# The Microstructures of Social Tagging: A Rational Model

Wai-Tat Fu
Applied Cognitive Science Lab
Human Factors Division and Beckman Institute
University of Illinois, Urbana, IL 61807
wfu@illinois.edu

**ABSTRACT**
This article presents a rational model developed under the distributed cognition framework that explains how social tags influence knowledge acquisition and adaptation in exploratory ill-defined information tasks. The model provides integrated predictions on the interactions among link selections, use and creation of tags, and the formation of mental categories. The model shows that the quality of tags not only influences search efficiency, but also the quality of mental categories formed during exploratory search. In addition, the model shows that aggregate regularities can be explained by microstructures of behavior that emerged from the adaptive assimilation of concepts and categories of multiple users through the social tagging system. The model has important implications on how collaborative systems could influence higher-level cognitive activities.

**Author Keywords**
Exploratory search, tagging, categorization, rational model

**ACM Classification Keywords**
H5.4. Information interfaces and presentation (e.g., HCI): Hypertext/Hypermedia. I.2.8 Problem solving, control methods, and search. J.4 Social and behavioral sciences

**INTRODUCTION**
Social tagging systems (STS) such as del.icio.us have attracted researchers from different areas to study their various characteristics. A major function of social tagging systems is to help people share and explore information contributed by other members of the system. The critical feature is the assignment of *tags* to a bookmark of a web page. These tags are "social" because once assigned, other members can see these bookmarks and the associated tags to get a general idea of the contents. Members of social tagging systems can therefore utilize these social tags to navigate, re-find information, or explore unfamiliar topics.

Previous research has shown the existence of regularities in STS [6,8] when data are aggregated across a large number of users over a certain period of time. One interesting aggregate pattern is that the frequency-rank distributions of tags tend to follow a power law, resembling those found in natural language and written text [28]. Another pattern is that as the number of bookmarks to a particular web page increases, the relative frequency of a tag's use converges to a constant. In other words, although initially users may assign different tags to a web document, as more and more users bookmark the same page, the long-term probability that any particular tag will be assigned to the same web document tends to stabilize. These patterns generally support the *social nature* of tag uses, i.e., the use of tags tends to be influenced by the tags created by others [24]. In other words, these social tags not only can function as some forms of metadata to web contents, but they may also play an active role in exchanging the knowledge structures of users through repeated interactions, a process analogous to that in the development of schemas and knowledge adaptation, such as concept assimilation and accommodation [17]. To a certain extent, these results suggest that STS can be considered shared external knowledge structures that allow exchange and assimilation of conceptual structures through the interpretation and creation of tags to web documents.

The current article adopts a user-centered focus to understand how the observed aggregate usage patterns could be related to the interactions between internal and external representations of knowledge structures at the individual level. I choose to call these interactions *microstructures* of social tagging. To identify these microstructures, detailed protocols were collected from participants over an extended period of time as they performed an exploratory task using a STS (Del.icio.us). A rational model is developed at the individual level to keep track of the incremental changes in both internal (concepts) and external (tags) representations of the participants during the task. Given that the goal of the model is to capture the emergent microstructures of behavior at the functional level over time (as opposed to, e.g., predicting what tags are most popular), and the rational assumption has been well tested to show low sensitivity to individual differences, depth of data is preferred over breadth. The microgenetic approach (e.g., [25]) is therefore chosen over population sampling to collect repeated samples from 4 participants over a period of 8 weeks to test how well the model can capture the dynamics in the interactions over time.


*CSCW'08,* November 8–12, 2008, San Diego, California, USA.

## STS AS DISTRIBUTED COGNITIVE SYSTEMS

STS are excellent examples of distributed cognitive systems (DCS) [9, 10, 27]. In contrast to the traditional definition of cognition, a DCS encompasses all flow of information among individuals and the resources in the environment. The idea is that the functional unit of analysis of behavior in a DCS should include all elements that bring themselves into coordination to accomplish some tasks, and any isolated analysis of its parts is insufficient to understand how the system works. A classic example is the demonstration of distributed memory systems in the cockpit by Hutchins [10], who showed that the encoding and retrieval of critical information by pilots rely on various displays inside the cockpit as much as individual memory. In addition, information from the external environment provides more than simply a cue to internal memories, but provides opportunities to reorganize the internal and external representations in the DCS.

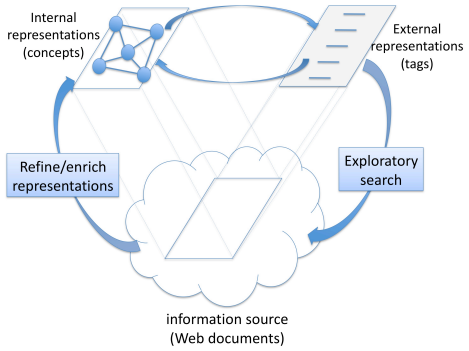


Figure 1. A distributed knowledge representations framework in a social tagging system during exploratory search.

Under this DCS framework, the current analysis of STS will focus on the intricate interactions between internal and external representations of concepts, tags, and documents as a user is engaged in an iterative exploratory search-and-comprehend cycle. Figure **1** shows a notational diagram of this theoretical framework. During exploratory search, the user may have only a rough idea what he or she is looking for, and the social tags created by others can be utilized as useful cues to select and navigate to the documents pertaining to the topic of interest [12,26]. Through this process of exploratory search, the user gains a better understanding of the topic through the enrichment of internal representations of concepts relevant to the topic (often called sensemaking activities, e.g., [23]). The user may then create their own tags for the web documents based on their own understanding as well as the existing social tags, and may choose to perform another cycle of exploratory search, refinement of concepts, and so on. The major characteristics of this DCS 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 the web document may influence both the internal (concepts) and external representations (tags).

