Getting Our Head in the Clouds: Toward Evaluation Studies of Tagclouds

A.W. Rivadeneira
Department of Psychology
University of Maryland
College Park, MD, USA 20742
arivadeneira@psyc.umd.edu

Daniel M. Gruen, Michael J. Muller, David R. Millen
Collaborative User Experience
IBM Research, One Rogers Street
Cambridge, MA, USA 02142
{daniel_gruen, michael_muller, david_r_millen}@us.ibm.com

ABSTRACT

Tagclouds are visual presentations of a set of words, typically a set of “tags” selected by some rationale, in which attributes of the text such as size, weight, or color are used to represent features, such as frequency, of the associated terms. This note describes two studies to evaluate the effectiveness of differently constructed tagclouds for the various tasks they can be used to support, including searching, browsing, impression formation and recognition. Based on these studies, we propose a paradigm for evaluating tagclouds and ultimately guidelines for tagcloud construction.

Author Keywords

Tagclouds, tagging, folksonomy, social software, evaluation, visualization

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Tagclouds are visual presentations of a set of words, typically a set of “tags” selected by some rationale, in which attributes of the text such as size, weight, or color are used to represent features of the associated terms (Figure 1). Tagclouds are becoming an increasingly familiar feature of social software sites in which content is categorized through evolving folksonomies. Examples of such sites include the social bookmarking site del.icio.us [5], the flicker photo-sharing site [9], the LibraryThing book recommendation site (e.g., [8]), and several enterprise-scale systems such as Dogear [10], Hermes [6], and Onomi [3].

In typical usage, Tagclouds are created by mapping a dimension associated with a term in an underlying data to a dimension parameter determining how that term should be displayed. For example, the prevalence of a term in the set could be represented by its size. Additional enhancements include the use of spatial algorithms to pack the words in a tag cloud into a smaller area, and clustering algorithms so tags which are used together or which have similar meanings are placed near each other.

Tagclouds support navigation to the underlying items, serving as automatically created tables-of-contents or indices into a block/batch/set of content. And, much as a table of contents or index can do for a book and a menu of categories can do for a website, they provide a means for users to form a general impression of the underlying set of content and a “gist” of what the book or site is about. In social software sites, where the tagclouds can represent the terms assigned by or associated in other ways with a person, the tagclouds can provide an impression of that person and his or her interests and expertise [4].

Despite their increasing popularity, we have seen no experimental studies evaluating the effectiveness of tagclouds. For example, a recent paper [7] discusses an algorithm for semantically-clustered tagclouds and provides an example of its use, but does not discuss formal evaluation of its effects on users.

This note is our first step in developing a paradigm and set of experimental data to evaluate the effectiveness of tagclouds for the various tasks they should support, considering the

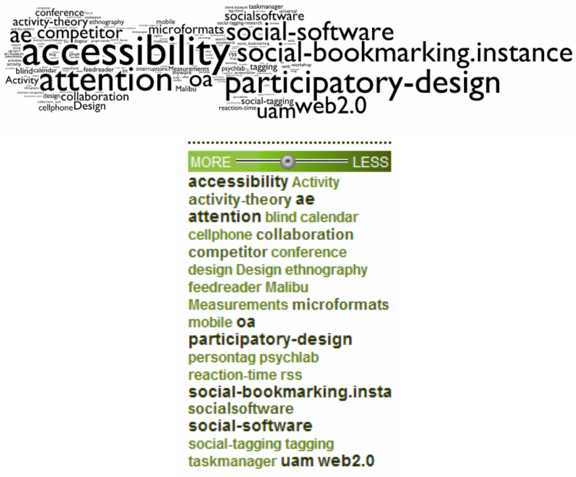


Figure 1. Two versions of a Dogear tagcloud for one of the authors.

CHI 2007, April 28–May 3, 2007, San Jose, California, USA.

contexts in which they appear. We are interested in understanding how the various dimensions used to construct a tagcloud affect different tasks. We believe studies of this type will stimulate the development of guidelines on how to design tagclouds for different settings. Other studies in this series may include additional display parameters and a reinvention/appropriation analysis of emergent, unanticipated uses of tags and tagclouds.

TASKS TAGCLOUDS CAN SUPPORT

Depending on the context, tagclouds can support user tasks ranging from locating specific items or groups of items, to providing an overview and general impression. Following is a list of tasks that we believe tagclouds can support:

Search: Locating a specific term or one that represents a desired concept (or determining that it is not there), often as a means to navigate to underlying content. .

Browsing: Using tagclouds as a means to browse, often with no specific target item or topic in mind.

Impression Formation or Gisting: Looking at the tagcloud as a means to form a general impression of the underlying data set or entity associated with it. This impression should include awareness of the most prevalent topics, but also knowledge of those that appear less frequently.

