

Performance comparison between tagging and categorization as personal information management strategies

Abstract

The present paper examined performance of collaborative tagging in a micro scope. Through a preliminary think-aloud test and a controlled experiment, performances of tagging and categorization in information organization and retrieval under personal information management context were compared. The experimental result shows that tagging imposes higher mental workload than categorization in both information organization and retrieval. This result is contrary to previous dominating viewpoint as well as our hypotheses. A detailed discussion of cognitive activities and paths involved in tagging and categorization was presented to review and modify the current model. Several characteristics contribute to higher workload in tagging include the mental engagement needed to generate context-dependent tags, the need to filter and modify tags, and heavier memory usage in tagging.

Keyword: collaborative tagging, categorization, mental workload, folksonomy

1. Introduction

It is common to organize information by its metadata, also called keywords or tags, to enhance future retrieval. Keywords have a long history in academia and traditionally, such organization is performed either by authorities or authors themselves. However recently the term “collaborative tagging” was brought into popularity by a number of online social applications, such as del.icio.us[1], flickr[2], and CiteULike[3]. Take del.icio.us as an example. It states in its help page that: “On del.icio.us, you can use tags to organize and remember your bookmarks, which is a much more flexible system than folders.”

In collaborative tagging, the collection of a specific user’s tags forms his or her personomy, or say *personal taxonomy*. And personomies of all users in the system merge into *folksonomy*, a blend of the words ‘folk’ and ‘taxonomy’. [4] Compare to traditional taxonomy systems that are imposed by experts, not by users, folksonomy lacks exclusiveness and hierarchy: an object can belong to many tags at the same time and all tags are equal. Some claim that folksonomy provides more accurate, truthful and democratic information infrastructure than the assigned by a single authority [5][6], while some others attacks that its

uncontrolled, inconsistent, and context-dependent vocabulary damages the information retrieval quality. [7] Nevertheless, tagging spreads quickly into different domains, like photo sharing (flickr.com), academic publication reference (CityUlike), and even the latest Microsoft Windows Vista and Mac OS X.

The popularity of tagging calls for in-depth analysis of its behavior and characteristics. Although a considerable amount of researches have been conducted regarding tagging usage and user behavior, these researches focus on macro and holistic issues of complex tagging systems, such as tag distribution[6][7][8][9], long-term stabilization[8][10], and user vocabulary.[11][12][13] While these researches provided valuable insights of how a collaborative tagging system function and behave as a whole, little attention has been put on the micro aspect of tagging, which may include cognitive analysis of tagging process, performance of tagging as an information organization and retrieval method, or its related usability consideration. From the perspective of methodology, previous studies mainly used log analysis to handle large amount of data generated by crawler or API of certain websites, while few use controlled experiment to obtain behavioral and qualitative data.

In an exploratory blog post, Sinha presented a comparative analysis of the cognitive process between tagging and categorization [14]. She argues that tagging does not require users to choose the best category, as the exclusiveness of categorization requires. Her conclusion that tagging demands less mental workload than categorization was accepted by many researchers [15][16], without further justification.

In this paper we compared performance of information organization and retrieval tasks in tagging and categorization system. This paper consists of two studies. The former one was a think-aloud test with six Chinese taggers followed by interview. The latter one was a controlled experiment to compare performance of tagging and categorization in information organization and retrieval tasks. The dependent variables include mental workload, satisfaction, and learning effect. The experimental result is contrary to sinha's conclusion, revealing lower mental workload in information organization tasks using categorization. In-depth discussion is provided to establish a cognitive model of tagging.

2. Literature Review

2.1. Characteristics of tagging as folksonomy

The name *folksonomy* arises as a result of increasingly application of collaborative tagging, it is a system of classification generated from the mass action of the users, a classification scheme for resources, and may challenge the role of established *taxonomies*, or say *ontology*.

The collection of a specific user's tags forms his or her own personomy, and the collection of all personomies in a collaborative tagging system constitutes the folksonomy of the system, thus the name "folksonomy" come from "folks' taxonomy", which indicates its essence: non-authoritative, personal, and collaborative. [17]

Familiar taxonomies include the Linnaean system of classifying living things, the Dewey Decimal Classification (DDC) for libraries, and computer file systems for organizing electronic files. Taxonomy systems are hierarchical and exclusive, in above systems, each animal, book, file and so on, is in one unambiguous category which is in turn a sub-category of a superior category. There are several good reasons to maintain such a hierarchical structure. An efficiently organized hierarchy neatly and unambiguously bounds a category's contents. And unlike a keyword-based search, wherein the seeker cannot be sure that a query has returned all relevant items, a hierarchy assures the seeker that all the contents it contains are in one stable place. At the same time, it is context-independent. A perfect taxonomy system should not change for different time, place, and people. That is, a book on Africa's geography should be in the Dewey Decimal system category 916, no matter it is in a US library or a German one, categorized by librarian Tom or Lucy.

Folksonomy lacks such hierarchy and exclusiveness. An object can belong to many tags at the same time and all tags are equal. The flat tagging system could associate an article with a great variety of attributes simultaneously. Certainly disagreement does exist. One could tag a picture "cat" while others may tag "animals", "lovely", or "to_own_one". And folksonomy is context-dependent for sure, since it directly reflects the different vocabularies and previous knowledge and experiences of each tagger. [18]

Even the tagging system for personal information management shows these two characteristics which bring difference in both information organization and retrieval tasks. [14], and the retrieval performance of tagging system is challenged by the inconsistency of vocabulary used. However, at the same time the non-exclusiveness of tagging system allows the existence of very personal and flexible set of contents, e.g. the tag "cool" and "toread". This flexibility makes it possible to customize personal organization patterns or special groups.

