Enhancing Information Scent: Identifying and Recommending Quality Tags

Shaoke Zhang
College of Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16802 USA
suz114@ist.psu.edu

Umer Farooq
Connected Systems Division
Microsoft
1 Microsoft Way
Redmond, WA 98052
umfarooq@microsoft.com

John M. Carroll
College of Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16802 USA
jcarroll@ist.psu.edu

ABSTRACT
We described a scenario of tag use and an empirical study of tags as socio-cognitive artifacts providing information scent. We articulated a three-step use scenario of tags, and used it to conceptualize tag "quality" as determined by use. We designed and conducted a user study to explore what attributes of tags and taggers predict the user-rated "quality" of tags. We found that frequency best predicted tag quality, while information entropy provided further refinement. We found that people rated our identified quality tags higher in quality than general tags. But these identified quality tags were not perceived better than self-generated tags. We derived a regression model for tag quality and discussed implications for social computing.

Categories and Subject Descriptors
H5.3 [**Information interfaces and presentation**]: Group and Organization Interfaces-*collaborative computing*.

General Terms: Experimentation

Keywords: Social bookmarking, sense-making, quality tags.

1. INTRODUCTION
Social tagging systems such as del.icio.us and Flickr are popular among web users and have recently attracted attention from researchers as a focus of study in Human Computer Interaction and Computer Supported Cooperative Work (e.g., [1-4]). Tagging systems allow users to generate labeled hyperlinks (i.e., tags) to web content for the purposes of further retrieval. These tags are typically keywords or short phrases assigned to any piece of information (e.g., website, photo, video, document, etc). In this sense, tags serve as user-generated metadata, allowing web content to be browsed and searched later. However, compared with traditional metadata that is typically generated by experts, tags are assigned freely by a large number of end users. In fact, not all tags are of high quality because most users are not experts [4]. Further, end users have different tag vocabularies [3, 5] based on their own mental models and assumptions. Herein lies the problem: with the abundance of tags generated by a diverse and large number of users, how can quality tags be identified, and if they can be identified, how can these tags benefit the users?

Our paper addresses the above-stated problems. Viewing tagging system as socio-cognitive artifacts designed to aid user's information foraging [6, 7] and sense-making [8] tasks, we articulate a scenario of tag use as a design representation [9]. The empirical part of our paper parallels the scenario, and explores how quality tags can be identified.

GROUP'09, May 10–13, 2009, Sanibel Island, Florida, USA.

2. MOTIVATION AND LITERATURE REVIEW
As social tagging systems scale up, the vocabulary of tags increasingly stabilize to reach statistical regularity and form tag patterns [10]. There have been many studies on probability distribution of tags [11], their growth patterns [10], and attributes such as tag frequency and entropy [1], tag correlation [12], tag similarity [3], and tag non-obviousness [2]. These patterns and attributes were drawn from large-scale, macro-level statistics, which reflect collective thinking and other social trends.

However, these studies of aggregate patterns and statistics provide little guidance on helping individual users better navigate social tagging systems and facilitate their information foraging behaviors. Few studies have discussed how these attributes at the macro level can be used at the micro level to help, for example, discriminate tags.

In fact, there is no clear definition of "quality" for a tag. The statement by MacGregor and McCulloch—"terms assigned to resources that are exhaustive will result in high recall at the expense of precision…, terms that are too specific will result in high precision, but lower recall" [13]—implies at least two aspects of quality: precision and recall. The characteristics of tags such as exhaustiveness and specificity may be implied or predicted by those statistical attributes. For example, tags applied to too many items may be too general; tags applied to fewer items may be more discriminate; and tags applied for too few times may be too obscure [2, 5, 14].

Although these studies imply the quality of tags, few studies discuss it explicitly. One exception is the paper by Sen and colleagues [4], which allowed users to rate the quality of tags for movies in MovieLens. Their results suggested a trend that more frequently applied or searched tags were usually rated better.

However, we have not yet seen any conclusive study that examines which and how tag attributes can predict tag quality with statistical significance.

Quality tags can be recommended to users. Studies have found that as more tags exist, it is less likely that the next assigned tag is new [5]. Therefore, recommending tags can facilitate tag reuse, which is one goal of social tagging systems for achieving a converging vocabulary [2].

Recommending good tags may also improve the quality of the tagging vocabulary. Fu [15] suggested that the quality of tags influences the formation of user's mental categories as well as information search efficiency. Furnas et al. [16] claimed that different users may use different terms to describe the same thing. Supporting this "vocabulary problem" [16, 17] is where social tagging systems outweigh conventional taxonomies [18]. Recommending tags can "induce conforming behaviors" by adapting users to better shared vocabulary [5]. Sen et al. [4] found that recommending tags changes the user's selection of tagging vocabulary and thus changes the distribution of tag classes.

3. TAGS AS INFORMATION SCENT ARTIFACTS

Since Vannevar Bush's vision of Memex [19], researchers have been studying the use of information systems to facilitate information foraging and sense-making. In information foraging theory, user strategies and technologies for information seeking, gathering, and consumption are adapted to the flux of information in the environment according to their costs and benefits [6]. Just as animals rely on scents to forage, users rely on information scent provided by various cues in judging information sources and navigating through information spaces. Tags serve as proximal cues that provide information scent. They can be considered as an external representation of users' mental concepts activated by web items. In later information foraging tasks, by re-activating these concepts, they provide "the imperfect perception of the value, cost, or access path of information source" [7].

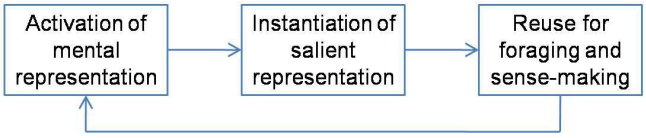

Figure 1. A scenario of tag use.

Figure 1 represents a typical scenario of tag use for information foraging and sense-making. We use the example of web documents as typical web items that can be tagged. First, when the user is viewing certain web content, mental concepts are activated. The user may instantiate salient concepts as tags. After these tags are generated, the user and others can use them for later information retrieval, sense-making, and information foraging.

