Semantically Structured Tag Clouds: An Empirical Evaluation of Clustered Presentation Approaches

Johann Schrammel* **Michael Leitner*** **Manfred Tscheligi*‡**

*CURE - Center for Usability Research and Engineering
Hauffgasse 3-5, 1110 Vienna, Austria

‡ICT&S Center, University of Salzburg
S.-Haffner-Gasse 18, 5020 Salzburg, Austria

{schrammel, leitner, tscheligi}@cure.at

ABSTRACT

Tag clouds have become a frequently used interaction technique in the web. Recently several approaches to present tag clouds with the tags semantically clustered have been proposed. However, it remains unclear whether the expected gains in performance and advantages in interaction actually can be realized as no empirical evaluations of such approaches are available yet. In this paper we describe a series of experiments designed to evaluate the effects of semantic versus alphabetical and random arrangements of tags in tag clouds. The results of our work indicate that semantically clustered tag clouds can provide improvements over random layouts in specific search tasks and that they tend to increase the attention towards tags in small fonts compared to other layouts. Also, semantically structured tag clouds were preferred by about half of the users for general search tasks. Tag cloud layout does not seem to influence the ability to remember tags.

Author Keywords

Tag clouds, folksonomy, clustering, visualization

ACM Classification Keywords

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

INTRODUCTION

Tag clouds are visual displays of set of words (or 'tags') in which attributes of the text such as size, color or font weight are used to represent relevant properties e.g. frequency of documents linked to the tag.

Tag clouds have become a frequently used interaction technique in the web. The popularity of tag clouds can be explained by their ability to help users in getting a fast overview of a certain area.

CHI 2009, April 3–9, 2009, Boston, MA, USA.

Tag clouds are typically presented in alphabetical order, but also different approaches such as random or importance based tag arrangements have been used. Recently different methods to arrange tags semantically have been proposed [3,6,8] but no user-based evaluations of these approaches are available yet. It therefore remains unclear, whether these new presentation approaches are actually suited to improve the usefulness of tag clouds, and what the specific strengths and weaknesses of semantic arrangements in comparison to alphabetic and random tag layouts are.

RELATED WORK

Visual features of tag clouds

Concise research results regarding the importance of visual features within tag clouds are available. Both Bateman et al. [2] and Rivadeneira et al. [11] agree that font size, font weight and intensity prove to be to most important variables. However, the importance of tag position remains unclear, as the two studies come to different conclusions. Whereas [2] report no influence of tag position [11] found that tags in the upper-left quadrant had more influence than tags in the lower-right quadrant. This discrepancy may be a result of the different tasks (impression formation versus visual importance) and tag cloud sizes used in the studies.

Tag clouds and information seeking tasks

Different studies impose that tag clouds are a good visualization technique to communicate an 'overall picture'. Sinclair et al. [12] studied the usefulness of tag clouds versus search interfaces for different types of tasks (general versus specific searches) and came to the conclusion that tag clouds are particularly useful for non-specific information discovery and that they can provide a helpful visual summary of the contents. Similarly, comparing the visualization of search results using tag clouds in contrast to hierarchical textual descriptions Kuo et al. [10] found that users were able to answer overall questions better when using the tag clouds. However, both studies showed disadvantages of tag clouds for specific search tasks.

Layout of tag clouds

Halvey and Keane [7] investigated the effects of different tag cloud and list arrangements comparing the performance for searching specific items.

The setup included random and alphabetically ordered lists and tag clouds, semantic ordering was not part of there setup. They found that respondents were able to more easily and quickly find tags in alphabetical orders (both in lists and clouds).

Rivadeneira et al. [11] compared the recognition of single tags in alphabetical, sequential–frequency (most important tag at the left-upper side), spatially packed (arranged with Feinberg's algorithm) and list-frequency layouts (most important tag at the beginning of a vertical list of tags). Results did not show any significant disparity in recognition of tags. However, respondents could better recognize the overall categories presented when confronted with the vertical list of tags ordered by frequency.

Hearst and Rosner [9] discuss the organization of tag clouds. One important disadvantage of tag cloud layouts they mention is that items with similar meaning may lie far apart, and so meaningful associates may be missed.

Hasan-Montero and Herrero-Solana [8] proposed an algorithm using tag similarity to group and arrange tag clouds. They calculate tag similarity by means of relative co-occurrence between tags. Likewise, Fujimura et al. [6] use the cosine similarity of tag feature vectors (terms and their weight generated from a set of tagged documents) to measure tag similarity. Based on this similarity they calculate a tag layout, where distance between tags represents semantic relatedness. Another very similar approach is proposed by [3]. However, none of these approaches was evaluated systematically, and little is known about the actual effects of semantic tag layouts.