Some claim that folksonomy is more suitable for common internet users and their communities, and the collaborative characteristics makes tagging a democratic approach to merge human knowledge [5]. This statement needs further justification.

2.2. Tagging behavior

Tagging behavior is complicated. Although often defined as a user generated metadata to describe an object, the actually usages of tags vary greatly given the large user number. Despite the most common descriptive and taxonomic tags, a significant phenomenon is that tags are often context-related, which is of great difference with traditional categorization methods. For example, the tags "cool", "toread" and "tobuy". Another situation is the use of compound words, such as "Develop/C++", "Develop/Java". Some extreme use cases include using tag "*", "**", to "*****", to give rating about content quality.

There have been practices to category tags. Golder et al [8] has outlined 7 individual types (or functions) of tags, as:

- Identifying what it is about
- Identifying what it is
- Identifying who owns it
- Refining categories
- Identifying qualities or characteristics
- Self reference
- Task organization

[11] describes a frame of tags based on ad hoc categories. He divided the tags into two groups: taxonomical and ad hoc. The taxonomical group is similar to formal taxonomy, like “politics”, “animal”, “hotel”, and the ad hoc categories are like “things to sell tomorrow”, “paper readed”. Based on this dichotomy, he identified 8 major types of tags and categorizations:

- Taxonomic (parrot/flower/tree)
- Functional (toy/vehicle/weapon)
- Functional collocation (furniture/cutlery/clothing)
- Reason collocation (leftovers/refreshments/groceries)
- Function+origin (vegetable/cereal/medicine/herb)
- Adjective (cool/social)
- Verb (search/tagging/todo)
- Proper name (New York)

Although different classification existed, we may take a holistic view to these dimensions, and consider a more generic classification. Here we classified tags into three types:

- **Identification**

This category includes the first four types as well as self reference type in Golder’s model and first five types as well as proper name type in Csaba’s model. Generally it is used to identify what the object is, to name it, to clarify basic characteristics or describe an obvious fact. These tags are generally nouns, like taxonomic: “cat”, “animal”, or exact name: “NYC”, or basic abstract concepts: “love”. These tags are relatively consistent and context-independent.

- **Description**

This category includes the “Identifying qualities or characteristics” type in Golder’s model, and “Adjective” type in Csaba’s model. It is used to describe the qualities or characteristics of objects, which is a context-dependent process and related to tagger’s personal opinion and previous knowledge. For example, one may tag a photo of dogs “lovely” while another may tag it “scary”.

- **Action**

This category includes the “task organization” type in Golder’s model, and “Verb” type in Csaba’s model. It is used to describe possible interactions between the user and the object. Therefore this type is highly context-dependent and personal. The description and action categories of tags together reflect ad hoc tags.

2.3. Tagging consistency

As various types of tag exist, and the process itself is highly personal and context-dependent, the consistency of tags becomes a heated issue in both practice and research. Here consistency includes both vocabulary and occurrence frequency. The consistency of tags can be classified into three different levels: personal consistency, means that a specific user labels consistent tags to similar content; collaborative consistency, means that individual resource should be labeled by the majority of users with a consistent set of tags; and inter-application consistency, means to maintain interoperability of tags among different applications. [19][20]

The major causes of inconsistencies include polysemy (one word has multiple meanings), synonymy (different words have the same meaning), and different level of abstraction (e.g. cat, animal, Persian cat) [9], misspellings, plurals, symbols, and compound words [7]. Although Shirky claims that no such things like as synonyms because users employ different tags for different specific reasons, the inconsistency vocabulary do degrade information retrieval performance. [5]

There have been various efforts to improve tagging consistency. Two possible ways to improve tagging consistency in a folksonomy have been proposed: educating users to make consistent tags, and to improve the system design to encourage adding consistent tags. [7] Regarding system design, some researchers believe information retrieval performance can be facilitated by a controlled vocabulary in tagging choice by matching the vocabulary of taggers and searchers.

While analyses of crawled data of large collaborative tagging systems revealed that tag distribution tends to stabilize into power law distributions. And they claim that given sufficient active users, over time a stable distribution with a limited number of stable tags and a much larger “long-tail” of more idiosyncratic tags develops [10]. This stabilization provides an optimistic expectation of collaborative consistency issue. Some believe this stabilization is due to the similarity of human fundamental mental architecture and shared knowledge [19], while some believe imitation plays a key role to form stabilization. [8]

2.4. Cognitive process of tagging and categorization

So far few studies have addressed the cognitive processes in tagging. At the same time, many just claimed that tagging has a lower cognitive cost compare to categorization. The main reason is that tagging is a free and unrestricted process. For example, [18] stated the

barriers of categorization are simply too high, using quote of [21] that “(tagging) It’s like 90% of the value of a proper taxonomy but 10 times simpler”. Wu et al state that tagging has “dramatically” lower costs without hierarchical structures, that “Users simply create and apply tags on the fly” [22]. And in a blog post, Sinha argues that tagging imposes a lower mental cost to users than categorization since it does not require the user to choose the “best” category, as the exclusiveness of categorization requires. She highlights the ability of tagging to avoid the condition of so-called ‘post activation analysis paralysis’. According to Sinha, such a condition places users in a state of cognitive paralysis and is triggered when users attempts to tag an information resource to ensure future retrieval. [14]

Two basic principles have been proposed for the formation of categories. The first one asserts that the task of category system is to provide maximum information with the least cognitive effort, and the second asserts that the perceived world comes as structured information rather than as arbitrary or unpredictable attributes, therefore maximum information with least cognitive effort is achieved if categories map the perceived world structure as closely as possible. [23]

Rosch et al stated that categories within taxonomies of concrete objects are structured such that there is generally one level of abstraction at which the most basic category cuts can be made. Basic objects are the first categorization of concrete objects made by children, and objects may be first seen or recognized as members of their basic category. Further, the most useful and thus most used name for an item is the basic level name.

