.

The assumptions of the word-concept-document distributions in the current model were based on previous research on semantic representations, and although they have been tested, they are of course subject to future improvements. For example, we plan to use the model simulate how polysemous tags are used in different context to understand their impact in social information systems. In fact, research in various fields, such as in cognitive science, has been providing insights on refining these knowledge representations of users. This again highlights the value of integrating research from multiple areas in understanding behavioral and computational constraints in social systems.

The current semantic representations, although probabilistic, do not change during the interactions. Our previous studies found that mental concepts do incrementally adapt to external knowledge structures as users interact with a social tagging system [7, 9]. We are currently working on incorporating the learning mechanisms into the model to simulate how different learning cognitive models may predict to different behavioral patterns, how they may facilitate exploratory learning through the system, and what would be the optimal settings that facilitate exchange of ideas by groups of users with different knowledge backgrounds. These simulations will lead to useful practical guidelines of the development of social information systems.

REFERENCES

- [1] Anderson, J.R., *The Adaptive Character of Thought*. 1990, Hillsdale, NJ: Erlbaum.
- [2] Anderson, J.R., D. Bothell, M.D. Byrne, S. Douglass, C. Lebiere, and Y. Qin, An Integrated Theory of the Mind. *Psychological Review*, (2004), 111, 1036-1060.
- [3] Blei, D., A. Ng, and M. Jordan, Latent Dirichlet Allocation. *Journal of Machine Learning Research*, (2003), 3, 993-1022.
- [4] Cattuto, C., V. Loreto, and L. Pietronero, Semiotic Dynamics and Collaborative Tagging. *Proceedings of National Academy of Sciences*, (2007), 104, 1461-1464.
- [5] Chi, E.H. and T. Mytkowicz, Understanding the Efficiency of Social Tagging Systems Using Information Theory, in *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*. 2008, ACM: Pittsburgh, PA, USA.
- [6] Eggenberger, F. and G. Polya, *Über Die Statistik Verketteter Vorgänge*. *Zeit. Angew. Math. Mech*, (1923), 1, 279-289.
- [7] Fu, W.-T., The Microstructures of Social Tagging: A Rational Model, in *Proceedings of the ACM 2008 conference on Computer Supported Cooperative Work*. 2008, ACM: San Diego, CA, USA.
- [8] Fu, W.-T. and J.R. Anderson, From Recurrent Choice to Skilled Learning: A Reinforcement Learning Model. *Journal of Experimental Psychology: General*, (2006), 135, 2, 184-206.
- [9] Fu, W.-T. and T.G. Kannampallil, 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, Editors. In Press, IGI Global: Hershey, PA.
- [10] Golder, S.A. and B.A. Huberman, Usage Patterns of Collaborative Tagging Systems. *J. Inf. Sci.*, (2006), 32, 2, 198-208.
- [11] Griffiths, T.L., M. Steyvers, and J.B.T. Tenenbaum, Topics in Semantic Representation. *Psychological Review*, (2007), 114, 2, 211-244.
- [12] Kintsch, W., *Comprehension: A Paradigm for Cognition*. 1998, New York: Cambridge University Press.
- [13] Landauer, T.K. and S.T. Dumais, A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, (1997), 104, 211-240.
- [14] Luce, R.D., *Individual Choice Behavior: A Theoretical Analysis*. (1959).
- [15] Marlow, C., M. Naaman, D. Boyd, and Davis, M., Ht06, Tagging Paper, Taxonomy, Flickr, Academic Article, to Read, in *Proceedings of the seventeenth conference on Hypertext and hypermedia*. 2006, ACM: Odense, Denmark.
- [16] McFadden, D., Conditional Logit Analysis of Qualitative Choice Behavior, in *Frontiers of Econometrics*, P. Zarembka, Editor. 1974, Academic Press: New York. p. 105-142.
- [17] Pirolli, P., The Infoclass Model: Conceptual Richness and Inter-Person Conceptual Consensus About Information Collections. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, (2004), 11, 197-213.
- [18] Pirolli, P., Rational Analyses of Information Foraging on the Web. *Cognitive Science*, (2005), 29, 343-373.
- [19] Pirolli, P. An Elementary Social Information Foraging Model. in *ACM SIGCHI 2009*. (2009). Boston, MA.
- [20] Pirolli, P. and S.K. Card, Information Foraging. *Psychological Review*, (1999), 106, 643-675.
- [21] Roediger, H.L. and K.B. McDermott, Creating False Memories: Remembering Words Not Presented in Lists. *Journal of experimental psychology: Learning, Memory, and Cognition*, (1995), 21, 803-814.
- [22] Steels, L., Experiments on the Emergence of Human Communication. *Trends in Cognitive Sciences*, (2006), 10, 8, 347.
- [23] Sun, R., ed. *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*. 2006, Cambridge University Press.
- [24] Wilkinson, D. Strong Regularities in Online Peer Production. (2008).
- [25] Zipf, G., *Human Behaviour and the Principle of Least Effort*. 1949, Cambridge, MA: Addison-Wesley.