Recognition/Matching: Recognizing which of several sets of information or entities a tagcloud is likely to represent. For example, determining which of two John Smith's is the one you met at a conference based on their personal tagclouds.

TAGCLOUD FEATURES

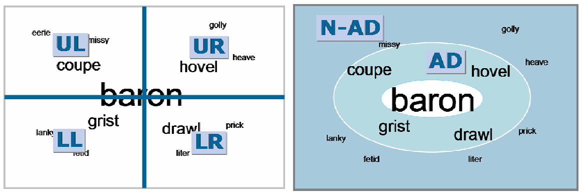
There are two types of features that are used to construct a tag cloud: text features and word placement.

Text features

Font Weight: The weight or bolding of text to represent frequency of an underlying quantity. It could potentially be used as a cue that denotes the grouping of items.

Font Size: The size of text as to represent a quantity, such as frequency of underlying items.

Font Color: Including the use of a single color to distinguish items across a single dimension, for example those the user has read or not read, or to represent or different underlying categories.



(a) (b)

Figure 2. (a) Tagcloud depicting Quadrants (UL=Upper-Left, LL=Lower-Left, UR=Upper-Right, LR=Lower-Right). **(b)** Tagcloud depicting Proximity-to-largest-font (AD=Adjacent, N-AD=Non-Adjacent).

Word Placement

Sorting: Words can be sorted alphabetically, by frequency or by a predetermined algorithm.

Clustering: Words can be sorted semantically or the users can specify their clustering preferences.

Spatial Layout: Words *can be located in sequential* line*s* or in a "bin-packed" cloud such as in the top of Figure 1.

Experiment 1

Method

We began with an examination of the influences of tagcloud attributes on low-level cognitive processes [1], manipulating font size and word location. All tagclouds had a spatial layout. The experimental design was a 3x4x2 repeated-measures design, with Font Size (High, Medium and Low), Quadrants (Upper-Left, Lower-Left, Upper-Right, Lower-Right) and Proximity-to-the-largest-font (Adjacent, Non-Adjacent) as independent variables, and recall as the dependent variable.

Subjects: Thirteen participants were recruited. All subjects had normal or corrected-to-normal vision.

Stimuli: Two-hundred and eight words were obtained from the MRC Psycholinguistic Database [11], with the following characteristics: 5-lettered words, 1-2 syllables, less than 5 phonemes, and a Kucera-Francis written frequency of 2 (this frequency is the mode in the database). Thirteen words (1-High, 4-Medium, 8-Small Font Size) were randomly sampled and located in predetermined locations in order to appear as a spatial tagcloud. A tag-per-person analysis for repeat-users of Dogear through April of 2006 reveals a median of 12 tags/person.

Procedure: Subjects performed one practice trial and ten experimental trials. Each trial started with a blank screen that was shown for 1 s. A tagcloud was then presented for 20 s. In order to eliminate any recency effects, a distracter task followed for 30 s (participants had to count backwards in threes starting from a random number). The trial ended with 60 s of free recall.[1]

Results

Table 1 shows means and standard errors. Recall for words with a larger font size was significantly higher than for words with a smaller font size ($F(2,24)=60.36$, $p<0.001$, $\eta^2=.83$). Words in the upper left quadrant were recalled significantly more than words in other quadrants. ($F(3,36)=3.52$, $p<0.05$, $\eta^2=.23$). There was no significant effect of proximity on recall. ($F(1,12)=3.07$, $p>0$).

[1] A working memory measure was obtained after all trials were completed. There were no significant effects of this measure. Thus, it will not be discussed in this paper.

	Mean	SE
Font Size		
High	0.725	0.049
Medium	0.413	0.043
Low	0.218	0.033
Quadrant		
Upper-Left	0.387	0.030
Lowe-Left	0.335	0.053
Upper-Right	0.290	0.035
Lower-Right	0.262	0.035
Proximity-to-largest-word		
Adjacent	0.290	0.033
Non-Adjacent	0.349	0.036

Table 1. Percent of correct recall for Font Size, Quadrant and Proximity-to-largest-word.

Discussion

The effect of font size was robust and expected; people recall words with larger fonts. We had expected words whose location was close to the largest font to be recalled more; we thought that attention would be drawn to this word and that participants would start scanning the other words from that starting point. We did not expect to see an effect of quadrant. This effect is usually expected on stimuli that require westernized reading (left-to-right and top-to-bottom). We believe the tagcloud was too sparse to induce this type of scanning, resulting in an upper-left quadrant effect. We hope to conduct future studies using eye-tracking devices to explore this hypothesis further..