3. Research questions and hypotheses

3.1. Study 1

This study focuses on user’s behavior, cognitive process and use cases in tagging, following research questions have been raised:

RO 1: Do users have a stable vocabulary or pattern for their tagging? Will they revise, add or delete their tags? Is there a hierarchy in everybody’s personal tag systems?

We hypothesize that expert users will form relatively stable vocabularies and acquire tagging patterns, while novice users may present higher inconsistency. And we also want to find out whether users modify tags to achieve better consistency. In addition, we want to test on the non-hierarchical characteristics in tagging practice.

RO 2: Is there any difference between users who tag for themselves and those who tag for others?

Since tagging may be used for personal information management or collaborative information management and exchange, we hypothesize that users will use more

personalized and casual words when tagging for themselves, and more generic wording when tagging for the public.

RQ 3: What's the user-perceived difference between categorization and tagging? And which do users prefer?

We want to directly compare these two information management tools. Whether user can perceive the non-exclusiveness and non-hierarchy properties of tagging? And how these differences influence users' preference?

3.2. Study 2

In this study, following hypotheses were formulated for experiment and analysis.

HYPOTHESIS 1: Tagging requires less mental workload than categorization does in information organization.

If the cognitive process behind tagging is really as Sinha described, the mental workload required from tagging users should be less than that from categorization users since tagging requires less cognitive activities. From a usability perspective, mental workload is an important factor of software and website efficiency, which is an important indicator of information system performance.

HYPOTHESIS 2: Information retrieval performance of categorization strategies is better than that of tagging strategy.

Given the presumption that categorization may require more cognitive engagement with the objects in information organization stage. And this cognitive engagement is spent to find the most favorable category for best retrievability. Then we would expect users can achieve higher information retrieval performance in hierarchical categorization systems. Higher performance may be indicated by lower mental workload and higher satisfaction level. Moreover, since the inconsistency of tagging vocabulary is often criticized to degrade re-findability of information, it is also plausible to expect worse information retrieval performance in a tagging system.

4. Methods

4.1. Study 1

This study was conducted using the Chinese version of MyWeb on Yahoo!China (<http://myweb.cn.yahoo.com/>), a social bookmarking service similar to del.icio.us.

Think-aloud technique was applied first to observe and analyze behavior and cognitive processes of tagging. [Think-aloud technique has been proven to be an effective method in

collecting data to evaluate one's mental strategies to accomplish tasks. It was first introduced by Ericsson & Simon[8].] Subjects were asked to tag 15 web pages which were stored on local server. Each page was generally a short article including news, political critics, digital devices and foods. We choose these common topics to minimize the influence of previous knowledge and learning. Subjects were asked to vocalize their thoughts, opinions and feelings while performing these tasks.

In the second step, we interviewed users with pre-structured questions. The questions were organized into three categories: 1) Comparison between online bookmarking service and traditional "favorites" option in browsers; 2) Strategies of tagging in social bookmarking service; 3) Objects, contexts, purpose and prospect of this service.

Finally, given the topic of each article, users were asked to recall their tags on each of them.

4.2. Study 2

4.2.1. Experiment Design

The experiment was a two-group design with information organization method as a between groups variable (tagging or categorization). We designed a two-phase experiment including both information organization and retrieval stage. In the information organization phase, subjects read a number of short articles and organized them by taxonomic categories or tags. In the information retrieval stage, subjects were presented with a series of questions, and asked to find the corresponding article for each question.

Short articles were used as objects of this study. Although many major tagging services are about photo (Flickr), URL (Del.icio.us) and Video (Youtube), textual material can provide greatest clarity and be easily retrieved and edited to fit the required style, length, and topic in our controlled experiment. All articles were written in Chinese describing an item, with a clear title, in an encyclopedic style and not longer than 500 characters.

An important standard of choosing material topics was to provide least ambiguity among different subjects, with little reliance to previous knowledge. Western Classic music was used as the main topic of test texts, considering generally most Chinese people have little knowledge on classic music, but generally they have no negative attitude to it. In the training session, we used a series of articles regarding basic computer components. The purpose is that participants will have some background knowledge on basic computer concepts, and based on our daily experience, this topic is natural and common.

Test materials were selected to contain latent hierarchical structures. For example, some of them were about musicians in different countries and eras. Which provide possibility for user to establish different categorization schemes among them.

4.2.2. Dependent variables

To assess mental workload in the experiment, we deployed time estimation and NASA-TLX scale. Time estimation is a secondary task technique to assess mental workload, it does not use the same sensory-motor pathways as a large range of operators' tasks so it is not inclined to interfere with operators' tasks. [24][25][26] In our experiment, subjects were asked to reproduce a 10-second interval by stepping down a mounted pedal.

NASA-TLX scale is a common used subjective measure of mental workload. [27]. It uses six dimensions to assess mental workload, namely mental demand, physical demand, temporal demand, performance, effort, and frustration. A paper and pencil version of the NASA-TLX scale was used, and the scale has been translated into Chinese.

The numbers of clicks that users performed to finish information retrieval tasks were recorded as an indication of information retrieval performance.

