Tag Clouds and Keyword Clouds: evaluating zero-interaction benefits

Mathew J. Wilson
FIT Lab
Computer Science Department
Swansea University
Swansea, SA2 8PP, UK
csmathew@swansea.ac.uk

Max L. Wilson
FIT Lab
Computer Science Department
Swansea University
Swansea, SA2 8PP, UK
m.l.wilson@swansea.ac.uk

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.
ACM 978-1-4503-0268-5/11/05.

Abstract
Tag clouds are typically presented so that users can *actively* utilize community-generated metadata to query a collection. This research investigates whether such keyword clouds, and other interactive search metadata, also provide measureable *passive* support for users who do not directly interact with them. If so, then objective interaction-based measurements *may not be* the best way to evaluate these kinds of search user interface features. This paper discusses our study design, and the insights provided by a pilot study that led to a series of improvements to our study design.

Keywords
Search, Sensemaking, Tag clouds, Keyword clouds.

ACM Classification Keywords
H5.2. User Interfaces (e.g., HCI): Prototyping. H5.4. Hypertext/Hypermedia: Navigation

General Terms
Design, Human Factors

Introduction
Tag clouds are often provided to help users issue new or improved queries with community-generated metadata, and so their benefits are typically measured

by clickthrough or success in finding results. Our research, however, aims to investigate whether keyword clouds help us passively, without direct user interaction, to make sense of information spaces, and potentially learn as we search. If true, such findings would have significant implications for the evaluation of advances made in search user interface design, as they may have a significant impact on searchers but *without* creating any observable and objectively measureable interactions. Marti Hearst, for example, suggested that the quality of facets are more important than the way we interact with them [5]. Our research should help to validate and quantify these distinctions.

In the following sections we first summarise this position in related work and then describe our initial pilot study into the passive benefits of user interface design issues. We describe three user interface conditions that include interactive variations of a keyword cloud within a typical web Search Engine Result Page (SERP). We then discuss the challenges faced in analysing 'learning' and present our plans for completing this work.

Related Work

Much recent research into information seeking, beginning more formally with the special issue in 2006 [11], has focused on users in exploratory conditions, with White and Drucker [10] noting that up to 80% of web searches involve at least some exploratory behaviours. It is within these terms that we have begun investigating how user interfaces passively help users make sense of information as they try to understand their problem or the unfamiliar domain that they are working within. Dervin [3] described the human process of sensemaking as trying to bridge a newly identified gap in knowledge. Models of information seeking, however, typically cover all forms of searching behaviour, and the study of Exploratory Search focuses more firmly on the cases where people have to investigate and learn [8]. Notably, Bates' research into search tactics [1] highlighted several idea and reflective tactics, such as 'surveying ones options'.

In studying exploratory search and sensemaking, studies often create a mix of 'known tasks' and 'exploratory tasks' and measure them in different ways. While simple known-target tasks are often measured by how quickly users can find information, Capra et al [2], for example, did not measure time for exploratory tasks, noting that a good system may encourage people to explore for longer. Similarly, in the evaluation of a system called MrTaggy [7] that provides two tag clouds to help people search for information, both quality of subsequent summary writing and *high* subjective reports of cognitive load were deemed as positive results. In this case, high cognitive load was only deemed positive if the quality of subsequent topic reports was also strong.

MrTaggy provided tag clouds to be *actively* used for searching, where as the work we describe below compares conditions where users can and cannot interact with keyword clouds, and our early pilot study results found positive evidence that people may experience roughly equal improved learning in both conditions, and very little interaction when they could anyway. Some secondary insights from our previous research indicated that well structured inter-relating facets may help people make sense of information spaces [12]. We saw people able to recall additional facts from metadata presented by the user interface

but did not directly measure whether users had interacted with these items. Consequently, this research aims to separate out information from possible interaction to measure any effect.

Hearst and Rosner [6] studied tag clouds in social tagging sites and concluded that people perceived them to be valuable because they provided personal or social content. Other research has studied how tag clouds might be used to search and explore. Sinclair and Cardew-Hall [9] compared a keyword search interface with a tag-cloud interface indicating that a tag cloud was more useful during general searches but was not an effective way of searching a web archive. Gwizdka [4] studied how tag clouds, created from delicious tags, were used by different cognitive-types of users and during different tasks, indicating that tag clouds were better for people with verbally-oriented cognitive styles, but didn't save them much searching time.

Study Design

Our motivating hypothesis was that keyword clouds[1] provide a lot of benefit for searchers but not necessarily for interactively searching. We have designed a study, and performed a pilot, that should show the benefits for sensemaking provided by both the presence and the use of a keyword cloud. Three controlled user interfaces have been developed, which differed only by the form of keyword cloud they incorporate: 1) an interactive keyword cloud that issues new queries when a keyword is clicked (Figure 1), 2) a static keyword cloud that cannot be clicked, and 3) no keyword cloud. We do not mean to suggest that keyword clouds should not be interactive, but aim to show that they provide significant benefits without interaction.