3.1 Spreading Activation of Semantic Concepts

A visual web item uses symbolic representations to convey information. Web documents consist of semantic terms, which can activate users' cognitive processing of corresponding mental concepts. These concepts as internal representations can be represented with nodes in a knowledge network, with properties of the concepts represented as labeled relational links from the nodes to other concept nodes [20]. Reading and learning a new web document can be understood as changing the relationship between certain conceptual nodes, that is, the activations spread by tracing an expanding set of links in the network of these concepts. Such underlying cognitive mechanisms are used to explain how social tagging system influences knowledge acquisition and adaptation [15].

Information acquisition behavior gives increment to or reconstructs human knowledge. Therefore the information forager may have different (or different level of) activation of mental concepts on the same web content in different times. For example, when he or she gets access to the same document after a period of time, the forager may have deeper understanding on the same problem; it is equally possible the forager has forgotten something where certain "cognitive artifact" [21] (e.g. tags) may be of help.

Similarly, different information foragers may have different knowledge backgrounds and thus different activations on the same web content. Furthermore, if others' understandings (as instantiated by tags) are provided in combination with the web content, the user's mental activation may be influenced, where we call tags as "socio-cognitive artifact".

3.2 Instantiation of Concepts as Tags

Tags, as cognitive and socio-cognitive artifacts used to augment users' cognitive capabilities in information foraging and sense-making tasks, are usually intentionally generated by users for further use. The question is which semantic terms would be assigned as tags. Theoretically, all semantic terms having certain activation can be assigned as tags for information scent. However, not all semantic terms or tags are of equal quality, given their varying levels of activation. It is reasonable that users will identify information of interests, and appropriately instantiate salient representations as tags for further reuse. Therefore, users' interest, salient activation, and expected reuse are critical to determine such prospective generation of tags.

It is notable that this instantiation process is problematic due to bounded rationality and choices based on satisficing [22]. When users are making decisions in assigning tags to a certain document, they may not always come up with the highest quality tags, given their limited knowledge, skill, and time. They may just choose "good enough" options (see also "naturalistic decision making" in [23]).

Weick [24] pointed out that in sense-making processes, people have inability to shift representations easily due to the inertia of their representations, and inability to find and use appropriate data. These inabilities can be reflected by individual user's preferences and bias of selecting tags, which can be complemented by social tags in a community level. Therefore, identifying and recommending others' quality tags can alleviate such problems.

3.3 Reusing Tags for Information Scents

Tags not only reflect the internal mental representations but also provide external information scent to reconstruct certain representation in later information processing tasks. They are usually generated intentionally and prospectively to reduce the cost of later operations. Therefore, tag quality can be defined as

how effectively and efficiently it helps later information retrieval, information foraging, and sense-making.

Such quality can be examined in the reuse phase of tags. Broadly, there can be two different kinds of usage for tags: information retrieval and exploratory search [15]. For information retrieval, tags provide information scent to relocate the document (e.g. keywords to search) and restore activations that have decayed. For exploratory search, tags can give users main topics of a document, and give information scent to locate related interested documents. These tags provide "imperfect information at intermediate location" [7] to decide on paths through online databases to target information.

As moderated by prior knowledge, different users may gain different information scent from the same tags. The same user may also modify his or her tags with reconstructed knowledge in information foraging tasks. Furthermore, the tags become a part of the semantic terms of the document, which provides different activation and information scent.

4. RESEARCH QUESTIONS

We seek to answer three research questions based on the scenarios of tag use presented earlier. The first research question is related to the identification of quality tags. Tags, as part of a language, come from practical use and natural understanding [25]; the underlying knowledge is also socially constructed [26]. Thus, the tags are ontologically subjective (semantic terms with different activation levels according to the scenario). Users may have different understandings and vocabularies for the same referred items. Therefore, users may choose tags based on their personal tendency, preferences, and beliefs [5]. Indeed, this will result in individual differences when identifying which tags are of high quality. This creates the need to objectively identify quality tags by assessing the attributes of the semantic terms in a tag set. Our first research question (RQ) can be stated as:

RQ1: Which attributes can objectively identify quality tags?

Several attributes can be used to objectively assess the quality of tags. Based on our scenario and prior literature, we have identified three such attributes to explore as part of RQ1: centrality, frequency, and entropy. Each of these objective measures is explained below with an associated hypothesis.

Reading a document activates the network of several mental concepts. Users select salient internal representations as tags. These selected tags are mentally linked, forming a network that depicts the main relationships in the original network of all activated mental concepts. Therefore, we can construct a network of tags representing their semantic relationships for each paper. Some researchers have tried to visualize tag correlation networks (e.g., [27]). This provides us opportunities to describe attributes of tags based on network analysis. Centrality is one such measure to assess the relative importance of a node within a network.

Tags with high centrality in the tag network are usually the most salient terms. Taking eigenvector centrality, for example, the centrality score of a node is proportional to the sum of the scores of all nodes that are connected to it. Therefore, a tag with high eigenvector centrality implies that, first, this tag is salient because it is related with many other tags; and second, related tags are also salient enough. Therefore, tags with higher centrality may have higher activation, which leads us to articulate our first hypothesis:

RQ1 (H1): Tags with higher centrality have higher quality.

If a tag is frequently assigned to a document, this tag has salient activation for many users, which suggests that it is highly possible it will have salient activation for an additional user. Recommender systems also recommend options with high usage to users [28]. This leads us to articulate our second hypothesis:

RQ1 (H2): Tags with higher frequency have higher quality.

Different documents may activate the same mental concepts, especially when these documents are from the same area. Consequently, same tags may be assigned to different documents. Given a tag X assigned to a set of documents, each of which occurred with probability $p(x_i)$, the entropy H(X) is defined as:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log(p(x_i))$$

Here x_i represents the event that tag X is assigned to certain document d_i. Entropy measures uncertainty about a particular event associated with a probability distribution. When a tag is assigned to only a few particular documents, that is, when the information entropy is low, the tag has a high discriminating value for foraging these documents. Chi and Mytkowicz found social tagging systems become harder and harder to navigate because of the increasing entropy [14]. If a tag is dispersed in more documents, there is a higher uncertainty to determine which document is associated with that tag; therefore, this tag has higher entropy. Tags with high entropy may be "too general" [5] without "precision" [13] and "discrimination value" [2]. On the other hand, tags with low entropy may be too "specific", too "hard to recall"[13], and may be too "obscure" [5]. From this perspective, both high and low entropy of tags imply lower information scent and thus lower quality. This leads us to articulate our third hypothesis:

RQ1 (H3): Tags with medium entropy have higher quality.