A modified version of the satisfaction measure utilized by Cook [28] was used to assess the satisfaction score of each information organization and retrieval task.

A questionnaire was used to assess learning effect of these two information organization methods. The questionnaire consists of 19 multiple-choice questions, each regarding a fact in an article.

4.2.3. Experiment interface

Two screen captures of experiment interfaces, categorization and tagging, are presented below correspondingly. Articles are shown on the left side of interfaces, user's current information structure is shown to the right, and no initial structure was provided.

In tagging interface, an input field was provided, allowing user to input multiple tags separated with a space. User can also reuse previous tags by checking them in the information structure. Each tag was displayed as an expandable object and can be renamed or deleted.

In categorization interface, user can create, rename, delete, or create sub-categories freely. The number of possible category levels is not limited. The established categorization structures will be shown in tree view.

4.2.4. Experiment procedure

After signing consent form, subjects were first given a 1-minute calibration session of time estimation. The calibration was done with a colored block on screen switching its color every 10 seconds. Then subjects were asked to step down the pedal every 10 seconds, with no other concurrent tasks, for a period of 3 minutes.

Subsequently subjects were given a guided warm-up session, they were asked to use tagging or categorization method to organize 10 short articles which describe basic computer components, synchronously stepping down the pedal every 10 seconds. After

the warm-up session was the formal information organization session. Subjects were asked to tag or categorize 38 short articles about classical music, synchronously continuing time estimation. No time limit was given.

After the information organization session subjects were immediately given a NASA-TLX questionnaire and satisfaction questionnaire. They were asked to consider the information organization session only when finishing the questionnaires. This break up is also to avoid recency effect in following information retrieval session. In subsequent information retrieval session, subjects were presented 17 questions, one at a time. Each question is about a fact in one article. Subjects were asked to use the organization they just created to find which article contains the answer. Questions were designed to include the target article name directly. Subjects were instructed to avoid mis-clicking as precise as possible. The number of clicks of each subject was recorded. When subjects find the target article in their information structure, they were asked to double-click the article to enter the full text and click "next step" to next question.

After information retrieval session, subjects were again given a NASA-TLX questionnaire and satisfaction questionnaire, and instructed to consider information retrieval session only. Subsequently they were asked again to perform 3 minutes time estimation task with no concurrent activities. The learning effect questionnaire was given at last to finish the experiment.

5. Results

5.1. Study 1

6 subjects, all male and aged from 21 to 29, participated in our study. Their average computer usage is 8.5 years, and average online time is 6.67 hours per day. Four of them are considered to be expert users, with over 2 years experience of tagging. Other two subjects knew the concept of tagging but never used it in practice. All subjects have bachelor or above education

Key findings include:

- Users behave differ largely when they are tagging for themselves and tagging for others.

Only one subject reported that he would consider the sharing aspect of tagging. He claimed that he will choose more personal word when tagging for himself and general words when tagging for others. A summary of social and personal tags is presented below:

Social tags	Personal Tags
Brands	Description and Action

(Sony, Creative, Ipod)	(to buy, to read, used in bathroom)
Categories	Adjectives
(MP3, TV, Food, Mobile, Education, Travel, IT, etc.)	(funny, ridiculous, important, cute)
Characterization	Proper noun
(political, design, digital, social, etc.)	(desinification)
Sources	
(Sina, engadget, "author")	

- Tagging vocabularies vary largely between users, but there are some vocabularies win all users' favor.

Tags identifying the topic of bookmarked items proved to be overwhelmingly used, like digital, education, politics, etc. Following that, tags are frequently used to identify what kind of a thing the item is, like news, blog, or article. Descriptive tags of the item's characteristics are also commonly used, such as fun, interesting, etc. Although fun/cool rated unusual in some researches, they are used by all subjects in the experiment, even if some subjects are not using task related tags like "toread" or "tobuy". In addition, subjects reported the tag "fun" was also useful to retrieve information when they want to recommend something interesting to friends or entertain themselves.

- Few users try to manage their tags, leading to an unorganized tag system.

Our subjects all reported that they would keep the consistency of tags, but few behave it in the experiment. Tagging is free and uncontrolled. Though users said they always pay attention to the consistency of tags, few reported they would coordinate or manage their tags regularly. Plural words, misspelled words did occur frequently.

5.2. Study 2

5.2.1. Subjects

40 subjects (17 of them are female) participated in the experiment and were paid 50 RMB Yuan as remuneration. All subjects have used tagging before. A summary of other characteristics is shown in [table]. All subjects have college or above education level. These 40 subjects were assigned to either tagging or categorization group evenly. No significant difference in characteristics between these two groups was found.

Variables	Tagging (N=20)		Category (N=20)	
	Mean	SD	Mean	SD
Age	22.55	1.36	22.65	1.18
Computer Exp. (yrs)	8.3	2.77	8.1	2.27
Internet Exp. (yrs)	6.9	1.62	7.2	1.15

5.2.2. Mental workload

- *NASA TLX scale*

A summary of TLX scores in information organization and retrieval sessions is given in [table and]

Table: summary of TLX overall and subscale scores in information organization session

Subscales	Tagging (N=20)		Categorization (N=20)	
	Mean	Std Dev	Mean	Std Dev
TLX1 overall score	59.14	7.04	54.54	8.95
TLX1 mental	67.375	12.152	58.05	14.841
TLX1 physical	43.5	26.124	48.375	28.554
TLX1 temporal	44.5	16.674	44.625	16.766
TLX1 performance	45.275	12.277	41	18.503
TLX1 effect	71.15	11.408	72	11.169
TLX1 frustration	48.775	17.457	35.55	21.123

