Keyword Clouds: Having Very Little Effect on Sensemaking in Web Search Engines

Mathew J. Wilson
Future Interaction Technology Lab
College of Science,
Swansea University, UK
csmathew@swansea.ac.uk

Jonathan Hurlock
Future Interaction Technology Lab
College of Science,
Swansea University, UK
csjonhurlock@swansea.ac.uk

Max L. Wilson
Future Interaction Technology Lab
College of Science,
Swansea University, UK
m.l.wilson@swansea.ac.uk

CHI'12, May 5–10, 2012, Austin, Texas, USA.
ACM 978-1-4503-1016-1/12/05.

Abstract

Tag clouds are typically presented so that users can *actively* utilize community-generated metadata to query a collection. This research investigates whether such metadata representations also provide *passive* support for sensemaking without any direct interaction. Previous work reported potentially significant results from a pilot study of three variations of *keyword* cloud support (interactive, non-interactive, and absent), built from related query terms. Our full study, however, found no significant differences in learning across the three conditions. We concluded that the sensemaking and learning mainly occurred outside of the search engine, where the keyword cloud no longer provided support. Our future work will study the passive support that may be provided by keyword clouds in more integrated systems like digital libraries.

Keywords

Search; Sensemaking; Tag clouds; Keyword clouds

ACM Classification Keywords

H.5.2 [Information Interfaces And Presentation]: User Interfaces - Prototyping; H.5.4 [Information Interfaces And Presentation]: Hypertext/Hypermedia - Navigation;

General Terms

Design, Human Factors

Introduction

Tag clouds help users to issue new or improved queries with community-generated metadata, and so their benefits are typically measured by click-through or search success. Our research, however, aims to investigate whether such representations help searchers passively, without direct interaction, to make sense of information spaces, and potentially learn as they search. If true, such findings would have significant implications for the evaluation of metadata representations in search user interfaces, as they may have a significant impact on searchers but *without* creating any observable, measureable interactions. Our previous work described the results of a pilot study, which showed initial support for this hypothesis [15]. This paper, however, describes the full study, where we were unable to find support for our hypothesis.

Related Work

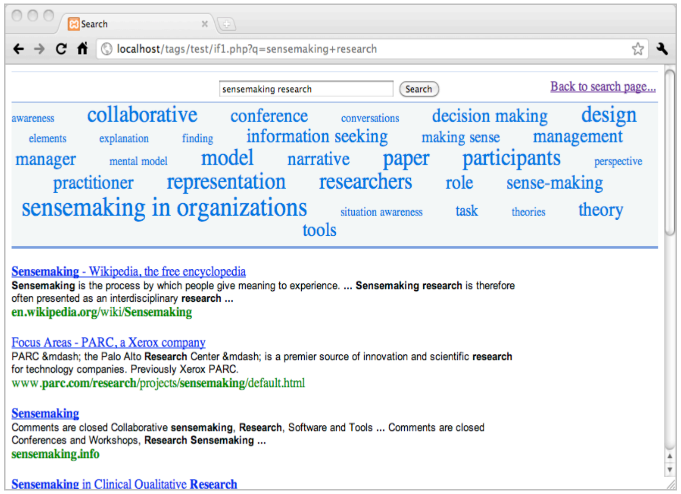
Much work has focused on the development and use of tag clouds in information systems. Schrammel et al., for example, compared different structure of tag clouds, indicating that tag clouds should be organized to place semantically similar tags close together [10]. Hearst and Rosner [7] studied tag clouds 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 [12] 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 [6] studied how tag clouds, created from delicious[1] 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.

Like many systems that provide metadata to help people search, a system called MrTaggy provided tag clouds to be *actively* used for searching [8]. They found that the tag-based system helped users to write better summaries, find more related domain terms, and experience higher cognitive load while working. The authors concluded that this increased cognitive load was good because these participants wrote better summaries. Rivadeneira et al. [9] believe that tag clouds can provide support for various types of task including searching, browsing, recognition and "gisting". The aim of this work is inline with this last type – to find out whether metadata representations can help people to learn *passively* as they search by giving them a general impression of the underlying content. In studying support for sensemaking and learning, for example, Sharma et al. found that participants who were provided with an initial framing for a topic, were able to learn faster and produce better presentations [11]. Further, secondary insights from Wilson et al. indicated that well structured facets may help people make sense of information spaces [16].