In addition to attributes of tags, attributes of users can also be used to identify quality tags. Studies found that experts usually gave more accurate descriptions of a task than novices [29]. Expertise recommendation systems help locate experts to provide recommendations [30]. Therefore, users with higher related expertise may also provide better tags, which are also descriptions at a meta-level. Our second research question is:

RQ2: Can experts identify quality tags better than average users?

Based on RQ1 and RQ2, we may identify quality tags with the attributes of tags and taggers. However, users may have different preferences of tags [5] due to their different understandings and vocabularies for the subjective semantic items. The question then becomes: is there value in identifying and recommending others' quality tags? As stated earlier, users are typically not completely rational. In order to determine values of recommending tags, we need to empirically validate how people perceive the quality of their own and others' tags, and most importantly, how our identified quality tags will be perceived.

RQ3: How do people perceive our identified quality tags?

5. METHODS

In this paper, we selected the tagging of scholarly papers as the domain context for social bookmarking. Scholarly papers activate abundant meaningful mental concepts, which is a direct application of the spread activation theory in our scenario.

To address the research questions, we conducted an online survey-based user study with two phases aligning with our scenario of use. In the first phase, participants were asked to read four scholarly papers (corresponding to the activation process in the scenario). Then they assigned tags to each of those papers (corresponding to the instantiation process in the scenario). In the second phase, participants were asked to evaluate how well the tags described each paper (corresponding to the reuse process in the scenario).

5.1 Materials

As researchers usually read and use scholarly papers in their own areas, we chose HCI as our area of focus because of our familiarity with the literature. Three of the four papers ([31-33]) were selected from proceedings of the ACM Conference on Computer Supported Cooperative Work (ACM CSCW); the fourth paper [34] was selected from proceedings of the SIGCHI conference on Human factors in computing systems (ACM CHI). They were all popular papers that had been cited for more than two hundreds times. According to Google Scholar, the Dourish and Bellotti paper [31] had been cited 1345 times; the Gutwin et al. paper [32] had been cited 209 times; the Grudin paper [33] had been cited 618 times; and the Rodden and Wood paper [34] had been cited 221 times.

These four papers were chosen based on the variance of their similarity. We chose "CSCW" as one discriminator of similar papers, and "group awareness" as a further discriminator. Accordingly, sixteen popular papers (cited more than 100 times) from the pool of all ACM CHI and ACM CSCW proceedings were selected. A panel of five HCI experts discussed their similarities and finally narrowed the set down to these four papers. According to our discussion, Dourish and Bellotti paper and Gutwin et al. paper were the most similar because they both discussed awareness in CSCW; the Grudin paper had some similarity because it dealt with CSCW issues; the Rodden et al. paper was quite different but in the HCI area.

The user study consisted of two online surveys corresponding to the two phases. Both surveys were conducted in Surveymonkey (www.surveymonkey.com). Counterbalancing was done to ensure that presentation order of the papers did not affect user behavior in both phases. All participants entered a lottery to win a $100 gift card; two winners were randomly selected.

5.2 First Phase: Assigning Tags

As our tagging objects were scholarly papers in HCI, we selected HCI professionals as our participants. We provided participants titles, abstracts, and hyperlinks to full papers, asking them to assign between five to seven tags to each. We confined the number of tags to ensure we get a consistent and sufficient amount of data while avoiding burdening the participants by asking them to provide more than several tags. In addition, according to our experience in social tagging systems, users generally assign about five to seven tags to a particular document.

We recruited participants by sending a recruitment email to the ACM CHI and AIS HCI mailing list. 90 participants completed our survey, providing 1,790 tags for all papers. 525 tags were distinct tags. Of these 90 participants, 68 of them were from academia (with 24 graduate students, 12 research associates or post doc, and 32 professors), 17 of them were from industry or government; there were 5 missing values. 44 participants were male and 40 were female, with 6 missing values. The average time participants had been involved in the HCI area was 12 years with a standard deviation of 8.9 years.

Two researchers independently performed data cleaning and consolidation. As the tags were rather subjective, we consolidated tags conservatively under the following five conditions: 1) abbreviation or acronyms (e.g., HCI vs. human computer interactions); 2) capitalized letters (e.g., user interface vs. User Interface); 3) misspelling (e.g., collaboration vs. colaboration); 4) words with hyphens or dashes (e.g., human-computer interaction vs. human computer interaction); 5) plurality (e.g., widget vs. widgets). The data cleaning and consolidation was discussed between the two researchers to reach full agreement. After this process, 119 distinct tags for the Dourish and Bellotti paper were consolidated to 107 tags; 119 distinct tags for the Gutwin et al. paper were consolidated to 105 tags; 179 distinct tags for the Grudin paper were consolidated to 161 tags; and 197 distinct tags for the Rodden et al. paper were consolidated to 171 tags.

5.3 Second Phase: Evaluating Tags

To measure how well these tags provide information scents, we asked participants to evaluate these tags in the second phase. Specifically, we asked participants to evaluate how well the assigned tags described the papers. We sent a survey via email to 77 of the 90 participants from the first phase who agreed to take part in the second phase of the study by providing valid email addresses. The second phase survey was sent out one month after completion of the first phase. We thought one month would be long enough for certain decay of their memory, and short enough for certain re-activation. 46 participants completed the survey. 40 of them provided valid email addresses so that we could match their data from the first phase.

As the consolidated tags from the first phase followed a power law distribution, we categorized these tags according to their frequencies into high, medium, and low groups. Tags with the highest quartile of frequencies (i.e., larger than 8) were categorized as high frequency; tags with the second quartile of frequencies (i.e., less than 8 and larger than 3) were categorized as medium frequency; tags with the third and fourth quartiles of frequencies (i.e., frequency of 1, 2, and 3) were categorized as low frequency.

For each paper, from each of the three frequency groups, we randomly selected 9 tags to be evaluated to be consistent across the four papers. In this way, we had 27 tags for each paper. We asked participants to rate the quality of these 108 tags on a 7-point Likert scale from "not good at all" to "very good". This rating scale was based on Sen et al.'s study [4]. In our scenario, we defined quality tags as those providing the most information scent for later information foraging tasks. Thus, asking participants to evaluate the tags on a scale was a practical and feasible approach to assess perceived quality of tags.