One important characteristic of exploratory search is that the user may not know exactly what is needed, and may involve weak predictions about something new based on generalization of previously encountered concepts – such as predicting whether a particular bookmark/link will lead to something useful. As the user gains knowledge about types of documents related to certain concepts, the internal representations of these clusters of concepts are *refined and enriched*, allowing better judgment of the relevance of tagged bookmarks. By observing tags created by others, this enrichment process allows the user to assimilate concepts and ideas that emerge in the social tagging system to their own knowledge structures. *In other words, through the iterative explore-and-comprehend cycles, the interactions between internal concepts and external tags gradually lead to sharing and assimilation of conceptual structures as more and more people assign social tags to represent ideas or concepts that they extract from the massive amount of web documents.* This arguably is a major strength for the development of most Web 2.0 technologies. Surprisingly, little is known about how these new technologies may directly interact with individuals at the knowledge and cognitive level. It is hoped that the current analysis will provide some insight into how the DCS framework is useful in capturing the emergent behavioral patterns from these Human-Web2.0 interactions.

## HUMAN CATEGORIZATION

The goal of this paper is to provide a formal model of exploratory search in a social tagging system in the DCS framework, and show that regularities observed at the aggregate level may be explained by the inherent nature of the cognitive mechanisms that are responsible for humans to learn, assimilate, and create new concepts. I will focus on the concept formation process during exploratory search. A core mechanism for human concept formation is the process called categorization. Categorization is justified by the observation that objects tend to cluster in terms of their attributes in the environment (possibly because of the inherent process of natural speciation). Thus, if one can establish that an object is in a category, one is in a position to predict other unobserved features of the object. For example, if one knows that an animal can fly, it is likely that it is a bird, has feathers etc. From an adaptive point of view, mental categories can be interpreted as our internal representations of the structures of the environment (clusters of features and objects), which allow us to better predict features of new objects. The basic claim is that mental categories exist because categories exist in the external world (e.g., because of natural speciation), and our internal representations of these categories allow us to exploit the inherent correlations of features or similarities within that category [14, 15].

One may argue that tagging is really about assigning labels to categories of web documents, and not necessarily about category formation. Although there may be an interesting theoretical distinction between the two, for the current

purpose this distinction is not very important. *In fact, the current model will treat a category label just as another feature of the category.* The notion is that so long as category labels (tags) correlate with other features of the objects (contents of web documents) within that category (and they are supposed to), then they should promote the process of categorization: The more that can be predicted by category membership, the more advantage there is to creating such a category. Hence, category labels (tags) can simply be treated as features of the category.

The current rational assumption is that people will naturally categorize web documents as they go through and comprehend them (with the tags helping by adding addition features), and *the reason why mental categories are formed is that this is an adaptive (rational) response to the inherent structure of the stimuli from the external world to our minds that allow humans to predict features of new objects better*. Tags assigned to documents are just another set of features that allow us to predict the unobserved contents of the documents, and with the formation of mental categories, the tags will not only inform the user what they literally refer to, but also *other unobserved features* of the documents.

## A RATIONAL MODEL OF SOCIAL TAGGING

The model is based on the rational analysis framework [1, 16, 19], which has been used to explain a wide range of human behavior, including various memory effects, categorization, and problem solving. In a nutshell, the rational analysis is a reverse engineering approach to derive the possible mechanisms underlying an evolving system. The major assumption is that the mechanisms are optimally adapted to fulfill some functions in the environment. Recently, the same approach has been successfully applied to understand information-seeking behavior [4,7,11,19,20]. The current model adopts the same underlying assumption to explain exploratory, ill-defined information-seeking behavior. The major challenge is that in exploratory search, users do not know exactly what they are looking for, and both the internal representations and information-seeking behavior may change during the iterative search-and-comprehend processes [21,22,23]. As Figure 2 shows, there are three major classes of activities in the model: (1) Enrichment of concepts and mental categories: When a document is processed, the tags (T) and semantic nodes (S) are extracted; (2) Assignment/creation of tags to documents based on current set of mental categories; (3) Tag-based exploratory search: Evaluation of each link/bookmark by interpreting its tags based on its current set of mental categories, and selection of a link/bookmark.

### Enrichment of Concepts and Mental Categories

An ideal candidate to formally characterize the network of mental concepts, or schemas, is the spreading activation network, which has shown much success in characterizing human memory and conceptual network [1,2,19]. The theory of spreading activation represents concepts by a network of interconnected semantic nodes, with the activation value of each semantic node represents its likelihood of being retrieved, and the strength of association between nodes represent how strongly two semantic nodes are related to each other. Given a set of source nodes, the spreading activation mechanism predicts how likely certain representations (sets of semantic nodes) will be active. Refinement of representation occurs as experiences change the strengths of connections between semantic nodes.

When the user finished reading a web document, a set of semantic concepts were extracted. These semantic concepts, together with the social tags (S and T respectively in Figure 2), acted as input to the spreading activation network (the circled "1" in Figure 2). *Note that once in the spreading activation network S and T would be treated equally as semantic nodes.* In the actual implementation, semantic nodes were manually extracted from participants' verbal utterances as they were processing each document. All stop words and common words not related to the task were excluded during the extraction. For example, if a participant said, "This is a page about the history of Kosovo", then only "history" and "Kovoso" were extracted as semantic concepts (e.g., $S_n$, $S_m$ in document $D_i$) from the page. All existing tags as well as newly created tags were also extracted and input to the spreading activation network. For example, figure 2 shows that tags $T_j$ and $T_k$ were associated with $D_i$, therefore the set of tags and semantic nodes associated with $D_i=\{T_j,T_k,S_n,S_m\}$ will be fed to the spreading activation network.

