

With a Little Help from My Friends: Examining the Impact of Social Annotations in Sensemaking Tasks

Les Nelson¹, Christoph Held², Peter Pirolli¹, Lichan Hong¹, Diane Schiano¹, and Ed H. Chi¹

¹Palo Alto Research Center (PARC)

3333 Coyote Hill Road

Palo Alto, CA 94304, USA

{lnelson, pirolli, hong, dschiano, echi}@parc.com

²Knowledge Media Research Center

Konrad-Adenauer-Str. 40

72072 Tübingen / Germany

c.held@iwm-kmrc.de

ABSTRACT

In prior work we reported on the design of a social annotation system, SparTag.us, for use in sensemaking activities such as work-group reading and report writing. Previous studies of note-taking systems have demonstrated behavioral differences in social annotation practices, but are not clear in the actual performance gains provided by social features. This paper presents a laboratory study aimed at evaluating the learning effect of social features in SparTag.us. We found significant learning gains, and consider implications for design and for understanding the underlying mechanisms in play when people use social annotation systems.

Author Keywords

User studies, social annotation systems, social sensemaking

ACM Classification Keywords

H5.2 User Interfaces – Graphical User Interfaces.

INTRODUCTION

Social annotation systems such as SparTag.us [2] and del.icio.us have been designed to encourage individual reading and annotation behaviors that, when shared, accumulate to build collective knowledge spaces. In a recent longitudinal classroom study, Kalnikaite and Whittaker [3] report correlations suggesting a positive impact of social annotations on learning, but a causal relation remains to be shown. Other studies suggest that while people often seek out previously highlighted and annotated content [4], they can be adversely affected by inappropriate annotations, even when warned about such adverse effects [6]. In the present study, we test whether annotations designed to represent the output of a subject-matter expert accelerates the learning of users with access

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CHI 2009, April 4–9, 2009, Boston, MA, USA.

Copyright 2009 ACM 978-1-60558-246-7/08/04...\$5.00

to those annotations during a given sensemaking task.

Social Annotation in SparTag.us

SparTag.us uses keyword tags and highlights as a means to collect paragraphs of interest in web pages. A Click2Tag interface offers a low-cost option for the user to annotate paragraphs using simple interactions made directly on the content being read. SparTag.us automatically extracts the annotated paragraphs from the page and inserts them into a system-created notebook, along with the URL of the page. Moreover, users may subscribe to and follow the annotations of another user by designating that user as a friend. Figure 1 shows a portion of a friend's notebook as viewed by the user. The user's highlights are in yellow, and tags are in red. The friend's annotations are displayed in light blue with tags attached to the end of the paragraph. If there were multiple friends highlighting one paragraph, all the friends' highlights would have been aggregated.



Figure 1. A friend's notebook may be browsed or searched. A tag cloud emphasizes the most used tags.

STUDY SCOPE AND APPROACH

Our experimental approach to measure learning within a complex social annotation setting was to: (1) define an ecologically valid, realistic sensemaking task in a technical domain; (2) develop domain-specific knowledge tests as instruments to measure performance learning gains in that setting; and (3) vary the conditions to distinguish the impact of socially-constructed annotations.

The task we chose was a *task force* participation scenario, in which a person was given a general topic area, and had to research that topic and produce a document at the end. This is a general category of tasks familiar throughout the academic and professional world that includes the activities of task forces, work groups, classroom group projects, and so on.

Our experimental contrast compared three groups of participants who worked: 1) without SparTag.us (WS) but with traditional note-taking tools, 2) with SparTag.us only (SO), and 3) with SparTag.us with a Friend (SF). The conditions WS and SO were control conditions in which individuals read web content without access to others' annotations. To provide for an ecologically valid comparison, WS participants could take notes in MS Word or with pen and paper. In the SF condition, people independently read web content but also had access to social annotations created by an experimenter-simulated subject-matter expert (expressed as a SparTag.us 'friend').

Our hypothesis was that participants with access to tags and highlights made by an expert would perform better than participants without the same access. We thus evaluate performance measures between subjects in the experimental condition, SF, with those in control conditions, SO and WS.

METHOD

Participants

Participants ($N = 18$) were solicited from various sources, including participants from other studies, company interns and a local university job list. The only screening concerned having no previous use of the SparTag.us tool. As participants arrived, they were assigned to one of the three groups in round-robin order to obtain six participants per group. There was an even gender split. The average age was 28 (range 18-45). Six participants reported having some education in computing.