6. RESULTS

In this section, we first present manipulation checks to confirm the similarity variance that was taken into account while selecting the four papers. We then address our research questions and hypotheses.

Figure 2 illustrates a network analysis of tags. Due to space limitations, we selected tags entered in the second phase of the study as an illustration. The cluster of tags in the top left corner are from the Dourish and Bellotti paper; the top right cluster represents tags from the Gutwin et al. paper; tags on the bottom left are from the Grudin paper; tags from the bottom right are from the Rodden et al. paper; tags in the center are the ones shared by more than one paper. As a manipulation check, we can see that the links between the first three ACM CSCW papers are denser than with the ACM CHI paper.

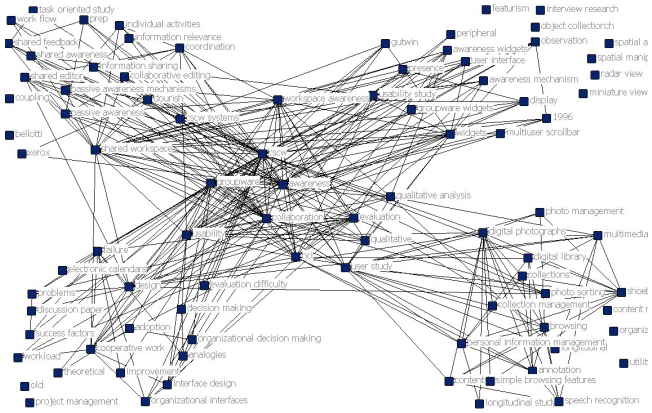

Figure 2. The network analysis of tags in four papers.

As another manipulation check, we also explored the cosine similarity based on the assigned tags. Two papers A and B have n tags in total. Each tag i appear a_i times in paper A, and for b_i times in paper B. Given vectors A={a_1, a_2, …, a_i, …, a_n} and vector B={b_1, b_2, …, b_i, …, b_n}, the similarity between A and B could be measured by the Tanimoto coefficient T(A, B), which is defined as:

$$T(A,B)=\frac{A\cdot B}{\|A\|^2+\|B\|^2-A\cdot B}.$$

Table 1 represents the cosine similarity between each of the four papers. The Dourish and Bellotti paper and Gutwin et al. paper has the highest similarity; the Rodden et al. paper has much lower similarities with the three ACM CSCW papers. This result is consistent with our expectation.

Table 1. The cosine similarity between each paper

	Dourish and Bellotti paper	Gutwin et al. paper	Grudin paper	Rodden et al. paper
Dourish and Bellotti paper	1	0.455	0.323	0.021
Gutwin et al. paper		1	0.250	0.026
Grudin paper			1	0.050
Rodden et al. paper				1

RQ1 (H1): Higher centrality tags have higher quality

In the network of tags for each paper, we calculated the eigenvector centrality of each tag as the measurement of their relevant importance in the concept network. We examined whether centrality identified tag quality (as rated by the participants) in the regression model. According to the result (see Figure 3), the centrality can identify the quality of tags. According to the Box-Cox method, the cubic regression model fits best. The equation is:

$$Quality=29.62\times centrality^3-27.23\times centrality^2+9.39\times centrality+3.53$$

This regression model is significant, $p<0.001$. Therefore, H1 is supported. Tags having higher centrality within the paper have higher quality. Furthermore, we found that the quality of tags is a cubic function of centrality. The regression model explains 23% of the total variation.

This result confirms our hypothesis. Furthermore, as shown in Figure 3, the cubic model for centrality suggests that only tags with extreme centrality can be distinguished. Only those quality tags with high centrality can be identified. We can see even when the centralities were between 0.4 and 0.5, the quality of some tags were still rated with scores below 4. Two examples were "evaluation" and "design" for the Grudin paper. On the contrary, tags such as "CSCW" and "awareness" had high centrality (about 0.6), and their qualities were also rated high. This is reasonable because tags that best describe a specific paper are not necessarily salient in a whole knowledge network. Therefore, only a small part of quality tags can be identified by centrality. This result suggests that centrality is not a very good predictor for identifying quality tags.

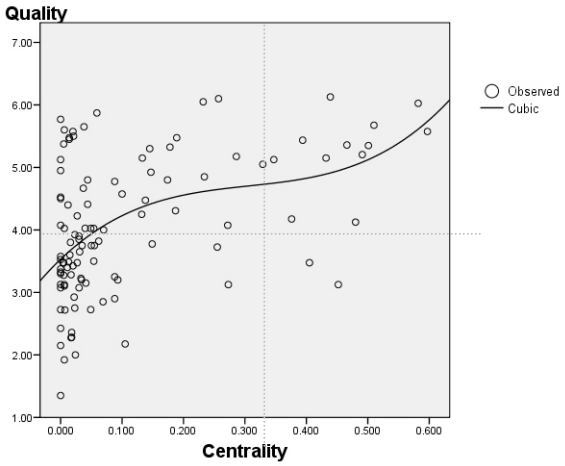

Figure 3. Centrality identifies quality of tags.

RQ1 (H2): Higher frequency tags have higher quality

We then examined whether frequency of tags within a paper identified tag quality in the regression model. According to the result, frequency can predict the quality of tags. According to the Box-Cox method, the logarithmic regression model fits best. The equation is:

$$Quality=3.39+0.46\times\ln(frequency)$$

This regression model is significant, $p<0.001$. Therefore, H2 is supported. Tags having higher frequency within the paper have higher quality. Furthermore, we found that the quality of tags is a logarithmic function of frequency. The regression model explains 27% of the total variation.