Many have also studied the process of exploring and making sense of information directly. Dervin [5] described the human process of sensemaking as trying to bridge a newly identified gap in knowledge. Similarly, the study of 'Exploratory Search' focuses

[1] http://delicious.com/

more firmly on the cases where people have to investigate and learn [13]. To investigate 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. [4], for example, did not measure time for exploratory tasks, noting that a good system may encourage people to explore for longer. Other studies, such as MrTaggy, constrain the time for learning and evaluate it by the subsequent quality of written summaries produced by participants.



localhost/tags/test/if1.php?q=sensemaking+research

sensemaking research

Sensemaking - Wikipedia, the free encyclopedia
Sensemaking is the process by which people give meaning to experience. ... **Sensemaking research** is therefore often presented as an interdisciplinary **research** ...
en.wikipedia.org/wiki/Sensemaking

Focus Areas - PARC, a Xerox company
PARC &mdash; the Palo Alto **Research** Center &mdash; is a premier source of innovation and scientific **research** for technology companies. Previously Xerox PARC.
www.parc.com/research/projects/sensemaking/default.html

Sensemaking
Comments are closed Collaborative **sensemaking**, **Research**, Software and Tools ... Comments are closed Conferences and Workshops, **Research Sensemaking** ...
sensemaking.info

Sensemaking in Clinical Qualitative Research

Figure 1: The keyword cloud enhanced Search Engine Results Page (SERP) used for the Interactive Cloud Condition.

Study Design

Our motivating hypothesis was that keyword clouds[2], as an example form of overview metadata, passively support sensemaking for searchers beyond the act of interactively searching. Consequently, a study was designed that would separate and then show the benefits for sensemaking provided by a) the presence and b) the use of a keyword cloud. Three controlled user interfaces were developed, which differed only by the form of keyword cloud they incorporated: 1) an Interactive Keyword Cloud condition (IntC) (**Figure 1**) that issued new queries when a keyword is clicked, 2) a Static Keyword Cloud condition (StaC) that could not be directly used to issue new queries, and 3) a condition with no keyword cloud (NoC). These keyword clouds were built using key terms returned with the top 100 results from the Yahoo BOSS API[3].

The aim of this study concept was not to show that a keyword cloud should not be interactive, but to determine whether it is the presence or the use of a keyword cloud that provides significant support for searching and sensemaking. Knowing the outcome of this tells us more about how tag clouds, or other forms of metadata like facets, should be designed according to how they are used. Should both the Interactive and Static keyword clouds provide significant gain for sensemaking, then we can conclude that it is more important to consider the metadata that they present. If a keyword cloud provides benefit only for the Interactive condition, then it is more important to make sure it is the interaction with keyword clouds, or similar forms of metadata, that needs to be carefully considered. The study described below was refined based upon the experiences of a pilot study [15].

Procedure

Participation took approximately 1 hour, including acquiring informed consent and demographic details. During the study, participants performed three 15-minute sensemaking tasks, one with each of the interface conditions. The exposure to interface condition was counterbalanced across all participants using the Latin-Square method. To establish a measure of existing knowledge, similar to the MrTaggy study, the 15 minutes began by participants spending up to

[2] Although much research has focused on social tags, we studied related terms from search results, and so are using the term ‘keyword cloud’ herein.

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

five minutes writing a short summary about the topic. Participants then had five minutes to search the web with one of the user interface conditions. After searching, to measure learning, participants spent the last five minutes (of 15) writing a new summary (without the search interface). After repeating these 15 minutes in each interface condition, participants were given a final survey and a short debriefing interview. While spending just five minutes researching may be considered insufficient, it was limited due to the number of tasks we asked participants to perform in the one-hour session.

Tasks

6 broad sensemaking tasks were generated on the topics of: childproofing, dog purchasing, home entertainment systems, E-book readers, anti-virus software, and web-applications. We wanted to gain some insight into whether existing knowledge affected the value of keyword clouds, so participants were asked to rate their current knowledge of each topic out of 7. Then, in a counterbalanced order, participants were issued tasks by alternating between high or low existing knowledge. Allowing participants to rate their own knowledge on a set of tasks allowed us to better control for existing knowledge (high or low) rather than have it as an uncontrolled or even confounding variable in our results. Tasks were presented to participants using a Simulated Work Task [3].