Table: summary of TLX overall and subscale scores in information retrieval session

Subscales	Tagging (N=20)		Categorization (N=20)	
	Mean	Std Dev	Mean	Std Dev
TLX2 overall score	47.39	14.00	42.66	13.81
TLX2 mental	52.575	20.768	48.450	19.068
TLX2 physical	36.650	22.706	32.550	20.725
TLX2 temporal	48.750	23.063	32.625	20.703
TLX2 performance	32.500	18.691	30.000	22.580
TLX2 effect	57.875	16.648	52.875	17.420
TLX2 frustration	33.250	17.054	29.050	22.104

Analysis of variance reveals a marginally significant difference on the overall score of NASA-TLX scale in information organization task ($F(1,39)=3.269$, $p=0.078$), and significant difference in mental demand ($F(1,39)=4.726$, $p<0.05$) and frustration ($F(1,39)=4.658$, $p<0.05$) subscales of the first NASA-TLX questionnaire, and in temporal demand ($F(1,39)=5.414$, $p<0.05$) subscale of the second NASA-TLX questionnaire in information retrieval tasks.

- *Time estimation data*

A summary of time estimation data is given in [table]. The recorded variables are lengths of time interval produced by subjects. The variables Base (1) and Base (2) denote the two 3-minute time estimation sessions without concurrent tasks. And Warmup, Formal, and Retrieval denote the average length of reproduced time intervals in warm-up, information organization, and information retrieval session, correspondingly.

Table: time estimation data summary

Variable	Tagging (N=20)				Categorization (N=20)			
	Grand Means	SD of Means	Max	Min	Grand Means	SD of Means	Max	Min

Base (1)	12.730	1.728	15.930	10.128	11.689	2.221	16.363	8.880
Warmup	16.713	6.914	34.071	10.513	15.691	6.889	33.769	7.791
Formal	17.140	6.945	33.235	10.743	17.103	8.480	42.341	8.570
Retrieval	13.584	4.454	28.760	8.234	12.798	5.803	28.361	7.824
Base (2)	12.183	1.960	16.180	9.349	11.982	2.973	18.958	6.995

Analysis of variance finds no significant difference of time estimation performance between tagging and categorization treatment in warmup, information organization, and information retrieval sessions. However, we identified a significant difference between the baseline performance (with no concurrent task) and working performance (with concurrent tasks)

Considering time estimation performance may largely depend on individual ability to perceive and estimate time duration, we performed ANCOVA using the baseline time estimation performance (the average of two baseline time estimation tasks) as a covariate. Result shows baseline time estimation significantly predicts the performance in information organization and retrieval tasks (all $p < 0.001$).

5.2.3. Information retrieval performance

We collected the number of clicks in information retrieval stage during the experiment. Note the double clicks required to access the correct article are not included. No significant difference was identified ($p = 0.807$)

Variable	Tagging (N=20)				Categorization (N=20)			
	Means	SD	Max	Min	Means	Means	Max	Min
clicks	28.6	10.97	61	19	27.85	6.45	45	19

5.2.4. Satisfaction and learning effect

A summary of satisfaction and learning effect is given in [table]. No significant difference between two treatments is found.

Table: satisfaction and learning effect score summary

Variables	Tagging (N=20)		Categorization (N=20)	
	Mean	SD	Mean	SD
Satisfaction 1	4.82	.51	5.00	.39
Satisfaction 2	5.15	.44	5.13	.32
Learning effect	11.5	2.82	10.2	3.16

6. Discussion

The present study examined possible difference of performance between tagging and categorization as two information organization and retrieval methods. Mental workload,

learning effect, and satisfaction were assessments of performance.

6.1. Mental workload

In information organization tasks, the overall TLX scores of tagging subjects are marginally significantly higher than those of categorization subjects and the mental demand and frustration subscale ratings of tagging subjects are significantly higher than categorization subjects. This difference indicates that the mental workload of tagging in information organization context is indeed higher than that of categorization, which, to our surprise first, is contrary to Sinha's conclusion as well as our hypothesis. This surprising result leads us to reconsider the cognitive processes behind categorization and tagging. In the following, we will argue that these two cognitive processes are different and more complicated than Sinha's model.

6.1.1. *A twofold concept activation process*

In Sinha's model, an implicit assumption is that given the same content, concepts activated in tagging and categorization processes should be the same. Only with them can the first stages of tagging and categorization become comparable.

An observational counterexample of this statement, however, is found in our experiment. Many categorization processes were completed very fast. Some users may just scan titles and topic sentences in articles, and then determine in seconds the suitable category (like "musician", "instrument", or "music history"). This happened frequently in the later stage of our experiments, when a stable hierarchical category structure has been established. However, while in tagging, users may determine the first tag very quickly (at most time also basic taxonomical tags like *musician* or *instrument*, however, they obviously slowed down when applying more tags.

Back to the theory of basic level name in categorization, two conclusions are interesting: objects may be first seen or recognized as members of their basic category, and the most useful and thus most used name for an item is the basic level name. These together indicate that the first generated concepts – basic level names – are very potential to be applied finally as the category.

It has been also spotted that in tagging system, statistically 5 [8] Also this pattern recurred in our experiment. Compare the earliest tags of each article to category names, high similarity was found. And category names are overwhelmingly taxonomical names such as "musician", "instrument", or "style".

And recall the three types of tags. Due to the context-independent characteristics, descriptive tags (such as "lovely" and "cool") and action tags (such as "tobuy") can hardly be used as categories. (Given another implicit assumption in sinha's model that each tag was once a candidate of category).