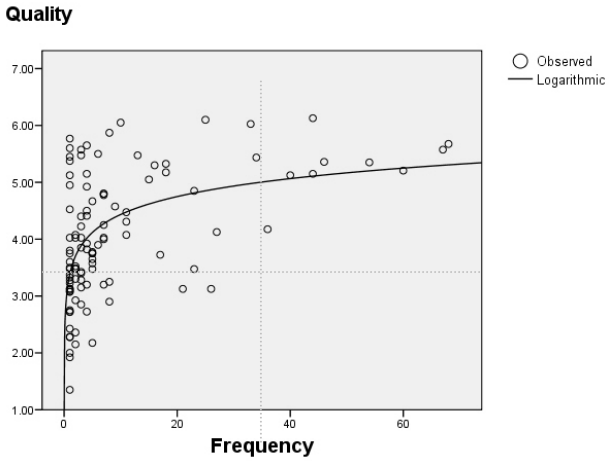

Figure 4. Frequency identifies quality of tags.

As shown in Figure 4, because the quality of tags is a logarithmic function of frequency, tags with higher frequencies were all rated as good tags. Further, there was no tag in the bottom right quadrant, which implies that tags with higher frequency were never of low quality. Therefore, the frequency of tags identifies tag quality reasonably well.

This result confirms our hypothesis. Furthermore, the logarithmic model for frequency makes intuitive sense because some repetition gives us confidence that people agree the tag is appropriate for a resource, but lots of repetition gives only a little more confidence.

RQ1 (H3): Medium entropy tags have highest quality

We also examined whether information entropy identified tag quality in the regression model. According to the result, entropy can predict the quality of tags. According to the Box-Cox method, the quadratic regression model fits best. The regression equation is:

$$\text{Quality} = -10.22 \times entropy^2 + 6.18 \times entropy + 3.78$$

The regression model is significant, $p<0.01$. Therefore, H3 is supported. Tags with medium entropy have the highest quality. Furthermore, we found that the quality of tags is a quadratic function of entropy. The regression model explains 13% of the total variation.

This result confirms our hypothesis. As suggested by the quadratic relationship in Figure 5, when the entropy is medium, the tags usually have the highest quality. When information entropy is 0.30, the mean quality reaches its highest value of 4.71. When the entropy is low and high, the mean quality is low. For example, tags such as "PREP" and "miniature view" had the lowest entropies; their qualities were rated low probably because they are too "specific" [13] and "obscure" [5]. Tags such as "HCI" and "user study" had highest entropies; their qualities were also rated low probably because they were too general terms without much "discrimination value" [2].

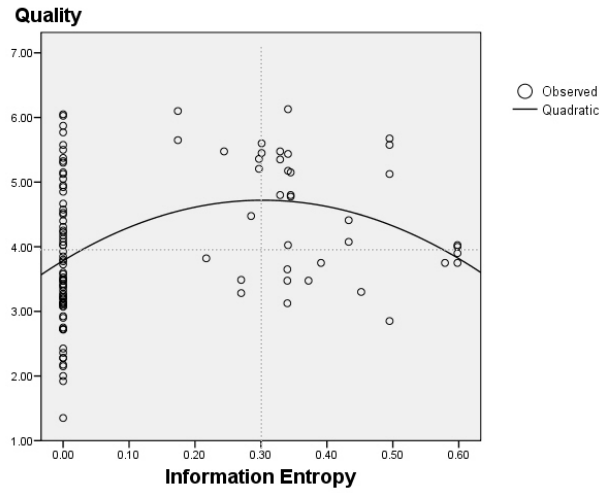

Figure 5. Information entropy identifies quality of tags.

RQ1: Which attributes objectively identify quality tags?

As the quality of tags is the cubic function of centrality, logarithmic function of frequency, and quadratic function of entropy, we put all of the $(\text{centrality})^3$, $(\text{centrality})^2$, centrality, ln(frequency), $(\text{entropy})^2$, and entropy variables into a linear regression model to examine their combination functions. We found ln(frequency), $(\text{entropy})^2$, and entropy entered into the regression model as significant with $p<0.001$. The regression equation is:

$$Quality = 0.41 \times \ln(frequency) - 7.54 \times entropy^2 + 3.90 \times entropy + 3.31$$

The regression model explains 31% of the total variation. Centrality was excluded from the model, which confirms our prior assessment that centrality is not a good predictor.

Figure 6 shows their relationships more directly. We divided the frequency into low and high groups with its median. We divided entropy into three groups as it had a quadratic relationship. As 64.8% of tags had entropy of zero (i.e., tags only applied to one paper), we categorized these tags in the low entropy group. For the tags with entropy larger than zero (i.e., tags applied to more than one paper), we categorized them into the medium and high entropy groups with the median of their entropy.

We can see that frequency is the main predictor of tag quality. The quality of tags with high frequencies (Mean=4.92, SD=0.88) were significantly higher than the quality of tags with low frequencies (Mean=3.75, SD=1.01), $t(106)=5.37$, $p<0.001$.

Entropy provides a refinement for identifying quality tags. For tags with high frequencies, tags with medium entropy had mean quality of 5.18 (SD=0.86); tags with low entropy had mean quality of 4.73 (SD=0.93); tags with high entropy had the mean quality of 4.85 (SD=0.88). However, their differences did not reach statistical significance.

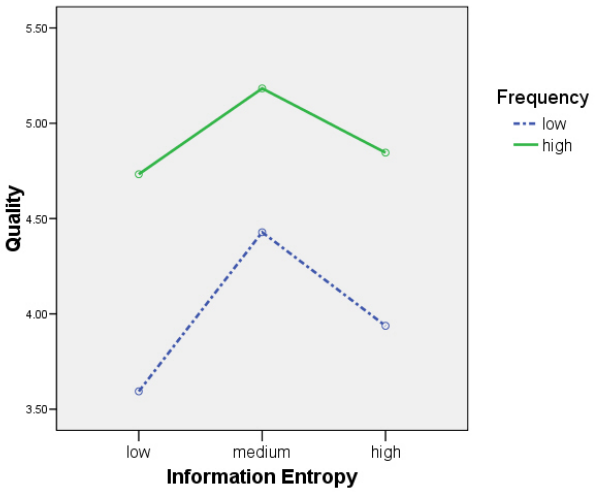

Figure 6. Frequency and entropy together predict quality tags.

RQ2: Can experts identify quality tags better than average users?

To determine participant's expertise in each of the four papers, we provided ACM SIGCHI keywords in the survey and asked participants to check all keywords that could be applied to their expertise and professional interests.