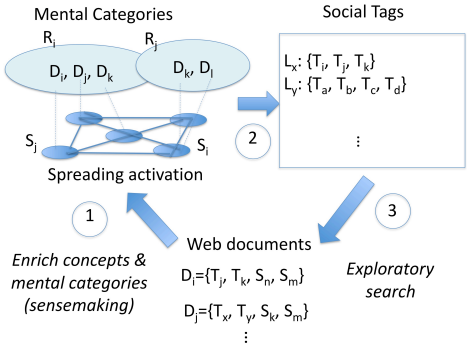


Figure 2. The main structures of the rational model. D=document, T=tags, S=semantic nodes, L=link/bookmark, R=mental categories.

#### *Rational Updating in the Spreading Activation Network*

When a document is fed to the network, it will first check if any of the semantic nodes already exists in the network. If a node is new, it will be added to the network. If it already exists, it will be merged to the old one, and its activation will be updated according to the activation equations (Eqn1 and 2 in Appendix). *The equations will increase activations of concepts that are recently or frequently encountered, assuming that recency and frequency reflect the need probability of the information in the future*. This assumption is adaptive (rational) in the sense that it aims at providing the *reason* why memory for items that are recently or frequently encountered is stronger is that they

tend to help us retain the most useful information in our memory system.

In addition to changes in activation of individual nodes, the network also updates the association between two nodes. The assumption is that if two semantic nodes co-occur in the same document, their association should be stronger. This change of association strength is updated by the associative learning equation (Eqn3 in appendix). Similar to activation updating of individual nodes, the strength of association reflects the log likelihood that two concepts co-occur in a document. Note that this is similar to models that utilize statistical language approaches (e.g., [5,7,11,13]) to calculate semantic relatedness from large text corpuses. However, in the current model, we assume that the user is naïve to the concepts to be found. In contrast, models that rely on statistical language approaches assume that the user's knowledge structure mimics the statistics of word co-occurrences, which obviously is less appropriate for exploratory search. Besides, statistical language approaches cannot handle learning, so the current association updating equation could also be considered an extension to models that rely on statistical language methods.

Although the spreading activation network provides precise characterization of the strength of the semantic nodes and their associations, it cannot explain the formation of abstract concepts or mental categories by generalizing across semantic nodes. Under the adaptive assumption, mental categories are useful because they allow predictions of unobserved features of a new document. For example, when the user finds that a bookmark/link is associated with a set of tags, the tags may allow the user to estimate how likely the document belongs to a certain category of documents that he or she has encountered before, thus allowing the user to predict what other information may be found in the document. I will first formalize how this estimation can be done, before I describe how each component in the analysis can be derived and how mental categories can be formed and refined.

*Predicting information content in the DCS framework*

Assume that a user has a set of mental categories $R$ and a set of semantic nodes $S$. The information goal is to predict whether node $S_j$ (some useful information) can be found by following a link with tags $T$, i.e., the user is trying to estimate this probability: $P(S_j|R,T_k)$ when deciding on links, which can be broken down into two components based on the DCS framework (see also [1,19]):

Predict internal rep from external rep

Predict information from a given mental (internal) category

$$P(S_j \mid R,T) = \sum_m P(R_m \mid T) P(S_j \mid R_m)$$

(Eqn4: Likelihood of finding information $S_j$ given R, T)

In other words, to predict whether node $S_j$ can be found in a particular document, one can first estimate $P(R_m|T)$: the probability that the document with tags $T$ belongs to a particular category $R_m$ 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 categories the document belongs. The second estimate $P(S_j|R_m)$ is the probability that $S_j$ can be found in mental categories $R_m$. This estimate therefore depends on the "richness" of the mental categories, i.e., the richer the mental categories, the better is the model able to predict whether the information can be found in the category $R_m$. The overall probability can then be estimated by enumerating the product of these two probabilities over all mental categories.

*Enrichment of mental categories*

Both components in Eqn4 will be updated in each explore-and-refine cycle as shown in Figure 2. First, the match between internal and external representations is improved:

$$P(R_m \mid T) = \frac{P(R_m)P(T \mid R_m)}{\sum_m P(R_m)P(T \mid R_m)}$$

(Eqn5: Prob. that a document with tags T belongs to $R_m$)

and each mental category is refined as $P(T|R_m) = n/n_m$,

(Eqn6: Probability that a tag belongs to mental category m)

where $n$ is the number of documents in category $m$ that contains $T$, $n_m$ is the number of documents in category $R_m$. Because $T$ and $S$ are treated equally as semantic nodes, the same equation can be used for $P(S_j|R_m)$ by replacing $S_j$ with $T$. The remaining variable to be estimated is $P(R_m)$: the prior probability for a document to belong to an existing category $m$. The prior probability $P(R_m)$ can be estimated by first assuming that there exists a prior probability $c$ for any two random objects to belong to the same category in a particular ecology (see [1,18]), and then estimate how likely the documents tend to *overlap in their information contents*. The higher the value of c, the lower the likelihood that any given document will belong to a new category. The exact equation for $P(R_m)$ can be found in the appendix (Eqn 7). For the current purpose, $c$ represents the *coupling probability* that any two documents belong to the same category for a particular information task. The value of $c$ therefore depends on the general structures of the information distribution. A higher $c$ value reflects a higher overlap of information contents across documents for a given information task (see Figure 3).