Therefore, we proposed that the concept activation stage is actually twofold. In the first phase the object is identified and basic level name is generated. In most cases, category name is chosen in this phase and earliest tags emerge. Then descriptive and action concepts are generated with higher level of cognitive process. Other tags are determined in this phase. For example, to tag an introductory article of the latest laptop computer “tobuy”, one should read through the article, probably compare its specifications to her expectation, calculate her monthly budget to determine the financial availability, then finally generate the idea.

Many users reported that categorization becomes easier when a well-defined categorization structure has been established. This can be explained by that the cognitive process becomes mainly category membership judgment process. Users do not need to activate and compare among new concepts, they just compare the object with a representing prototype of each category to determine category membership. However, tagging users are not benefitted much from this procedure, one reason is there often are too many tags to compare, and tags such as tobuy cannot be determined by comparison since it is context-dependent.

6.1.2. Users do filter and modify tags

Another cursory statement in Sinha’s model is that a tagging user can just write down whatever concepts emerged in her mind, without considering the amount or wording of tags. We found this statement especially questionable, from our observation and interview, and from our own experience.

Tagging is generally a bottom-up process, in which users perceive a large amount of concepts quickly. For example, in our experiment using textual materials, an article about symphony orchestra might refer to over 10 instruments in one paragraph. Even in pictorial setting, many different concepts can still be presented at the same time. Although theoretically and technically the number of tags that one object can have is not limited, obviously users still have to determine a proper amount, therefore some trivial concepts have to be dropped.

Also users do consider wording of tags. while we say tagging is a process with more freedom than categorization, it does not mean that users put no consideration to their choices of tags. In contrast, we will argue that compare to categorization, tagging needs more wording consideration as a context-dependent process.

In the experiment we often observed that user modified or deleted tags. The fact can be strongly convincing that users do care about tag wording even under personal information management context. Several types of possible actions are summarized below:

Deletion: users completely drop a tag from further consideration.

Combination: users combine two related tags into one, for example, to combine “blue”

and “sky” into “bluesky”.

Reinforcement: users add new modifier or supplement to current tags, for example, modify “hair” to “longhair”.

Why wording is not such an important issue to categorization? Firstly recall the discussion of basic level name; it is probable that users name the category from their previous knowledge. However tagging is largely context-dependent, additional effort might be used to incorporate the contexts. Moreover, even if to name a category needs, more mental effort than a tag, but generally there are far more tags, consider users may, and quite often label more than one tag to an object.

In collaborative tagging, the issue becomes even more complicated since users may consider how their tags appear to others. For example, in hope to be searched more, users may tag their photo using general and various tags, including different forms such as plurals, thesaurus or related words. From another perspective, one may want to establish a decent online image, so she may consider the wording of her tags, replace common concepts with advanced words. Above are just two exemplars of possible situations. These specific tagging patterns may also be a potential mental workload source.

6.1.3. Different memory usage

A practical issue that Sinha missed in her argument is human memory capacity. In our interview and experiment it has been proven that users tend to use consistent tags for similar content, given they can remember them. This approach can maintain personal consistency for better retrieval performance, and avoid mental workload to generate new tags. To successfully reuse tags, users have to search for previous used tags or categories in short term memory or long term memory. Performance of such searches is influenced by limited memory capacity, especially in a bundle of successive organization tasks. Of course this is a problem for both tagging and categorization, but generally in tagging there are far more tags stored in memory.

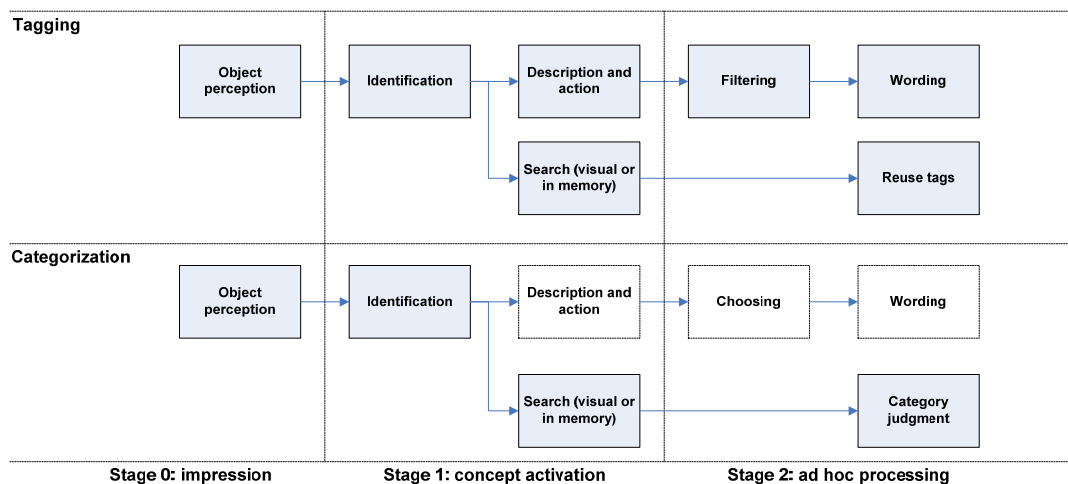
Similar problem occurs in visual information search. Most modern tagging systems adapted suggestion systems to help users maintain consistency. A suggestion system in tagging provides users with their previous tags, as well as tags of others. User can choose to reuse certain tags from the list. Generally the process is like to choose an appropriate category for one object, however compare to categorization system, there are far more tags to be presented. Easily a such list contains hundreds of tags. Though techniques like Tag Cloud, Semantic clustering have been applied or proposed to facilitate visual information retrieval in tag list, it is still a mental costly process to search for a specific tag in a large list. Our experiment system also includes suggestion system, which may be also a source of intense mental workload in tagging.