Sensemaking Domain and Task

We chose the study topic domain of "Enterprise 2.0 Mashups", which is a combination of the technology areas of "Enterprise 2.0" and "Web 2.0 Mashups". This choice required participants to find and understand many web pages because at the time of the study there was no single good source of information on the topic area. The area was relatively new, with visible activities in conferences and companies providing such capabilities.

Participants in all three groups were asked to find and read material in order to write reports on Enterprise 2.0 Mashups. There were two writing tasks, each specified by three questions that needed to be addressed. The writing tasks were aimed at eliciting what someone reasonably skilled in the area 'should be able to answer'.

The questions to be addressed in the writing tasks were derived from a survey of experts. Ten self-reported experts were solicited in a survey at an enterprise mashup industry

event (mashupcamp.com) and asked to name "key concepts" in the field. We clustered the 56 responses into six categories, leading to six questions for the writing tasks.

Expert Annotations

The way we constructed the SparTag.us friend for the study was to provide clear and succinct summaries of key content derived from social sources (del.icio.us aggregation and repeated expert elicitation on the topic). The overall organization of topics was then mapped to specific content that exemplified the topic. The representation of this knowledge was given *in-situ* to the source materials and through clipped, annotated collections of the relevant information with back links to the sources. The top 20 tags associated with the top 100 annotated URLs returned by a del.icio.us query on "enterprise mashup" were used as the target tag cloud in SparTag.us. A set of URLs covering these keywords (found by Google search with expert assessment of the resulting URLs in result order) were manually tagged using SparTag.us. A sufficient number of URLs were tagged with a given tag to produce the target tag cloud. This process created a repository of a friend persona named "mjones". Each SF participant was given access to this friend notebook at the start of learning.

Knowledge Pretest and Posttest

We measured learning by assessing domain-specific knowledge about enterprise 2.0 mashups prior to the task (pretest) and after (posttest), in all three groups. Two lists of 20 true-false questions were created, and each list was used as a Pretest for half the participants in each group and as a Posttest for the other half. The true-false questions were elicited from experts. Each list of 20 questions was designed to have an even distribution of easy and hard questions about enterprise mashups, as rated by 100 random people on Amazon Mechanical Turk [4] (together with control questions with 'obvious' answers to reduce spurious responses). Questions were designed to minimize prompting of participants' subsequent learning. Both tests were taken without access to tools or resources.

Procedure

A two-day test schedule was established, refined through pilot testing and applied to 18 participants covering all conditions. The first day involved a four-hour session of demographic and background survey, tool training, knowledge pretest, learning in the domain area, knowledge posttest, and essay writing. Training in SparTag.us involved a short video and then a guided walkthrough of features. Participants were given a brief written statement of learning objectives, instructing them to read from any sources and take notes as they felt appropriate regarding the definitions, standards, benefits, issues, and examples relating to the topic area. Participants had one hour of unsupervised learning, with a break for lunch, and then 50 more minutes of learning. The writing activity was separately prompted by a different set of questions and limited to 30 minutes. A

second session was held a week later involving a second writing task (40 minute limit).

Sessions were logged, including URLs visited, content scrolling, and words written. In debrief interviews, people were asked to talk about strategies used for learning, question answering, writing, and tool use. Participants were randomly assigned to computers configured with Windows XP and Mozilla Firefox (2.0.0.16).

EXPERIMENT RESULTS

We now outline the performance and behavioral measures taken in each condition.

Learning: Gains on the Knowledge Test

One measure of learning was obtained by computing gains in test scores from the Pretest to Posttest. Specifically, gain scores were calculated as:

$$Gain = \frac{\text{Posttest score} - \text{Pretest score}}{\text{Max score} - \text{Pretest score}}$$

This score has the advantage of normalizing the observed gain (the numerator) against the amount of possible learning that could be achieved (the denominator).

The mean gain scores were: SF group, $M=0.46$, $SD=0.22$; SO group, $M=0.13$, $SD=0.32$; WS group, $M=0.27$, $SD=0.23$. An analysis of covariance showed a significant effect of learning group, $F(2, 16) = 5.91$, $p < .05$, with the SF group showing significantly greater gains than the SO group, $t(16) = 4.66$, $p < .0005$, and the WS group, $t(16) = 3.93$, $p = .001$. The WS and SO groups were not significantly different.