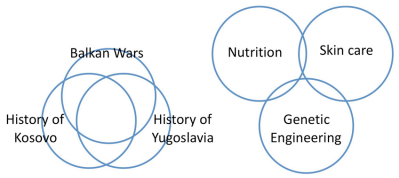


Figure 3. Notational diagrams showing high (left) and low (right) overlap of information contents across the two example sets of documents. Each circle represents a set of documents relevant to a particular concept. Note that these partitions depend on the information goal.

*Formation of new mental category*

To simulate naïve users, the model assumed minimal initial set of mental categories (a single category consisting of concepts from the task description). As the model processed a Web document, the set of semantic nodes was extracted based on the verbal protocols collected from the participant. The set of semantic nodes, together with the set of tags associated with the document, represented the *contents* of the document to be processed by the model (see step 1 in Figure 2). At this point, the model decided whether to assign this document to an existing category or to create a new category for this document. This decision was based on the value of *max[P(R|S)],* where *S* represents the contents (set of semantic nodes and tags) of the document and *R* represents the set mental category, and the max operation is performed in the set R (i.e., finding the value of P(R|S) for the most likely category). Specifically, a new category will be created only if

$$P(R_{new}) > \max[P(R|S)] \quad \text{(Eqn8: New mental category)}$$

i.e., the probability that the document belongs to a new category (from Eqn7) is larger than that for the category that has the highest value of *P(R|S)* among all existing categories. If this condition is not met, the document will be put into the category with the highest value of *P(R|S)*. The mental categories formed by the model could then be compared with the results of the categorization task given to the participants at the last session.

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

Given an existing tag $T_k$, the model will calculate the value of $P(T_k|R_m)$, where $R_m$ is the category to which the current document is assigned according to Eqn8. The model will assign this tag $T_k$ to this document only if

$$P(T_k|R_m) > \tau_{threshold} \quad \text{(Eqn9: Assign an existing tag)}$$

where $\tau_{threshold}$ is a free parameter to be estimated from the data. A new tag is created only if any of the semantic nodes associated with the documents in category $R_m$ is larger than the maximum of $P(T|R_m)$ for all existing tags, i.e.,

$$P(T_{new}|R_m) > P(T_{max}|R_m) \quad \text{(Eqn10: Assign a new tag)}$$

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. *Note the model does not predict which particular tags will be used, it only predicts how likely existing tags will or will not be used based on the relationship between the tags and the predicted mental categories formed.*

**Evaluating and selecting a link/bookmark**

The model assumes that the evaluation of a link/bookmark will be based on both the activation of the semantic nodes represented by the link text and the Bayesian estimate of finding $S_j$ given the link text and the tags associated with the link. Specifically, if *U(L)* represents the utility of a link *L* (the *goodness* of *L*), then

$$U(L) = \sum_i \sum_m \sum_k A_i * P(S_i | R_m, T_k)$$

(Eqn11: Evaluation and selection of link/bookmark)

where the summation is over all tags and link text ($k$) in link $L$, all mental categories ($m$), and all semantic nodes in the spreading activation network. $A_i$ is the activation of semantic node $i$ (Eqn1). Eqn11 implies that nodes that are more active (as calculated by the spreading activation mechanism) will be given more weight during link selection. This implies that the model will predict that topics that are studied more recently or frequently will more likely be included as part of the information goal in the next cycle of exploratory search.

## TESTING THE MODEL

To test the model's predictions, a set of exploratory information tasks was chosen. In all tasks, participants started with a rough description of the topic and gradually acquire knowledge about the topic through an iterative search-and-comprehend cycles. Participants were told to imagine that they had to write a paper and to give a talk on the given topic to a diverse audience who may ask all kinds of questions related to the topic.

After pilot studies, two general topics were chosen: (1) "Find out relevant facts about the Independence of Kosovo" (IK task), and (2) "Find out relevant facts about Anti-aging" (AA task). These two tasks were chosen because the IK task referred to a specific event, and therefore 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 AA task, on the other hand, was more ambiguous and was related to many disjoint areas such as cosmetics, nutrition, or genetic engineering. Web sites relevant to the IK task have more overlapping concepts than those relevant to the AA task. The other characteristic is that because the AA task was more general, the tags tended to be more generic (such as "beauty", "health"); in contrast, for the IK task, tags tended to be more "semantically narrow" (such as "Kosovo"), and thus dad higher cue validity than generic tags.

**Participants**

4 participants were recruited from the University of Illinois. Participants were undergraduate students and all had extensive experience with general information search and the del.icio.us Web site. Participants were randomly split and assigned to one of the tasks. From their self-reports they were all unfamiliar with the given topics. Participants were told that they should explore all relevant information to comprehend the topic using either the search function in del.icio.us or any other Web search engines, and they should create tags for Web pages they found relevant to the topic and stored them in their own del.icio.us accounts. Participants were told that these tags should be created for two major purposes. First, these tags should allow them to re-find the information quickly in the future; second, these

tags should allow their colleagues to utilize the relevant information easily in the future.

### Procedures

Each participant performed the exploratory information task for eight 30-minute sessions over a period of 8 weeks, with each session approximately one week apart. Participants 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. Participants 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. These tags and verbal descriptions were then input as contents of the document.