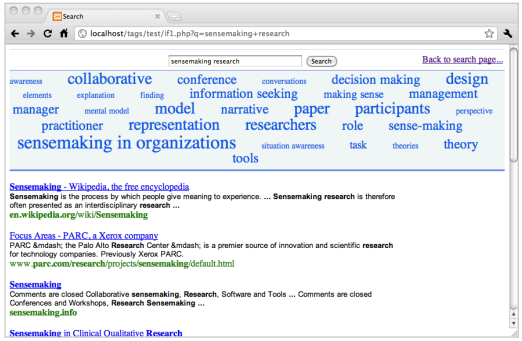

Figure 1. Screenshot of interactive keyword cloud condition.

Keyword clouds are generated dynamically using the Yahoo! Boss API[2], which returns associated keywords with each search result for a given query. Keyword clouds are based on the first 100 results, and remain consistent by query and across pagination. To control for load-time, all three interfaces generate a keyword cloud algorithmically and only vary by how they subsequently display keyword clouds. To control for number of results visible above the fold (the lowest vertical part of the interface shown before scrolling), in case this affects learning, the prototypes maintain a fixed-sized space in the user interface regardless of the size or presence of the keyword clouds generated.

Participation takes approximately 1 hour including acquiring informed consent and demographic details. Participants will perform 3 sensemaking tasks, one in each condition, where order is counterbalanced. To

[1] Although much research has focused on tag-clouds, as generated by social media systems, we studied the metadata that search engines had about search results, and so are using the term 'keyword cloud' herein.

[2] http://developer.yahoo.com/search/boss/

establish a measure of existing knowledge, similar to the MrTaggy study, we will give participants 5 minutes to a) judge their current expertise on the task topic out of 5, b) write a short summary (from memory), and c) list related keywords. Participants will then have 5 minutes to search the web with one of the user interfaces. After searching, to measure some form of learning, participants will be given 5 minutes to write a new summary (from memory) and list keywords. After repeating these 15 minutes in each interface condition, participants will be given a final survey and interview.

Time will not be measured, but instead controlled, and we intend to focus on the amount learned during the sensemaking task, while measuring a) the number of facts listed, b) the number of related keywords listed, and c) the quality of the summary. The system also logs all queries, words in the keyword clouds, keyword hovering, keyword clicking, and results viewed.

6 broad sensemaking tasks have been generated on the topics of: childproofing, dog purchasing, home entertainment systems, E-book readers, anti-virus software, and web-applications. We want to gain some insight into whether existing knowledge affected the value of keyword clouds, and so, in a counterbalanced order, participants will be asked to pick tasks by alternating between high or low existing knowledge.

Initial Pilot Study

As our results will be highly dependent on making sure we record learning, in some form, as accurately as possible, we ran a pilot study on our methods. 12 participants, 10 male, took part in the methodology described above. This number of participants allowed us to test each study condition and task description evenly. We acknowledge here that the sample for this pilot investigation was biased towards students under the age of 23, but the pilot was mainly for testing our methodology, measurements, and plans for analysis.

Initial Findings

From our trial analysis of the pilot study results it would appear that there is some support for our hypothesis. Focusing on the post-study summaries, we found that the quality of summaries produced from using the interactive keyword cloud was not significantly different from that of the non-interactive keyword cloud. We saw a marginally significant ($p=0.063$, $F(35)=3.13$) ANOVA result in quality of the post-task summaries. A post-hoc Tukey test indicate that the differences were between the control and each experimental condition (non-interactive: $p<0.05$ $t(11)=2.34$; interactive: $p<0.05$ $t(11)=2.16$), but not between the two experimental conditions. This indicates that the presence of the tag cloud was more important than using it for search.

It is worth noting that in the pilot study every participant showed some active interaction with the keyword cloud when available. This leads us to believe that, regardless of the interactions that occurred with the keyword cloud, the benefit that *interaction* provides toward learning is minimal. These initial findings appear to suggest that our full study will provide significant results in support of our hypotheses.

Lessons Learned and Changes

Measurements/Recordings. In the pilot, participants were asked to handwrite their summaries on paper. This created unexpected, although potentially foreseeable, logistical problems such as the text being unintelligible, the amount of space needed for the

varying size of handwriting and, with the task being timed, the speed at which the participants write. These factors will be easily overcome by having participants type their summaries in our main study.

Casual observational analysis of the summarization tasks while being performed also made it obvious that participants would often forget or overlook the section that required them to list keywords. We undertook a closer, albeit brief, analysis of the keywords that were listed to see whether this stage of the task should be improved or removed. When comparing the keywords listed before the search phase to those listed after we found that the number of keywords rarely changed, but the quality of keywords appeared to potentially improve. As this provided little additional insight into learning from the summaries, we decided to simplify the method and exclude the keyword stage.