At the same time, two HCI experts decided which of these keywords could be applied to each of the four papers. We used the number of keywords checked by the participants as a percentage of the keywords identified by the two HCI experts as an index of expertise. Although only a subset of tags from the first phase was selected for the second phase, 87 of 90 participants had at least one tag evaluated. We calculated the average quality for each participant's rated tags as an estimation of his or her assigned tags. We found that for each of the four papers, there was no correlation between participant's expertise and their tag quality. The Pearson correlation coefficients were -0.07, -0.05, 0.21, and -0.01 respectively.

We also measured participant's familiarity with each paper on a 5-point Likert scale. No correlation with tag quality was found. The Pearson correlation coefficients were -0.21, -0.13, 0.02, and 0.03 respectively. Further, we examined participant's tagging experience and years in the HCI area, which also reflect some of their expertise. They still had no correlation with the quality of their tags. All of the Pearson correlation coefficients were between -0.14 and 0.16.

RQ3: How do people perceive our identified quality tags?

Self-assigned tags are better than other's tags

Users assigned several tags in the first phase; in the second phase, they evaluated not only their own tags but also others' tags. We compared the evaluated quality of self-assigned tags with the quality of tags assigned by others.

For the Dourish and Bellotti paper, we calculated the mean quality of self-assigned tags (Mean=5.55, SD=1.39) for each participant, and the mean quality of tags assigned by others (Mean=4.19, SD=0.98). According to the paired-samples t test, these two had significant difference ($t(39)=8.34$, $p<0.001$). The quality of self-assigned tags was significantly higher than quality of tags assigned by others.

Similarly, for the Gutwin et al. paper, the quality of self-assigned tags (Mean=5.31, SD=1.38) was significantly higher than quality of tags assigned by others (Mean=4.03, SD=1.00), $t(39)=7.38$, $p<0.001$. For the Grudin paper, the quality of self-assigned tags (Mean=4.79, SD=1.63) was significantly higher than quality of tags assigned by others (Mean=3.36, SD=0.93), $t(37)=6.21$, $p<0.001$. For the Rodden et al. paper, the quality of self-assigned tags (Mean=5.18, SD=1.51) was significantly higher than quality of tags assigned by others (Mean=3.96, SD=0.82), $t(35)=5.44$, $p<0.001$. Therefore, we conclude that self-assigned tags were generally perceived as higher quality than tags assigned by others.

It is striking that, one month later, participants still rated their self-generated tags as significantly better than other-generated tags. This suggests that the utility of tags in general is a highly personalized matter of designing effective scent.

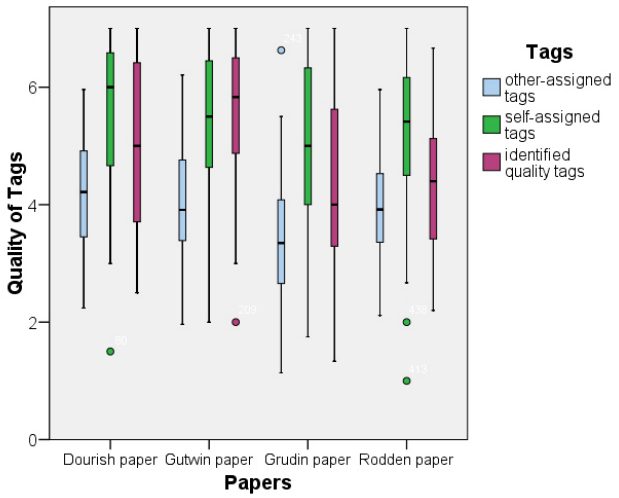

Figure 7. Comparison of qualities of other-assigned tags, self assigned tags, and identified quality tags.

The value of our identified quality tags

A subsequent question is whether our identified quality tags are perceived as better than other's tags. Based on our regression model, we selected five tags with the highest predicted quality as "identified quality tags", and examined how participants perceived these tags.

Our "identified quality tags" were perceived as much better than other-assigned tags in all of the four papers, as presented in figure 7. According to the paired-sample t tests, all of the significance coefficients were smaller than 0.01. This result suggests that our regression model can help identify tags that will be perceived as quality tags by users.

When it comes to compare self-assigned tags and our identified quality tags, it is hard to say which one is perceived better. For

Dourish paper, the perceived quality of identified tags (Mean=5.07, SD=1.40) was lower than that of self-assigned tags (Mean=5.55, SD=1.39), $t(39)=2.70$, $p<0.01$. Similarly, for Rodden paper, the perceived quality of identified tags (Mean=4.39, SD=1.08) was lower than that of self-assigned tags (Mean=5.18, SD=1.51), $t(34)=2.97$, $p<0.01$. However, for Gutwin paper, there was no significant difference between our identified quality tags(Mean=5.54, SD=1.23) and self-assigned tags (Mean=5.31, SD=1.38). Similarly, our identified quality tags in Grudin paper (Mean=4.41, SD=1.53) were not perceived as having different quality as self-assigned tags (Mean=4.79, SD=1.63). See Figure 7 for a boxplot of the comparisons.

We further compared the contents of participants' self-assigned tags with our identified quality tags. We found that 1) identified quality tags have overlaps with self-assigned tags. Therefore, recommending the identified tags could provide tags that users want to assign themselves. 2) identified quality tags can complement or improve users' own tags. Table 2 represents an example of self-assigned tags and identified quality tags for a certain user *i* for the Dourish and Bellotti paper. We can see that user *i* also assigned the identified tag "CSCW" and "awareness". Furthermore, user *i* did not cover tags such as "shared workspace" and "groupware" but rated them high in quality. So others' tags may complement users' own tags. In addition, this particular user rated the identified tag "workspace awareness" better than self-assigned tag "awareness" or "passive awareness", probably because he considered the former was more accurate. This suggests a user's own tags may also be improved by recommending others' tags.

Table 2. An example of self-assigned tags and top five other-assigned tags.