As an example, assume that the participant is processing a page that has two tags: "Kosovo" and "independence". Assume that this participant gave this verbal description after reading this document: "This page discusses reasons for the independence of Kosovo". The analysis will first start with excluding all stop words and task-irrelevant words from the description, yielding "discusses reasons independence Kosovo". It will then perform stemming on the words, yielding "discuss reason independ Kosovo". Because the words "independ" and "Kosovo" overlaps with the existing tags, the set of semantic nodes associated with this document D={discus reason independ Kosovo}. These will be fed to the existing spreading activation network (Eqn 1-3) as well as the mental categories refinement (Eqn 5-7), as shown in Figure 2.

One week after the last session, participants were asked to come back to perform a sorting 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 categories formed by the participants were then matched to those predicted by the rational model (i.e., the set of *R* in Figure 2).

## RESULTS

Participants on average created 88.5 bookmarks (IK1=93, IK2=84) and 379.5 tags (IK1=392, IK2=367) for the IK task, and 58 bookmarks (AA1=52, AA2=64) and 245 tags (AA1=256, AA2=234) for the AA task. Participants in the IK task created more bookmarks and assigned more tags than those in the AA task, but the average number of tags per bookmark is about the same (4.3 tags per bookmark) for the two tasks. As expected, finding relevant information for the AA task is more difficult, as reflected by the fewer number of bookmarks created. Given that distribution of information was more disjoint in the AA task (e.g., there is little overlap of information between web sites on skin care and genetic engineering), the results were consistent with the assumption that the average rate of return of relevant information was lower for the AA task then the IK task.

Three sets of measures were extracted from the data and compared with the main predictions of the model: (1) Tag use, (2) link/bookmark selection, and (3) formation of mental categories.

### Use of Unique Tags

The left panel of Figure **4** shows the cumulative number of unique tags assigned across the 8 sessions by each of the four participants. As expected, the number of unique tags approached asymptotes across sessions, and they were higher for the IK task than the AA task. This again was likely due to the lower rate of information return in the AA task. The right panel of Figure **4** shows the simulation results for each individual model. All pages bookmarked by the participants were used as input to the model; therefore each individual model had the same experiences as the corresponding participants. The model fit the data well ($R^2$=0.95), as confirmed by the almost identical trends of growth for each participant. The mismatches occurred mostly in the first sessions, in which the model under-predicted the assignment of tags. This was likely 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. The value of $\tau_{threshold}$ was set to be the same value (0.2) for all participants. $\tau_{threshold}$ was originally intended to be a free parameter to control the different levels of willingness for people to assign tags, but in this task all four participants were equally motivated to assign tags so we set this value to be the same. The lower number of tag assignments by the model in the AA task simply followed from the fact that there were fewer bookmarks created.

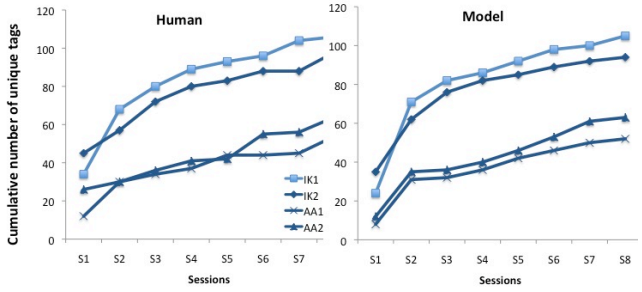


Figure 4. The number unique tag assignment across the 8 sessions by participants (left) and models (right).

Figure 5 shows the proportion of new tags created by the participants (left) and model (right). Perhaps the most

interesting pattern was that even though participants assigned fewer tags, but the *proportions* of new tag creation over total number of tag assignment were higher in the AA task than in the IK task. This was consistent with the lower rate of return of relevant information in the AA task, and this lower rate could be caused by fact that the existing tags on del.icio.us was less informative for the AA task. Indeed, concepts extracted from the documents by the participants in the AA task were more often different from the existing tags (and had a higher $P(T|R)$ value, see Eqn10) than in the IK 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 ($R^2$=0.75). Again, the major mismatches were found in the first sessions, where the model tended to under-predict the creation of new tags, especially for the IK task. A random model that randomly assign 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 are significantly above the chance level.

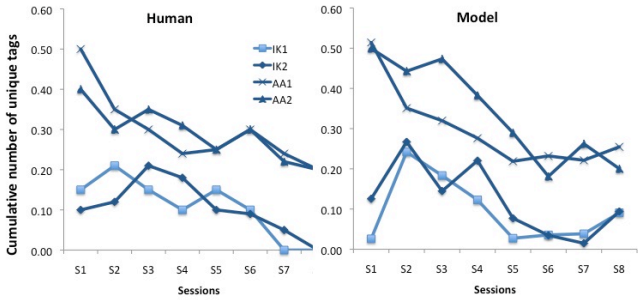


Figure 5. Proportions of new tags created over total number of tags assigned across the 8 sessions by the participants (left) and models (right).

### Link/Bookmark Selections

Participants made on average 357 link selections throughout the sessions (IK1=242, IK2=331, AA1=454, AA2=401). Figure 6 shows the frequency-rank plots matching the link selections by the participants to the model. The plots were first created by extracting all pages in which participants clicked on one of the links. The values of U(L) (see Eqn7) were then calculated for all links on each page, and the link that had the highest value of U(L) on a particular page would be given a rank of 1, the second highest a rank of 2, and so on (see [7]). The ranks of the actual links selected by the participants could then be found. If the model were perfectly correct, then all links selected by the participants would have had a rank of 1. The frequency-rank plots in Figure 6 could then be obtained by plotting the frequency distributions of the ranks of the selected links.