Experiment 2

Method

In the second experiment we examined the effect of font size and word layout on impression formation and memory. Experiment 2 augments Experiment 1 by investigating both high-level and low-level processes [1]. The experimental design was a 5x4 repeated measures design, with Font Size (five levels: F1 to F5, big to small), and Layout (Sequential with Alphabetical Sorting, Sequential with Frequency Sorting, Spatial Layout (Feinberg's algorithm) and Single Column List with Frequency Sorting) as independent variables, and gist and recognition as dependent variables.

Subjects: Eleven participants were recruited; some had participated in the previous study. All subjects had normal or corrected-to-normal vision.

Stimuli: Forty-four categories, totaling 728 words, were obtained from the Information Mapping Project [2]. Categories are obtained by the distribution of co-occurrences between a word and some set of content-bearing terms; they were based on the New York Times corpus (mid 1990's). Four categories appeared per Tag; one related to an occupation and the other three were either hobbies or travel locations.

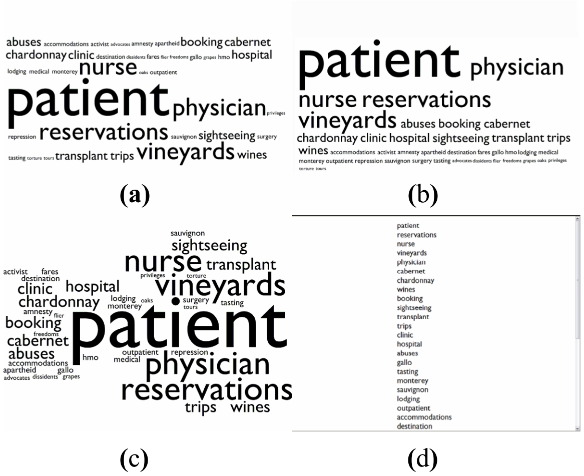

abuses accommodations activist advocates amnesty apartheid booking cabernet
chardonnay clinic destination dissidents fares flier freedoms gallo grapes hmo hospital
lodging medical monterey nurse oaks outpatient
patient physician privileges
repression reservations sauvignon sightseeing surgery
tasting torture tours transplant trips vineyards wines
(a)
patient physician
nurse reservations
vineyards abuses booking cabernet
chardonnay clinic hospital sightseeing transplant trips
wines accommodations activist amnesty apartheid destination fares gallo hmo lodging medical
monterey outpatient repression sauvignon surgery tasting advocates dissidents flier freedoms grapes oaks privileges
torture tours
(b)
sauvignon
sightseeing
nurse transplant
activist fares
destination
privileges torture
clinic hospital
vineyards
chardonnay
lodging oaks
monterey
surgery
tours
tasting
amnesty flier
booking
patient
freedoms
cabernet
abuses
outpatient repression
hmo medical
physician
accommodations
apartheid gallo
advocates dissidents grapes
reservations
trips wines
(c)
patient
reservations
nurse
vineyards
physician
cabernet
chardonnay
wines
booking
sightseeing
transplant
trips
clinic
hospital
abuses
gallo
tasting
monterey
sauvignon
lodging
outpatient
accommodations
destination
(d)

Figure 3. (a) Sequential – Alphabetical, **(b)** Sequential – Frequency, **(c)** Spatial, **(d)** List - Frequency

The category prototype was not used as stimulus. Ten words per category were used for each tagcloud, for a total of 40 words. A tag-per-person analysis for repeat users of Dogear reveals a mean of 39 tags/person.

Procedure: Subjects performed one practice trial and twelve experimental trials. Each trial started with a blank screen for 1 s. A tagcloud was then presented for 30 s, after which the participants had to describe the principal interests of the "tagcloud owner". We assumed that this task could be considered a distracter task and eliminate any recency effects. A recognition task followed that contained targets, semantically related distracters, and unrelated distracters.

Results

We assigned a score to measure impression formation (gist). A point was given each time a subject identified correctly one of the categories presented in each tagcloud. This scoring procedure was performed by two judges. The inter-rater reliability was high (0.96). The Tag Cloud presented as a list provided better comprehension of the interests of its "owner" (F(3,30)=4,26, $p<0.05$, η^2=.30).

	Mean	SE
Sequential - Alphabetical	2.258	0.169
Sequential - Frequency	2.174	0.138
Spatial	2.409	0.147
List - Frequency	2.682	0.090

Table 2. Impression Formation Scores by tagcloud layout**.**

Recognition for words with a larger font size was significantly higher than for words with a smaller font size (F(4,40)=49.37, $p<0.05$, η^2=.83). There was no significant effect of layout on recognition, for either hits or false alarms (F(6,60)<1). Semantically related distractors had more false positives than unrelated distractors (F(2,20)=74.87, $p<0.001$, η^2=.88).