Analysis. Since the intention of the study was to look at learning, and we considered three approaches to analysing the written summaries. One was to look at the overall length of the summary as a quick telling of how much the participant had learned during the study period. Second, we simply counted the number of facts the participant had given to represent the validity of what the participant had learned, as in the study performed by Wilson et al [12]. Finally, we looked at the overall quality of the summary. Although the first two were more objective, it quickly became clear that both were unsuitable for this study, where some participants would fill their summaries with information that, while factual, was redundantly obvious ("A labrador is a dog"). Consequently, we have used this pilot study data to generate a careful likert scale to judge the quality of summaries, and will soon establish a high inter-relater reliability with Cohen's Kappa.

System Improvements. The system itself was designed with simplicity in mind, but feedback from pilot participants highlighted some limiting flaws while participating. First, participants noted the poor quality of keyword clouds. In an attempt to improve the keyword clouds, we revised the system to combine returned Yahoo! keywords with common terms in the text-snippets linked to the returned results, as shown in Figure 2. To maintain reasonable performance this new process only parses the first 50 results, rather than the top 100, but produces much richer keyword clouds. Further, additional filtering was included to remove repeated or redundant terms from the cloud.

Figure 2. Comparison of the old style keyword cloud (top) and improved cloud (bottom).

For analysis, the system logs several forms of interaction to a database, but upon looking at the results, it became obvious that this was the biggest problem with the system. All of the logging was done in plaintext with timestamps that was then exported to an Excel spreadsheet. To make analysis easier, each user

is now assigned a randomly generated cookie ID for the duration of the study. Further, a web interface was created that allows the information to be displayed in an easy-to-read timeline and also to rebuild the keyword clouds that were displayed during each logged interaction.

Conclusions and Future Work

Following the design of our study, the inclusion of a pilot was important as the approach to analyzing learning is neither objective nor easily predictable. The pilot study allowed us to trial different approaches to our analysis as well as provide us with real-world experience with the interfaces that would lead to improvements. As well as this practical feedback, however, our trial analysis also found some initial support for our hypothesis that leaves us optimistic about the forthcoming study.

Shortly, we will begin performing the main study, using the methodological enhancements described above, on approximately 36 participants (similar to the MrTaggy study [7]). If our hypotheses are correct, our findings should have a significant impact on the way we *evaluate* search user interfaces, placing emphasis on the implicit and less-tangible benefits that such features can provide. Further, this places importance on carefully choosing *and* representing metadata in order to support exploratory search. We hope in the future to explicitly delineate the contributions of both information and interaction in search user interfaces.

References and Citations

[1] Bates, M.J., Information search tactics. *Journal of the American Society for Information Science*, *30*(4), 205-214. 1979.

[2] Capra, R., Marchionini, G., Oh, J.S., Stutzman, F. and Zhang, Y., Effects of structure and interaction style on distinct search tasks. In *Proc. JCDL 2007*, 442-451. 2007

[3] Dervin, B., Foreman-Wernet, L. and Lauterbach, E. Sense-Making Methology Reader: Selected Writings of Brenda Dervin. Hampton Press, 2003.

[4] Gwizdka, J., What a Difference a Tag Cloud Makes: Effects of Tasks and Cognitive Abilities on Search Results Interface Use. *Information Research*, *14*(4), Paper 414. 2009.

[5] Hearst, M.A., Clustering versus faceted categories for information exploration. *Communications of the ACM*, *49*(4), 59-61. 2006.

[6] Hearst, M.A. and Rosner, D., Tag Clouds: Data Analysis Tool or Social Signaller? In *Proc. HICSS 2008*. 2008

[7] Kammerer, Y., Nairn, R., Pirolli, P. and Chi, E.H., Signpost from the masses: learning effects in an exploratory social tag search browser. In *Proc. CHI 2009*, 625-634. 2009

[8] Marchionini, G., Exploratory search: from finding to understanding. *Communications of the ACM*, *49*(4), 41-46. 2006.

[9] Sinclair, J. and Cardew-Hall, M., The folksonomy tag cloud: when is it useful? *Journal of Information Science*, *34*(1), 15-29. 2008.

[10] White, R.W. and Drucker, S.M., Investigating behavioral variability in web search. In *Proc. WWW 2007*, 21-30. 2007

[11] White, R.W., Kules, B., Drucker, S.M. and schraefel, m.c., Introduction. *Communications of the ACM*, *49*(4), 36-39. 2006.

[12] Wilson, M.L., André, P. and schraefel, m.c., Backward Highlighting: Enhancing Faceted Search. In *Proc. UIST 2008*, 235-238. 2008