In general, one can see that links selected by the participants had low ranks (higher on the right side than the left side of each freq-rank plots), indicating that the model had done a good job predicting link selections by the participants. Simple t-tests conducted on the slope of the best fit lines obtained by regressing the frequencies to the ranks showed that they were all significant ($p<0.05$), indicating that the predictions were significantly above chance for all four participants. The results showed that the rational model was successful in capturing how the incremental changes in internal concepts and mental categories influenced the use of social tags during exploratory search.

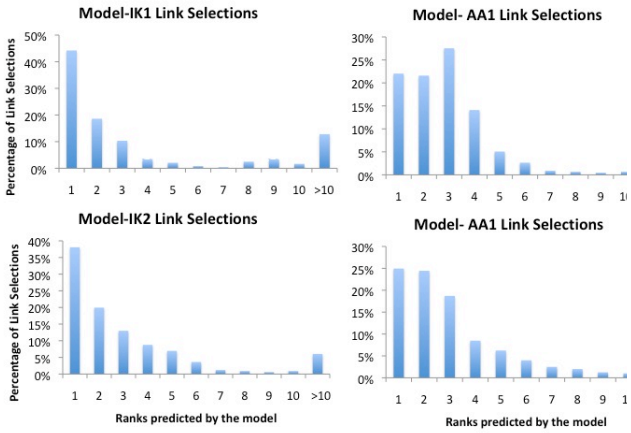


Figure 6. Frequency-rank distribution of the link selected by the participants and ranked by the model.

### Formation of mental categories

One core assumption of the rational model was that the assignment of tags and the selection of links were both based on the set of mental categories formed from observing existing tags assigned to documents that they processed. It is therefore critical to verify that the set of mental categories formed by the model match 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 category will be given a value 1, otherwise a 0. Table **2** shows two possible categorizations for {a,b,c,d,e} and their corresponding match tables. In this case the correlation is r=0.102. Similar correlations were calculated between the participant and model.

| | {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 |

Table 1. Example match tables for different categorizations.

The major determining variable for mental category formation in the model is the value of the coupling parameter, $c$ (see Eqn6). This was set to 0.6 for the IK task and 0.3 for the AA task to best fit the data. Because the information distributions are more disjoint for the AA task, the value of $c$ was set to a smaller value to reflect this property in the model. Table **2** shows the number of

categories formed by each participants and model, as well as their correlations. As predicted, participants formed more categories in the AA task, reflecting the structures of the information sources. However, as shown earlier, participants in the AA task had lower rate of return in their information search, suggesting that they spent more time looking for relevant information. Although the number of categories formed was higher in the AA task, the quality of these categories (in terms of how much they help in finding information) was lower than those in the IK task (results shown next). The correlations between the participants and the models were high in both tasks, suggesting that the model roughly formed similar mental categories as participants, even though the inherent information structures were different between the two tasks.

| | #categories (Human) | #categories (Model) | r |
|---|---|---|---|
| IK1 | 6 | 6 | 0.71 |
| IK2 | 5 | 6 | 0.68 |
| AA1 | 12 | 13 | 0.59 |
| AA2 | 10 | 11 | 0.67 |

Table 2. Number of categories formed by each participant and model, and the correlations (r) between the partitions of categories by human and model calculated using the match tables.

## EXPLAINING AGGREGATE PATTERNS: FROM INDVIDUAL TO SOCIAL BEHAVIORAL PATTERNS

Finally, one important goal of the model is to see if the microstructures of individual behavior, as identified by the rational model, could explain the emergence of aggregate behavioral patterns in STS as identified by others [6,8]. *The idea is that, similar to how the relation between pressure and temperature characterized by the ideal gas law can be explained by the kinetic theory of gas molecules, patterns of aggregate behavior in STS may be explained by the cognitive mechanisms derived based on the rational principle at the individual level.* If so, the current model may bridge the gap between cognitive theory that explains individual behavior and the aggregate behavioral patterns observed in collaborative systems.

### Power relationship in freq-rank distributions of tag use

One aggregate pattern was identified by [6], who showed that the frequency-rank distributions of tag use on del.icio.us follow a power function with negative exponents. In addition, frequencies of semantically narrower tags (such as "ajax") tend to drop faster than that of generic tags (such as "computer"). One explanation is that semantically narrow tags tend to have higher cue validity, i.e., they tend to convey more information about the contents of the document. To test whether the rational model can predict such patterns, the values of *P(T|R)* (the conditional probability that external tag T belongs to mental category R) were pooled across all mental categories from participants in each task. This value reflects the likelihood that a tag will be used for a given category of documents. Therefore a more informative tag should have a higher value of *P(T|R)*.

The log-log plots of *P(T|R)* against ranks for both tasks were shown in Figure **7**. The slopes of the best-fit lines for the IK and AA tasks were -1.76 and -1.35 respectively, and the fits were good ($R^2$=0.86 and $R^2$=0.81 for the IK and AA tasks respectively). The ranges of the exponents were also consistent with the Zipf's law [28], which show that, similar to those found by Cattuto et al. [6], the frequency-rank distributions of the tags were similar to those found in natural language utterances. This shows that the model, although derived based on individual behavior, shows the same aggregate patterns as identified by Cattuto et al.

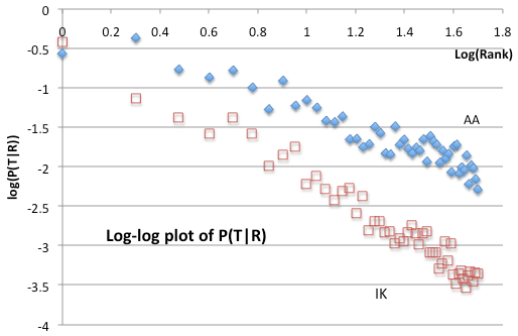


Figure 7. The frequency-rank distributions of P(T|R).