Self assigned tags		Identified quality tags	
Tag	**Perceived quality**	**Tag**	**Perceived quality**
CSCW	7	CSCW	7
Awareness	4	Awareness	4
Collaboration	5	Shared workspace	6
Shared feedback	7	Groupware	6
Passive awareness	5	Workspace awareness	7

7. CONCLUSION AND IMPLICATIONS

In this paper, we articulated a scenario of tag use. This three-step scenario resembles the three main processes in Russell et al's [8] learning loops in sense-making: generation loop, data coverage loop, and representational shift loop. In the generation loop, users search for appropriate representations to capture important regularities, which corresponds with the process of spreading activation of semantic concepts in our scenario. In the data coverage loop, users identify information of interests, and appropriately instantiate their mental representations as "encodons" [8]. Similarly, in the tag assignment process of our scenario, users select salient representations and instantiate them as tags. Representational shift loop is guided by the discovery of residue. Similarly, users' tagging patterns evolve as they assign more and more tags. Tagging fixes the vocabulary for sense making. Every tag and every potential tag narrows and sharpens the sense that people can easily make of documents. Moreover, online tagging is usually a social behavior, which helps the representational shift in a community level. This provides us a scenario to explore sense-making in group or community level, which has not been studied in detail.

Based on the scenario, we conceptualized "quality" as determined by use, and framed an investigation of tags as folksonomy versus a priori or authoritative definitions of quality. Furthermore, we were investigating whether simple characteristics of the user experience can be linked to perceived quality in a schematization of using and retrieving documents via queues: activation of mental concepts by web items, instantiation of mental concepts as tags, and later reuse of these tags for information foraging and sense-making. Our findings—for example, frequent tags are perceived as higher in quality—showed how tag sense-making consensus can work.

This study provided applicable regression models to identify quality tags, which could be directly used by social tagging systems to recommend tags to users. Our study results not only confirmed our hypothesis, but also gave more information about the relationship between tag quality and tag attributes. Centrality has a cubic relationship with tag quality, because tags that best describe a specific paper are not necessarily salient in a whole knowledge network, which suggests that centrality is not a very good predictor to identify quality tags. Frequency has a logarithmic relationship with quality, because the marginal effect of quality decreases as the frequency increase. Entropy has a quadratic relationship between tags with middle entropy has the highest value of discrimination.

Some social tagging systems provided recommendation based on users' collective evaluation of existing tags (e.g., [4]), where motivating their rating may be a problem. Our study used objective measures, in particular the attributes of tags, to identify quality tags. These regression models can be used to predict the quality of tags over time if certain tag attributes are modified. These models can be used by designers to understand how social bookmarking systems are evolving with respect to the quality of tagging vocabulary in their system and what they can do in terms of user recommendation to improve the quality.

We utilized the collective statistics of tags to identify quality tags. There have been many studies describing the probability distribution of tags [11], their growth patterns [10], and attributes such as tag frequency and entropy [1]. However, few studies have explored how to use these data in practice. Our study demonstrated a way to use such aggregate patterns and statistics for recommending tags, which will facilitate individual information foraging behaviors [6] and improve the "quality of a system's vocabulary of tags"[4].

We also showed the value of identifying quality tags. Participants perceived our identified quality tags based on our regression model as higher in quality than general tags. These tags could be recommended to predict, complement, and improve user's tags. In our study, 90 participants assigned more than 100 tags for each of the four papers in the first phase. Such a high number of tags seem to an inefficient model for information foraging. Just as our results showed, participants rated some tags high quality while rated others as low quality. Hence, there is value in discriminating higher quality tags from lower quality tags. This implies that given some attributes of tags that are deemed important for a social bookmarking system, lower quality tags may be allowed to

decay after some time whereas higher quality tags can be sustained in the system (e.g., through tag recommendation). The identified quality tags are recommendable according to our results.

8. LIMITATIONS AND FUTURE WORK

In this study, we explored three tag attributes to identify the quality of tags, which can explain 31% of the variance. The three tag attributes are not exhaustive as they were low hanging fruit identified from existing literature. Other attributes need to be folded into our regression model. For example, "similarity" as suggested by [3] can be used to explore the vocabulary of tags.

Another limitation is that we used only four documents– from HCI/CSCW - as tagging resources. However, the kind of study we conducted (Internet-based survey) makes it difficult to use a larger sample of documents. Our materials were carefully selected, and we presented a manipulation check in our results, which is beyond current method standards for Internet survey research.

In our study, participants' evaluation of tag quality may have had subjective bias due to the generation effect [35]. Users may remember their assigned tags and rated them high. However, the interval of one month between the two phases may be long enough to alleviate this bias. According to our results, even with the generation effect, participants still rated a subset of others' tags as good as or better than their own tags.

Our definition of quality is rather general, which can be different in other scenarios of use. The term "quality" is rather vague, which implies many metrics such as recall, precision [13], discrimination value [2], etc. According to our study, we can build appropriate algorithms to identify and recommend quality tags depend on which aspect of "quality" is valued in later information foraging by particular social bookmarking system. For example, tags with high entropy are good for recall; tags with low entropy have precision. If the goal of a social bookmarking system is to facilitate retrieval of all relevant documents based on searching for tags (i.e., recall is more important than precision), then high entropy tags are preferred.

While we recognized users may have different preferences of tags, we did not discuss personalization issues in this paper. We found 102 of all 108 tags in the second phase had the maximum value of 7 (i.e., perceived as having highest quality); at the same time, 89 tags had the minimum value of 1 (i.e., perceived as having lowest quality). This suggests that people have diverse preferences of tags. However, when we examined the standard deviations of the perceived quality for each tag, we found that the perceived quality does not vary too much for different participants. For all of the 108 tags, the mean of their standard deviations is 1.740; the standard deviation of the standard deviations is 0.277. This result suggests that although some of them may have quite different tag preferences for certain tags, users usually have a shared vocabulary. It is possible that users with similar previous tagging patterns may have similar tag preferences. However, we did not conduct such an analysis because the data was not sufficient. The measurement of tagging patterns would not be robust in the condition that every participant only had five to seven tags for each paper, and there were only four papers.

We found that users' expertise cannot identify quality tags, which is not consistent with other studies [29, 30]. This may be because the assignment of five to seven tags per paper was not an expert task. Further, only a subset of participants' tags from the first phase was rated during the second phase, which may have skewed the results. Lastly, we found some participants with high expertise rated specific tags very well that were rated low quality by other participants. For example, some experts assigned tags such as "PREP" and "Bellotti", while many other participants rated these two tags as low quality. There are some studies (e.g., [36]) claiming that we should not use expert's recommendation exclusively because they have different backgrounds and different concerns. In sum, although we did not find a significant effect of expertise on identifying quality tags, we cannot arbitrarily claim experts do not generate better tags. One way to check this claim is to improve our study with larger user samples with more diverse expertise and backgrounds.