Regression analyses identified two background questions showing significant relationships to the gains scores across all groups (other measures are correlated with these factors):

- IT Learning: I enjoy learning about information technology and new developments in this field
- Web Use: On average, how many hours per week do you spend on the World Wide Web?

Interest in IT Learning significantly reduced the gains, $t(16) = 4.30$, $p = .005$, whereas increasing web use had a positive effect on gain scores, $t(16) = 6.57$, $p = .02$.

Overall, the largest impact on gains was the presence of the expert friend annotations in the SF condition. Using SparTag.us without access to these expert annotations (SO) did not yield any learning gains over the use of the standard note-taking tools (WS).

Writing: Use of Domain Words

Although not statistically significant, the performance measure of the use of domain terms in writing show a trend towards favoring the SF condition (Table 1). This pattern was consistently seen after each round of six participants was run and preliminary analyses were made.

Domain words are determined by using (1) tags associated with URLs matching the del.icio.us query “enterprise mashup” as well as (2) terms gathered in each of three sessions of a domain conference event where the topic was “What is an enterprise mashup,” and (3) further including synonyms of those terms (e.g., for the general term “vendor”, a relevant specific term would be “IBM”).

| Group | All Words | SD | Domain Words | SD |
|-------|-----------|--------|--------------|-------|
| SF | 549.92 | 207.01 | 141.50 | 46.27 |
| SO | 528.92 | 202.08 | 136.00 | 58.68 |
| WS | 459.67 | 174.32 | 117.00 | 48.17 |

Table 1. People using SparTag.us used more domain words.

Trends in Behaviors

There is individual variability in the behavioral data, though also trending in interesting directions (Table 2). We see that on average SF participants visited fewer URLs, but spent more reading time on those they visited. The SparTag.us users (SF and SO) scrolled content more on average, indicating more reading was occurring. Again, these patterns were consistent after each round of participants.

| Group | URL Visits | | Time on URL | | Scroll on URL | |
|-------|------------|------|-------------|------|---------------|-------|
| | Mean | SD | Mean (sec) | SD | Mean | SD |
| SF | 59 | 23.7 | 144.5 | 73.0 | 642 | 110.3 |
| SO | 71.2 | 25.5 | 128.2 | 56.1 | 711 | 235 |
| WS | 79.3 | 35.9 | 87.3 | 24.0 | 476.3 | 132.9 |

Table 2. Trends suggest different reading behaviors between conditions may be measurable with more observation.

Qualitative Data

Participants in the SF condition all show activity with the simulated friend. All SF participants looked at the friend’s repository early (three as their first site, two after a visit to a Wikipedia page first, and one after two visits to Wikipedia). In debriefing (Table 3), not all showed appreciation for the actual friend-annotated content, particularly when they believed that the questions of the learning task were not directly addressed by the seeded content. Each did consider the information offered there, and followed it or found alternatives they thought were more appropriate.

Summary of Results

SF learning gains were significantly higher than those of SO and WS, with SO not significantly different from WS. Activity with the friend’s seeded information is visible during learning. Trends indicate SF participants visited fewer URLs, spent more time reading/scrolling those, and wrote more in the essay.

IMPLICATIONS

One design implication of these results is that designing for social processes can produce measurable learning benefits.

This work suggests that expertise delivered by social annotation mechanisms is helpful to users in learning unfamiliar domains. In getting this result, we encountered the tradeoff between ecologic validity and statistical power in the experiment design. Our priority was to increase ecologic validity, take multiple measures enabling the analysis of covariance, and control some sources of variability using a systematic procedure and the constant set of stimuli (e.g., the ‘friend’ being a constant set of social annotations). For example, by including IT Learning and Web Use as individual difference covariates in the reported analysis of covariance we were able to partial out a good portion of the variance due to participant differences and thereby reduced our error of measurement to get the significant result that was reported. So we gained statistical power through additional subject measurements.