In addition, consistent with Cattuto et al. [6], the more generic tags in the AA tasks showed a flatter curve (the value of *P(T|R)* dropped slower with rank). This finding, together with the lower number of bookmarks and tags created in the AA task, suggested that the social tags were less useful for the AA task then the IK task. Apparently participants were not successful in assimilating to the information structures based on the poorer social tags. To a certain extent, one can say that there was a lower level of conceptual coherence achieved by interacting with the STS in the AA task then in the IK task.

### Convergence of proportions of tag use

Another aggregate pattern was identified by Golder and Huberman [8]. They found that for a particular web document, as the number of bookmarks by multiple users increases, the proportion of use of any particular tag to this web document tend to converge to a constant. Golder and Huberman suggested that one possible reason for the convergence is that people tend to *imitate* others' behavior by using the same tags.

The rational model provides a more sophisticated explanation to the convergence. The selection of tags is governed by Eqn 9-10, which basically select that tag that has the maximum value of *P(T|R)*. Note the value of *P(T|R)* is updated by Eqn 6, which depends on the existing tags. Instead of explaining the convergence by the process of imitation, the model assumes that *existing tags directly influence the categorization of the document, which in turn influence the selection of tags*. The model therefore predicts that, if users share the same mental category, their use of tags should be similar, thus the convergence. Most importantly, the model predicts that the likelihood of assigning different tags will be higher *if* the user categorizes the document to differently, such as when the user extracts

different semantic concepts from the document because the information goal is different. This is a prediction derived from the model that can directly tested in the future.

## GENERAL DISCUSSION

The rational model was developed under the assumption that humans categorize web documents as they explore the Internet and represent knowledge extracted from web documents internally as mental categories. The processes conform to the adaptive principle, which states that the goal of categorization is to improve our ability to predict new objects by exploiting structures of the environment. The model, developed under the DCS framework, was successful in providing good quantitative predictions on the emergent behavior of four different individuals across an extended period of time. The model shows how internal representations slowly assimilate to the external informational distribution through the processing and assignment of social tags, and how individuals create new tags based on their internal representations. The dynamic interactions between internal and external representations captured by the model has also highlighted the value of the DCS framework, as they imply that isolated analysis of either the distributions of external tags or cognitive mechanisms of the user will unlikely lead to good characterizations of the dynamics that emerge from socio-technological systems.

When social tags convey useful information for formation of "good" mental categories, exploratory search performance was better, mental categories contain more semantically distinct concepts (as shown by the steeper frequency-rank relationship). On the other hand, when social tags were generic, mental categories formed were less useful, and search performance suffered. *The results imply that higher-level cognitive behavior, such as knowledge acquisition and adaption, or decision making that depends on the formation of mental categories, could be directly influenced by social tags.* Future research would provide more direct evidence showing the impact of social tags on other higher-level cognitive activities.

The current results also show that social tagging systems have the potential to facilitate not only collaborative indexing of the massive amount of information, but also as a means for social exchange of knowledge structures, and thus has the potential to promote collaborative activities that involve higher level cognitive processing, such as problem solving, decision making, or creative designs. For example, the formal analysis of the current distributed cognitive system can be implemented as software tool that facilitates extraction and exchange of mental categories for different groups of people who have different expertise in different domains. Can tags, for example, be organized by mental categories extracted from experts in different fields in ways that facilitate knowledge transfer? Will transfer or exchange of knowledge at the fact, concept, and category levels facilitate innovation because they encourage re-structuring of existing knowledge structures? 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.

Although the model was developed to keep track of individual performance, the outcomes were shown to be useful in understanding aggregate usage patterns found by others. This shows the potential of explaining emergent social behavior through a set of relatively stable distributed cognitive processes at the individual levels. Not only does this kind of multi-level analyses provide direct implications on designs, but may also provide a better understanding and prediction on how and when changes in individual behavioral patterns could be related to changes in interfaces or different social or cultural norms.


## ACKNOWLEDGMENTS

This research is supported by a Seed Grant from the Center for Healthy Minds, funded through the National Institutes of Health/National Institute on Aging under Award No. P30-AG023101, as well as funding from the Human Factors Division and Beckman Institute of the University of Illinois. The author thanks Thomas George Kannampallil and 4 anonymous reviewers for comments on a previous version of the paper.


## REFERENCES


1. Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Lawrence Erlbaum Associates.
2. Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y . (2004). An integrated theory of the mind. *Psychological Review 111,* 1036-1060.
3. Anderson, J. & Schooler, L. (1991). Reflections of the environment in memory. *Psyc Science, 2*, 396–408.
4. Blackmon, M. H., Kitajima, M., & Polson, P. G. (2005). Tool for accurately predicting website navigation problems, non-problems, problem severity, and effectiveness of repairs. *Proc. of CHI 2005*.
5. Budiu, R.; Royer, C.; Pirolli, P. L. (2007). Modeling information scent: a comparison of LSA, PMI and GLSA similarity measures on common tests and corpora. In *Proc. of 8th RIAO Conference*.
6. Cattuto, C., Loreto, V., & Pietronero, L. (2007). Semiotic dynamics and collaborative tagging. *PNAS, 104*, 1461-1464.
7. Fu, W.-T., & Pirolli, P. (2007). SNIF-ACT: A cognitive model of user navigation on the World Wide Web. *Human-Computer Interaction, 22*, 355-412.
8. Golder & Huberman, (2006). Usage Patterns of collaborative tagging systems. *J. Info. Sci., 32*, 198-208.
9. Hollan, J., Hutchins, E., Kirsh, D. (2000). Distributed Cognition: Toward a New Foundation for Human-Computer Interaction Research, *ACM Transactions on Computer-Human Interaction, 7* (2), 174-196.