9. ACKNOWLEDGMENTS

We would like to thank Craig Ganoe for his help in designing the user study. We would like to thank Jing Wang for helping tag consolidation and Haibin Liu for helping data cleaning. We appreciate all participants' efforts in this study, and the reviewers' valuable suggestions.

10. REFERENCES

[1] Chi, E.H. and Mytkowicz, T. Understanding the efficiency of social tagging systems using information theory. In *Proc. Hypertext 2008*, ACM, (2008), 81-88.

[2] Farooq, U., Kannampallil, T.G., Song, Y., Ganoe, C.H., Carroll, J.M. and Giles, L. Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics. In *Proc. GROUP 2007*, ACM, (2007), 351-360.

[3] Muller, M. J. Comparing tagging vocabularies among four enterprise tag-based services. In *Proc. GROUP 2007*, ACM, (2007), 341-350.

[4] Sen, S., Harper, F.M., LaPitz, A. and Riedl, J. The quest for quality tags. In *Proc. GROUP 2007*, ACM, (2007), 361-370.

[5] Sen, S., Lam, S.K., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F.M. and Riedl, J. tagging, communities, vocabulary, evolution. In *Proc. CSCW 2006*, ACM, (2006), 181-190.

[6] Pirolli, P. and Card, S. Information foraging in information access environments. In *Proc. CHI 1995*, ACM, (2008), 51-58.

[7] Pirolli, P. and Card, S. Information Foraging. *Psychological Review*, 106, (1999), 643-675.

[8] Russell, D., Stefik, M., Pirolli, P. and Card, S. The cost structure of sensemaking. In *Proc. CHI 1993*. ACM, (1993), 269-276.

[9] Carroll, J. M. Making use: a design representation. *Communications of the ACM*, 37, 12, (1994), 28-35.

[10] Golder, S.A. and Huberman, B.A. Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 32, 2 (2006), 198-208.

[11] Cattuto, C., Baldassarri, A., Servedio, V.D.P. and Loreto, V. Vocabulary growth in collaborative tagging systems. *Arxiv preprint arXiv:0704.3316,* (2007).

[12] Cattuto, C., Loreto, V. and Pietronero, L. Collaborative Tagging and Semiotic Dynamics. *Arxiv preprint cs.CY/0605015,* (2006).

[13] Macgregor, G. and McCulloch, E. Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review*, 55, 5, (2006), 291-300.

[14] Chi, E.H. and Mytkowicz, T. Understanding navigability of social tagging systems. In *SigCHI alt.chi* (2007).

[15] Fu, W. T. The Microstructures of Social Tagging: A Rational Model. In *Proc. CSCW 2008*, ACM, (2008), 229-238.

[16] Furnas, G.W., Landauer, T.K., Gomez, L.M. and Dumais, S.T. The vocabulary problem in human-system communication. *Commun. ACM*, 30, 11, (1987), 964-971.

[17] Furnas, G.W., Fake, C., von Ahn, L., Schachter, J., Golder, S., Fox, K., Davis, M., Marlow, C. and Naama, M. Why do tagging systems work? *Ext. Abstracts CHI 2006* ACM, (2006), 36-39.

[18] Shirky, C. *Ontology is overrated*. (2005). http://www.shirky.com/writings/ontologyoverrated.html.

[19] Bush, V. As We May Think. *Atlantic Monthly*, 176, 1 (1945), 101-108.

[20] Collins, A. M. and Loftus, E. F. A spreading-activation theory of semantic processing. *Psychological Review*, 82, 6, (1975), 407-428.

[21] Norman, D. A. Cognitive Artifacts. In Carroll J.M. (eds) *Designing Interaction*, Cambridge University Press, Cambridge, 1990.

[22] Simon, H. A. A behavioral model of rational choice. *Quarterly Journal of Economics*, 69, (1955), 99-118.

[23] Klein, G. and Klinger, D. Naturalistic Decision Making. *Human Systems IAC Gateway*, 11, 3, (1991), 16-19.

[24] Weick, K. E. *Sense-making in organizations*. Sage Publications, Newbury Park, CA, 1996.

[25] Wittgenstein, L. *Philosophical Investigations*. Blackwell Publishing, MA, 1953.

[26] Berger, P. L. and Luckmann, T. *The Social Construction of Reality: A treatise in the Sociology of Knowledge*. Anchor Books, NY, 1966.

[27] Halpin, H., Robu, V. and Shepherd, H. The complex dynamics of collaborative tagging. In *Proc. WWW 2007,* ACM, (2007), 211-220.

[28] Resnick, P. and Varian, H. R. Recommender systems. *Communications of the ACM*, 40, 3, (1997), 56-58.

[29] Vu, K. P. L., Hanley, G. L., Strybel, T. Z. and Proctor, R. W. Metacognitive Processes in Human-Computer Interaction: Self-Assessments of Knowledge as Predictors of Computer Expertise. *International Journal of Human-Computer Interaction*, 12, 1, (2000), 43-71.

[30] McDonald, D. W. and Ackerman, M. S. Expertise recommender: a flexible recommendation system and architecture. In *Proc. CSCW 2000,* ACM (2000), 231-240.

[31] Dourish, P. and Bellotti, V. Awareness and coordination in shared workspaces. In *Proc CSCW 1992)*, ACM, (1992), 107-114.

[32] Gutwin, C., Roseman, M. and Greenberg, S. A usability study of awareness widgets in a shared workspace groupware system. In *Proc. CSCW 1996,* ACM, (1996), 258-267.

[33] Grudin, J. Why CSCW applications fail: problems in the design and evaluation of organization of organizational interfaces. In *Proc. CSCW 1988,* ACM, (1988), 85-93.

[34] Rodden, K. and Wood, K. R. How do people manage their digital photographs? In *Proc. CHI 2003,* ACM (2003), 409-416.

[35] Slamecka, N. J. and Graf, P. The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology*, 4, 6, (1978), 592-604.

[36] Tory, M. and Moller, T. Evaluating visualizations: do expert reviews work? *Computer Graphics and Applications, IEEE*, 25, 5, (2005), 8-11.