Pilot Findings

One concern for the methodology was whether the two experimental conditions (IntC and StaC) would have a significant enough impact on learning. As reported in an extended abstract [15], our pilot study results did appear to show some support for our hypothesis. Focusing on the post-study summaries, but only using a relatively naïve single likert-scale measure of overall quality, we found that the quality of summaries produced using the IntC was not significantly different from the StaC condition. 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 indicated that the differences were between the baseline NoC condition and the experimental conditions (NoC vs StaC: $p<0.05$ $t(11)=2.34$; NoC vs IntC: $p<0.05$ $t(11)=2.16$), but not between IntC and StaC themselves. This indicated a) that the presence of the tag cloud was more important than using it for search, and b) that a full study and a detailed analysis of learning would likely identify significant findings.

Main study measurements

Beyond logged data, like queries and clicks, we took 8 measurements of learning from the written summaries. First, we counted number of facts (assigned 'F1' as an identifier in our analysis) and second, we counted number of statements (or sentences – F2). Further, we calculated F3 as the ratio of facts per statement. We also analysed the breadth (number of topics – T1) and depth of topics using a 4-point likert scale (T2). Using Bloom's taxonomy of learning levels [1, 2], we created further 3 likert scales. D1 measured the quality of facts, D2 measured how statements were synthesized to draw conclusions and deductions and D3 identified whether summaries exhibited Evaluation in Bloom's taxonomy. D1-D3 looked beyond simply the number of facts, which could be banal statements like 'a Labrador is a type of dog' to identifying signs of higher level learning, in a topic-agnostic way. The full details of these new scales are reported elsewhere [14]. We used kappa scores until we achieved ‘substantial agreement’ for inter-rater reliability in all 8 measures.

Main Study Results

The data from only 34 participants was analysed below, as two participants did not participate correctly. As perhaps can be expected, the average number of queries in the IntC condition (3.47), was slightly higher than the StaC condition (2.88) and the baseline (2.56). These differences however, were not significant. Consequently, in regards to interaction, we concluded that the neither the presence nor the interaction with the word cloud had a significant impact on the number of queries submitted. There was also no significant difference between the number of words per query, nor the number of pages visited (NoC: 3.56, StaC: 3.74, IntC: 3.62). Participants rarely viewed more than one page of results for any query.

P-scores per Measure	**Pre-task**	**Post-task**
D1	0.81	0.5
D2	0.82	0.68
D3	0.91	0.54
F1	0.54	*0.08**
F2	0.86	*0.06**
F3	0.6	*0.06**
T1	0.25	*0.1**
T2	0.8	0.37

Table 1: Table showing the significance scores of each measure, when comparing the three interface conditions.

* indicates marginal differences

D1-3 = Levels of learning
F1-3 = Fact and statement counting
T1-2 = Topic breadth and depth

Analysis of Learning by Interface

As expected, we saw no significant differences between the three conditions in the state of knowledge before performing the task, across all measures (D1-3, F1-3, T1/2). As shown in Table 1, the topic coverage, depth, number of facts, length, or levels of blooms taxonomy were all approximately the same across interface conditions. Unexpectedly, however, we saw only marginal differences between the three interface conditions after the task in a) the number of facts (F1: F(2)=2.59, p=0.08), b) the number of statements written (F2: F(2)=2.85, p=0.06), and c) the ratio of facts per statement (F3: F(2)=2.86, p=0.06). We also saw a very minor influence on the number of topics covered (T1: F(2)=2.4, p=0.1).

To check whether this lack of significant results was created by the inclusion of tasks where participants stated they already had high knowledge on a topic, we divided the data into high and low pre-knowledge tasks. We expected to see more significant differences between the three interface conditions when people chose a task where they had low existing knowledge. We did not see, however, any significant differences between the three interface conditions in either high or low existing knowledge tasks. These results appear to indicate that neither the presence of nor the interaction with a keyword cloud in a web search engine had a significant impact on learning.