A last reason is that users may try to remember certain patterns of tagging/categorization, which might be also more difficult in tagging. For example, For example, one may categorize Mozart as “musician”, for she remembered that she has categorized Bach as

“musician” too. Or she may tag Mozart as “musician, Austria, Requiem, Classical” since she tagged Bach as “musician, Germany, Brandenburg Concertos, Baroque”

6.1.4. A revised cognitive model

Based on above discussion, we proposed a modified cognitive model of tagging and categorization as personal information organization tool.



Stage 0: perception and impression

The object is perceived in this stage and its initial impression is made. For textual materials, it may be skimming of the title and first few sentences, for image material, it may be the first glance at the picture to form a general impression of its style and topic. Mental activities in this process are very low since it is mainly about perception.

Stage 1: concept activation

Generally this stage is what Sinha briefed as “multiple concepts are activated”. However, the stage is divided into two phases. The first one is identification of what the object is and formation of basic level name. In most cases, the basic level name is used as the category name, at the same time the earliest tags emerged. This phase ends shortly. Some mental activities are involved in this phase but generally with low workload. The second phase requires higher level of mental engagement, descriptive and action concepts are generated. Tagging process continues in this phase. In some cases categorization process also continues to this phase, and this continuation may due to ambiguity of topics or different usage pattern. This phase lasts for a longer time and may involve continuous concept activation and mental activities.

Stage 2: ad hoc processing

In this stage, concepts are processed to become desired outcomes: category or tag. In

tagging system, users usually need to determine an appropriate amount of tags, and determine each tag's wording. While in categorization, our discussion has revealed that the basic level name is often chosen as the category name, therefore these steps are skipped frequently.

In both stage 1 and 2, the lower paths are listed as alternatives for users to reuse tags or categories. In both systems users should search in memory or previous item lists to reuse such items. However, in categorization users have fewer items to search generally.

The significant difference of temporal demand in information retrieval task was also surprising at first, since no time-related instruction was given during the whole experiment, and the interfaces for both information retrieval systems are nearly the same. Thus we focused on the only difference, which is that there are far more tags than categories that subjects should performance visual and memory search within. As we argued above, to search a particular item in previous tags should be much difficult than in categories. This difference may be the source of the significant temporal demand distinction.

6.2. Discussion for other dependent variables

In information retrieval session there is no significant difference in required clicks between tagging and categorization users. This result is surprising since we expect that the hierarchical structure will need more click to navigate (i.e. to enter a sub-category one must enter the superior category first.). Through analysis of experiment log, we found that the additional clicks in categorization are caused mainly by hierarchical structure, while additional clicks in tagging are caused by a few questions. A possible explanation is that the retrieval process of categorization is top-down while that of tagging is bottom up. Thus when a tagger does not remember critical detail that she used to tag an article, the search process will likely consume many clicks. For example, a user may tag music works by their authors, however, if she did not remember the author of Requiem, she has to check each musician tag, while in categorization, she only has to remember Requiem is a music work. This may reflect that categorization is relatively robust to memory ambiguity.

Time estimation data shows no significant difference between two groups, and ANCOVA shows the time estimation performance significantly depends on personal differences. Nevertheless, time estimation data shows significant difference between baseline and concurrent task performance. This result indicates that time estimation may not be a very sensitive method in differentiating slight mental workload difference, as [25] also pointed out.

Although no significant difference of satisfaction and learning effect was found, satisfaction score of information organization task showed a slightly tendency that categorization > tagging (3% difference, $p=0.211$), and learning effect score showed a tendency that tagging > categorization (13% difference, $p=0.177$). These two tendency can

be also explained by above discussion, that low mental workload of categorization in information organization task results in higher satisfaction, however lack of high level mental engagement results in low learning score in categorization subjects.

While preparing this paper, we noticed a study focusing also on the mental workload of tagging in information organization and retrieval tasks [29]. Using pictorial material, they found tagging has a significant higher frustration subscale score than categorization in information organization task, as well as trend of higher overall mental workload, which is consistent with ours. However, they did not make improvement to Sinha's cognitive model, which as discussed above may have several flaws. It is also not very convincing to claim that the higher mental workload imposed by tagging is from the need to generate specific tags, since as a bottom-up process, it should be easier to extract detail information in tagging.

7. Conclusion

Through a preliminary think-aloud study and a controlled experiment, the cognitive process in tagging and categorization is explored. In the context of personal textual information management, tagging appears to impose higher mental workload to users. We may conclude that tagging is not preferred as the primary information management strategy due to the higher mental demand. And the trends of satisfaction and learning effect score indicate the high mental demand of tagging may degrade its satisfaction score but enhance learning effect.

Sinha's cognitive model of tagging and categorization was reviewed and modified. Several important considerations in these two strategies are discussed in detail, including basic level name, different cognitive level of tags, and alternative cognitive paths in matured systems. This largely modified model is to better describe the observed process and couple with experimental data.

However, we believe that the non-hierarchy and non-exclusiveness of tagging, as well as the collaborative methodology, will serve as excellent complement of categorization scheme. A notable fact is that in collaborative tagging system, the mental workload to generate tags is distributed to the user community. Most popular collaborative tagging systems such as del.icio.us have deployed suggesting system, which shows user the popular tags that other users applied to the same content. This should be able to provide users with strong cognitive clues. The non-exclusiveness of tagging enables users to apply multiple tags, which will together convey more detailed info about the object. The flexibility of tagging provides infinite possibilities for customized and personal information structures, for example, using tags `"*"`, `"**"`, ... `"*****"` to rate content, or using `"recommend_toread"`, `"for: username"` to establish a simple recommendation system across different users. The possibilities are only limited by imagination.