	Mean	SE
Font Size 1 (High)	0.838	0.041
Font Size 2	0.737	0.049
Font Size 3	0.598	0.051
Font Size 4	0.430	0.047
Font Size 5 (Low)	0.326	0.049

Table 4. Percent of correct recognition by Font Size**.**

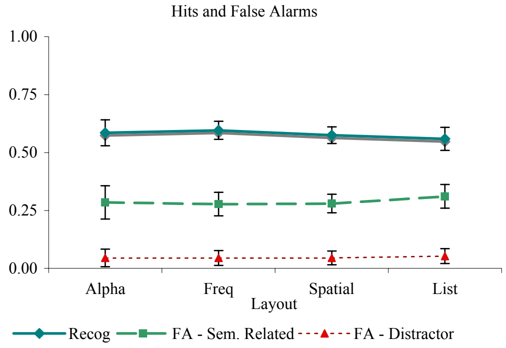


Figure 4. Proportion of Hits and False Alarms for Target Words, Semantically Related Words and

Discussion

The effect of font size was again robust and expected; people recognize words with larger fonts. There was no effect of layout on recognition; the layout of the tagcloud does not assist or hinder the recollection of the tags presented, though it is possible that such effects would be seen with shorter presentation times. There was a moderate, but statistically significant, effect of layout on impression formation. The list ordered by frequency resulted in a better identification of the categories present. The results for the spatial layout were second best behind list layout.

GENERAL DISCUSSION

Tagcloud designers can rely on established psychophysical principles in setting basic tagcloud parameters. These parameters produced results as predicted by perception theory: font size and location affected memory (low-level) processes. Instead, we propose that designers focus on layout, because this variable was shown to affect high-level processes, such as impression formation.

Our proposed paradigm for tagcloud evaluation consists of two-phased trials. The first phase is a presentation period that displays the tagcloud for a predetermined amount of time. This amount should depend on predicted tagcloud usage. We plan on manipulating presentation time to explore situations in which tagclouds are glanced at quickly. The second phase is an interpretative period in which impressions are elicited. Our study used open-ended questions to obtain participants' impressions; we believe multiple-choice questions or ratings are also applicable. The current paradigm presented the phases serially. We plan on investigating concurrent phases by asking participants to describe their impressions as they view tagclouds; this method will measure layout effects on the time and order in which concepts are formed.

There are additional manipulations to the impression formation task that we wish to investigate before generalizing our results, such as investigating semantically clustered layouts and employing eye-tracking devices to study effects of tagcloud layout on scanning, Finally, we plan to expand our set of tasks and corpus into a standard paradigm for evaluating tagcloud effectiveness in the full range of situations in which tagclouds are used.

For sparse tagclouds, such as in Experiment 1, designers may want to consider the upper-left quadrant as a focal point within the tagcloud. This recommendation can translate into different options: (a) locate smaller font words here to compensate for font size, while locating bigger font words in other quadrants; (b) locate tags that you want to emphasize in this quadrant.

The results from Experiment 2 imply that a simple list ordered by frequency may provide a more accurate impression about the tagger than other tagcloud layouts,.

Limitations

We investigated only one type of task in which tagclouds are used: impression formation. It is possible that our results were influenced by the introduction of a memory component in this task, and the choice of presentation time

REFERENCES

1. Ashcroft, M., *Cognition*. 3rd Ed. Prentice Hall, NJ, USA, 2002.
2. Computational Semantics Lab from Stanford University, "Information Mapping Project" http://infomap.stanford.edu/
3. Damianos, L., Griffith, J., & Cuomo, D., "Onomi: Social bookmarking on a corporate intranet," *Proc WWW 2006.*
4. Farrell, S., & Lau, T., "Fringe Contacts: People-Tagging for the Enterprise. *Proc WWW 2006.*
5. Golder, S., & Huberman, B.A., "The structure of collaborative tagging systems," Technical Report, Information Dynamics Lab, HP Labs, www.hpl.hp.com/research/idl/papers/tags/ tags.pdf (verified 26 Sept 2006)
6. John, A., & Seligman, D., "Collaborative Tagging and Expertise in the Enterprise," *Proc. WWW 2006.*
7. Hassan-Montero, Y., & Herrero-Solana, V., "Improving tag-clouds as visual information retrieval interfaces," *Proc. InfoSciT2006.*
8. Maness, J.M., "Library 2.0 Theory: Web 2.0 and its Implications for Libraries," *Webology 3(2)*, June 2006.
9. Marlow, C., Naaman, M., Boyd, D., & Davis, M., "HT06, tagging paper, taxonomy, Flickr, academic article, to read," *Proc Hypertext and Hypermedia 2006.*
10. Millen, D.R., Feinberg, J., & Kerr, B., "Dogear: Social Bookmarking in the Enterprise," *Proc CHI 2006.*
11. Wilson, M. "MRC Psycholinguistic Database: Machine Usable Dictionary. Version 2.00" (1987) http://www.psy.uwa.edu.au/ mrcdatabase/mrc2.html