Analysis by Prior Knowledge

We also analysed the data to see whether the measures could differentiate between high and low pre-knowledge tasks. As could be expected we saw some significant differences in the quality of summaries written before learning began, as shown in Table 2. Participants in high pre-knowledge tasks included: a) marginally more facts (F1: t(100)=1.49, p=0.06), b) significantly higher topic breadth (T1: t(100)=1.83, p<0.05), and c) significantly better quality facts (D1: U(51) = 888.5, Z = 2.75, p<0.005). After the tasks, however, we were unable to find any significant differences across the measures, between tasks where participants began with high and low existing knowledge. One possible explanation is that all participants reached approximately the same level of understanding after the 5 minute learning task, regardless of task and interface.

Conclusions

Despite finding promising results in our pilot study [15], a larger study of 36 participants was unable to

P-scores per Measure	**Pre-task**	**Post-task**
D1	_0.003**_	0.29
D2	0.4	0.37
D3	0.31	0.43
F1	_0.06*_	0.24
F2	0.15	0.3
F3	0.3	0.36
T1	_0.04**_	0.22
T2	0.37	0.46

Table 2: Table showing the significance scores from independent t-tests for each measure, when comparing tasks where participants said they had high or low existing knowledge.

* indicates marginal differences
** indicates significant differences

D1-3 = Levels of learning
F1-3 = Fact and statement counting
T1-2 = Topic breadth and depth

show that either the presence of or the interaction with a keyword cloud had a significant impact on learning while using a web search engine. Despite using three different approaches to measuring learning, none showed any significant results. Consequently, we are unable to confirm our hypothesis that keyword clouds provide passive support for learning. From an analysis of the querying and searching behavior, including time spent on results page and the results themselves, we believe that the keyword clouds were simply not present during the majority of the sensemaking process, because people learned most while on other websites. In our future work, we aim to study the passive support that may be provided by these metadata structures in more integrated systems. In such vertical search environments, like digital libraries or online shopping sites, the metadata structures may be present throughout the process, including when viewing results, where we hope to observe a more significant impact on sensemaking and learning.

References

[1] Anderson, L., Krathwohl, D., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., Raths, J. and Wittrock, M. A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, Abridged Version. Allyn & Bacon, 2000.

[2] Bloom, B.S. and Engelhart, M.D. Taxonomy of educational objectives : the classification of educational goals. Handbook I, Cognitive domain. Longmans, London, 1956.

[3] Borlund, P. and Ingwersen, P., The Development of a Method for the Evaluation of Interactive Information Retrieval Systems. *Journal of Documentation*, *53*(3), 225-250. 1997.

[4] 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

[5] Dervin, B. From the mind's eye of the user: The sense-making qualitative-quantitative methodology. in *Qualitative research in information management*, Libraries Unlimited, 1992, 61-84.

[6] 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.

[7] Hearst, M.A. and Rosner, D., Tag Clouds: Data Analysis Tool or Social Signaller? In *Proc. HICSS 2008*, IEEE Computer Society, 160. 2008.

[8] 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

[9] Rivadeneira, A.W., Gruen, D.M., Muller, M.J. and Millen, D.R., Getting our head in the clouds: toward evaluation studies of tagclouds. In *Proc. CHI 2007*, ACM, 995-998. 2007.

[10] Schrammel, J., Leitner, M. and Tscheligi, M., Semantically structured tag clouds: an empirical evaluation of clustered presentation approaches. In *Proc. CHI 2009*, ACM, 2037-2040. 2009.

[11] Sharma, N., Role of available and provided resources in sensemaking. In *Proc. CHI 2011*, ACM, 1807-1816. 2011.

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

[13] White, R.W. and Roth, R. Exploratory Search: Beyond the Query-Response Paradigm. Morgan & Claypool, 2009.

[14] Wilson, M.J. and Wilson, M.L., A Comparison of 3 approaches to measuring learning through written summaries. *JASIST*(in submission).

[15] Wilson, M.J. and Wilson, M.L., Tag clouds and keyword clouds: evaluating zero-interaction benefits. In *Ext. Abstracts CHI 2011*, ACM, 2383-2388. 2011.

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