One line of further inquiry that may guide us towards testable explanations of the gains seen using SparTag.us may be found in ‘schema theory’. If a user of SparTag.us does not already have the necessary background knowledge or schema, then she will be forced to obtain the background knowledge during reading. ‘Schema’ is often used to refer to units of knowledge that individuals internalize that contain elements of related information and provide a kind of structure for future information [1]. A possible hypothesis here is that the high cognitive load of having to produce the schema in addition to doing the reading task reduces learning performance. The SparTag.us expert notebook functions as a kind of scaffold for learning, and serves as an advantage over participants without the scaffold. The notebook not only provides sample reading material that might be useful, but the sample paragraphs and tag cloud also provides for a kind of preview of sample terms that might serve as navigational signposts. These previews are all different kinds of schemata, and address possible organizational/abstraction processes at work.

ACKNOWLEDGEMENTS

The research was funded in part by support from Office of Naval Research Contract No. N00014-08-C-0029 to Peter Pirolli. We thank the study participants and we thank Gregorio Convertino for critique of drafts of this Note.

CONCLUSION

In this paper, we describe a first step in grounding our understanding of the impacts of one social annotation technology on people’s information foraging practices. People with access to well structured artifacts left by others do show measurable learning improvements. Our future plans are to explore the social sensemaking processes in action and expand the unit of analysis to groups collaborating synchronously and asynchronously, examining the timeliness and manner of delivery of expertise in the context of ongoing reading activities; and looking at the role of expertise and social familiarity in perceiving annotations. We are currently investigating

| Id | Quote about Friend |
|-----|---|
| SF1 | I checked on that [Friend's Repository], but I couldn't find the kind of information I was looking for there... I went back to the Friend afterwards [after finishing first pass in the writing] and found another thing to add [to the writing]. |
| SF2 | I started off with the recommendations for websites that my friend saved taking into account that he seemed to be an expert. |
| SF3 | I was knowing that a person had done this, even if he's not my friend, but he's just like a somewhat trustful person... just the fact that it was a human being made a huge difference to me. But that being said, I never like looked at his tags, because he had a different tagging system, ...it was like this word, you know, "data sources", "client"... they were like too broad ... some of these could've been useful to me, maybe "SLA" and "SOA". But "software"? It's not what I would've thought of... |
| SF4 | The first time I checked the friend was about 20 minutes. Because first I wanted to get an idea myself, so I could evaluate how good my friend is. I found only very few information that was useful I hadn't found already. So I used him not very much. But still, if I saw an article in his listing that I also tagged that made me feel better. If Google thinks it's important, my friend thinks it's important, then it must be important. |
| SF5 | That one [Friend's notebook] I didn't find particularly helpful. Because for one thing this computer terminology is totally new to me. So I had to go back and read. I got what "RSS" stands for and all that stuff. I did find that after that I began to appreciate [the friend] better and especially a couple of places with making money. I found that interesting and helpful. |
| SF6 | I mean I did sort of surprise myself using the Friends Web links, because that was new. I did find myself highlighting a lot, but again I think it's more like the, it's more of the compulsive behavior, because when I highlight I really want to like physically write that stuff down. |

Table 3. In debrief all participants reported giving attention to the seeded information with varying degrees of appreciation.

different analytical approaches we may apply to the data collected to better understand the role of social annotations.

REFERENCES

1. Anderson, R.C. and Pearson P.D. A schema-theoretic view of basic processes in reading comprehension. In P.D. Pearson (ed.), *Handbook of Reading Research*, 255-291. New York: Longman, 1984.
2. Hong, L., Chi, E.H., Budiu, R., Pirolli, P., and Nelson, L., SparTag.us: A low cost tagging system for foraging of web content, *Proc. AVI'08*, ACM, pp. 65-72, 2008.
3. Kalnikaite, V. and Whittaker, S., Social summarization: Does social feedback improve access to speech data? To appear in *Proc. CSCW 2008*, ACM, Nov. 2008.
4. Kittur, A., Chi, E.H., and Suh, B. Crowdsourcing user studies with Mechanical Turk. *Proc. CHI 2008*, ACM Pres (2008), 453-456, 2008.
5. Marshall, C. Annotation: From paper books to the digital library. *Proceedings of the ACM Digital Libraries '97 Conference*, Philadelphia, PA, July 1997, 131-140.
6. Silvers, V.L. and Kreiner, D.S., The effects of pre-existing inappropriate highlighting on reading comprehension. *Reading Research and Instruction*, 36, 3, 217-223, 1997.