Certain limitations exist for present study. The proposed cognitive model needs further experiments to validate each step, and the interaction style of current experiment system is slightly different to popular tagging websites.

It might be interesting to explore possible cultural difference between Eastern and Western taggers, as it is believed Western culture focus on facts and reasoning while Chinese focus upon emotion and relations. Will these preference influence tagging pattern? It may be also interesting to compare tagging under personal and social contexts, will user experience different cognitive process, involve different tag patterns and considerations? Overall, in-depth and close examination of users' tagging behavior using controlled experiment would be greatly helpful to better assess this novel information management strategy.

8. References

[1] <http://del.icio.us>

[2] <http://www.flickr.com>

[3] <http://www.citeulike.org>

[4] Vander Wal, T. (2005) Folksonomy Definition and Wikipedia.

<http://www.vanderwal.net/random/entrysel.php?blog=1750>

[5] Shirky, C. (2005). Ontology is overrated: Categories, Links, and Tags.

Retrieved from: www.shirky.com/writings/ontology_overrated.html

[6] Wu, X., Zhang, L., & Yu. Y. (2006). Exploring Social Annotations for the Semantic Web. WWW 2006, May 23–26, 2006, Edinburgh, Scotland. Retrieved from: <http://www2006.org/programme/files/pdf/4071.pdf>

[7] Guy, M. and Tonkin, E. (2006). Folksonomies: Tidying up Tags?. *D-Lib Magazine*, January 2006, 12 (1). Retrieved from: <http://www.dlib.org/dlib/january06/guy/01guy.html>

[8] Golder, S. and Huberman, B.A. (2006). Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science*, 32(2): 198-208.

[9] Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006). Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, 31–40, New York, NY, USA. ACM Press.

[10] Halpin, H et al. (2007). The complex dynamics of collaborative tagging, *WWW 2007*, May 8-12, 2007. p211-220

[11] Veres, C. (2006). The language of folksonomies: What tags reveal about user classification. *Lecture Notes in Computer Science*, 3999, 58-69.

- [12] Noruzi, A. (2007). Folksonomies: Why do we need controlled vocabulary? *Webology*, 4(2), editorial 12. Retrieved from: <http://www.webology.ir/2007/v4n2/editorial12.html>
- [13] Chen, A. et al (2007) Predicting Social Annotation by Spreading Activation, *Lecture Notes in Computer Science*, 4822, 277
- [14] Sinha, R. (2005). A cognitive analysis of tagging. http://www.rashmishinha.com/archives/05_09/tagging-cognitive.html. Online; accessed March 13, 2008
- [15] Macgregor, G and McCulloch, E. (2006) Collaborative tagging as a knowledge organisation and resource discovery tool, *Library Review*, Vol. 55, No. 5, pp. 291-300
- [16] Hassan-Montero, Y. & Herrero-Solana, V. (2006). Improving tag-clouds as visual information retrieval interfaces. In *Proceedings of the International Conference on Multidisciplinary Information Sciences & Technologies*.
- [17] Hotho, A., Jäschke, R., Schmitz, C., & Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. *Lecture Notes in Computer Science*, 4011, 411-426.
- [18] Mathes, A. (2004). Folksonomies - Cooperative Classification and Communication Through Shared Metadata. *UIC Technical Report*.
- [19] Veres, C. (2006). Concept modeling by the masses: Folksonomy structure and interoperability. *Lecture Notes in Computer Science*, 4215, 325-338.
- [20] Gruber, T. (2005). Ontology of folksonomy: A mash-up of apples and oranges. *1st On-Line Conference on Metadata and Semantics Research (MTSR '05)*.
- [21] Butterfield, S. (2004). Sylloge
<http://www.sylloge.com/personal/2004/08/folksonomysocial-classification-great.html>
- [22] Wu, H et al. (2006) Harvesting Social Knowledge from Folksonomies, In *Proceedings of the 7th ACM/IEEE joint conference on Digital libraries* 107-116
- [23] Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization*. Hillsdale, NJ: Erlbaum. Reprinted in: Margolis, E. and Laurence, S. (Eds.) (1999). *Concepts: Core readings*. Cambridge, MA: MIT Press.
- [24] HART, S. G. (1975), Time estimation as a secondary task to measure workload. In *Proceedings of the 11th Annual Conference on Manual Control* NASA TM X-62.464, 64-77.
- [25] Liu, Yili and Wickens, Christopher D. (1994). Mental workload and cognitive task automaticity: an evaluation of subjective and time estimation metrics, *Ergonomics*, 37:11, 1843 – 1854
- [26] Zakay D., & Block, R.A. (1997). Temporal Cognition, *Current Directions in Psychological*

Science, 6(1), 12-16

[27] HART, S. G. and STAVELAND, L. E. (1988). Development of NASA-TLX (Task Load Index): results of empirical and theoretical research, in P. A. Hancock and N. Meshkati (eds), *Human Mental Workload* (Amsterdam: North Holland), pp. 139-183..

[28] Cook, J.R. (1991). Cognitive and Social Factors in the Design of Computerized Jobs, *Doctoral Dissertation*, Purdue University

[29] Pak, R., Pautz, S., & Iden, R. (2007). Information organization and retrieval: An assessment of taxonomical and tagging systems. *Cognitive Technology*, 12(1), 31-44.