10. Hutchins, E. (1995). How a cockpit remembers its speed. *Cognitive Science, 19*, 265-288.
11. Kitajima, M., Blackmon, M., & Polson, P. (2005). Cognitive architecture for Website design and usability evaluation: Comprehension and information scent in performing by exploration (pp. 343–373). *HCII.*
12. Marchionini, G. (2006). Exploratory search: From finding to understanding. *Comm. of the ACM*, *49*, 41-46.
13. Matveeva, I., Levow, G., Farahat, A., & Royer, C. (2005). Terms representation with generalized latent semantic analysis. In Proc. ranlp Conference.
14. Medin, D. & Schafer, M. (1978). Context theory of classification learning. *Psy Review, 85*, 207-238.
15. Medin, D., Lynch, E., & Coley, J. (1997). Categorization and reasoning among tree experts: Do all road lead to Rome? *Cognitive Psychology, 32*, 49-96.
16. Oaksford, M. & Chater, N. (2006). *Bayesian Rationality*. Oxford: Oxford University Press.
17. Piaget, J. (1963, 2001). *The psychology of intelligence*. New York: Routledge.
18. Pirolli, P. (2004). The InfoCLASS model: conceptual richness and inter-person conceptual consensus about information collections. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, 11, 197-213.
19. Pirolli, P. (2005). Rational analyses of information foraging on the Web. *Cognitive Science, 29*, 343–373.
20. Pirolli, P., & Card, S. K. (1999). Information foraging. *Psychological Review, 106*, 643-675.
21. Qu, Y. & Furnas, G. (2005). Source of structure in sensemaking. In *Proc. of CHI*, 1989-1992.
22. Qu, Y. & Furnas, G. (2008). Model-driven formative evaluation of exploratory search: A study under a sensemaking framework.
23. Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993). The cost structure of sensemaking. *Proc. of ACM INTERCHI'93*, 269-276.
24. Sen, S., Lam, S., Rashid, A.,Cosley, D., Frankowski, D., Osterhouse, J., Harper, F., & Riedl, J. (2006). Tagging, communities, vocabulary, evolution. In *Proc. of CSCW*.
25. Siegler, R. S., & Crowley, K. (1991). The microgenetic method: A direct means for studying cognitive development. *American Psychologist, 46*, 606-620.
26. White, R., Kules, B., Drucker, S., & Schraefel, M. (2006). Supporting exploratory search: Introduction. *Communications of the ACM, 49*, 36-39.
27. Zhang, J. & Norman, D. (1994). Representations in distributed cognitive tasks. *Cognitive Science, 18*, 87-122.
28. Zipf, G. K. (1949) *Human Behavior and the Principle of Least-Effort.* Addison-Wesley.

**Appendix – Equations of the rational model**

From the ACT theory of human memory [2], the activation of any particular semantic node $i$ in the network is represented by $A_i$, which can be calculated as:

$$A_i = B_i + \sum_j W_j \sigma_{ji} \quad \text{(Eqn1 -- spreading activation equation)}$$

In the above equation, $B_i$ stands for the base-level activation of node $i$, $\sigma_{ji}$ stands for the strength between node $i$ and $j$, and $W_j$ stands for the weight of node $j$ in the association. $W_j$ is simply set to 1 if node $j$ is the current input to the network; otherwise it is set to 0. Base-level strength will increase according to this base-level learning equation:

$$\beta = \log(\sum_j t_j^{-d}) \quad \text{(Eqn2: base-level learning equation)}$$

where the summation is over all previous encounters for the concept, $t_j$ is the time since the $j$th encounter of the concept, and $d$ is a decay parameter which is typically set to 0.5. Note that the base-level learning equation succinctly captures the recency and frequency effects of memory in a single equation, and has shown to be an optimal solution to the information need as imposed by the statistical structures of the environment [3].

As $S_i$ and $S_j$ co-occur in the same document, the strength of association between these two nodes will be updated as:

$$\sigma_{ji} = log\ (a) – log\ (M) \quad \text{(Eqn3: associative learning equation)}$$

where $a$ represents the number of times concept $j$ and $i$ are observed in the same document, and $M$ represents the total number of nodes in the network. This equation implies that the more often two concepts co-occur, the more strongly they will be associated with each other, but this association will be "diluted" as the size of the network grows [2]. The overall effect is that after a set of semantic nodes are activated from the web documents (the source nodes), the activation will spread to other nodes that are connected to this set of source nodes (as reflected by the value of $\sigma_{ji}$).

The prior probabilities for a concept to belong to category $R_m$ and or to a new category can be estimated as:

$$P(R_m) = \frac{cn_m}{(1-c)+cN}, P(R_{new}) = \frac{1-c}{(1-c)+cN}$$

(Eqn7: Prior probability of mental categories)

where $c$ is the coupling probability, which represents the probability that any two documents will belong to the same group. $n_m$ is the number of documents in category $m$ as before, and $N$ is the total number of documents in all categories. As $N$ increases, the likelihood that an object comes from a new category decreases. As $N$ increases the value of $P(R_m)$ approaches $n_m/N$, which implies that without any information cue, "popular" categories tend to be favored over "unpopular" categories, which in general is consistent with previous findings (e.g., [6,8]).