Figure 1. Example content displayed in four layouts

STUDY DESIGN AND MATERIALS

In detail we wanted to answer the questions, how semantic tag arrangement effects search time, the perception of tag clouds as well as the subjective satisfaction of the users both when searching for a specific tag and when performing searches for more general topics. Additionally we were interested to find out whether tag cloud layout influences the ability to remember tags. Three experiments were designed to answer these questions.

Our basic test content for all experiments was formed by 304 popular tags taken from 4 different thematic clusters of flickr. Tags were randomly assigned to one out of four groups, which then were used to generate the tag clouds. Each tag cloud used in the experiments consisted of 76 items arranged in 7 lines with 11 respectively 10 tags each. Every tag cloud included exactly 6 very big, 11 big, 22 small and 37 very small tags. Tags were randomly assigned to one of these font size groups. Four different tag cloud layouts were used in our experiments:

Alphabetic. Tags were placed alphabetically starting at the top left similar to reading text.

Random. Tags were placed randomly on the tag cloud by use of a random number generator.

Folksonomy-based. We used the getrelated-function of flickrs API[1] to retrieve a list of the tags most related to each word within the tag cloud. Then based on the number of co-occurring related tags a measure for the relatedness of two tags was calculated. An alternating least-squares algorithm to perform multidimensional scaling (ALSCAL)[2] was used to compute a two-dimensional arrangement of the tags. In the third step we used the value on the y-axis to form 7 groups of 11 resp. 10 tags each. Next tags within each group were sorted according to their value on the x-axis. The result provided an 11 times 7 arrangement that was used to generate the tag cloud.

Linguistics-based. In this approach we used WordNet [5] to calculate the relatedness of the used tags. Several algorithms to compute semantic distances based on WordNet data have been proposed, a comprehensive discussion can be found in [3]. We used the approach proposed by Banerjee et al. [1], as informal evaluation of the results of the different methods revealed the best results for our test content. After calculating a distance matrix the same procedure as described in the folksonomy-based approach was used to arrange the tags in rectangular shape.

Altogether 16 different tag clouds using all combinations of the four content sets with the tag cloud layouts were used. In all three experiments test conditions where presented in different order using content sets and tag cloud layouts that counterbalanced position and learning effects.

[1] http://www.flickr.com/services/api
[2] http://forrest.psych.unc.edu/research/alscal.html

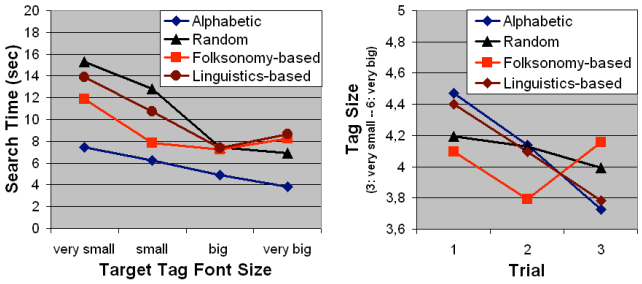

Figure 2: Interaction diagrams for experiment one (left) and experiment two (right)

EXPERIMENT ONE: FINDING SPECIFIC TAGS

The first experiment was designed to test how semantic tag arrangement influences search time and subjective satisfaction when searching for a specific tag.

The task for the 24 test participants (10 male, 14 female) was to find a specific tag within a tag cloud as fast and accurately as possible. The tag to be found was shown on the screen, on clicking 'Next' a tag cloud containing the target word appeared on the screen. The target word was also shown below the tag cloud. After locating the target tag participants had to click on it to proceed to the next task. Search time and clicked tag was logged.

Four different tag layouts as described above were used. For each layout 12 search tasks for different targets within the same tag cloud where performed. Target tags where evenly distributed across the four font sizes. We controlled for evenly distributed target position across the 4 quadrants of the clouds used in each condition. After each session with one tag cloud layout (i.e. 12 search tasks) test participants where asked to rate the difficulty of the tasks.

Effects of tag cloud layout and target size on search time

A two-way repeated measure analysis of variance (ANOVA) with search time as dependent variable shows significant effects for both factors target size ($F_{3,66}$=17.25, p<0.001) and tag cloud layout ($F_{3,66}$=17.16, p<0.001).

Follow-up analysis with paired samples t-test showed that the alphabetic layout is faster than all other conditions and that the folksonomy-based layout is faster than both, the random and the linguistics-based design. There is no difference between the random and linguistics-based layout.

In addition, there is also a significant interaction between target size and tag cloud layout ($F_{9,198}$=2.64, p=0.007). The interaction diagram in Figure 2 (left) shows that whereas for the conditions random and alphabetical the search time consistently decreases with increasing size of the target in the two semantically clustered conditions search times start to increase again for targets with very big fonts.

We think this effect is caused by users applying different visual search strategies to semantic tag layouts when they recognize the layout principle. Words in small fonts provide valuable information about the contents of an area. Therefore we can expect users to spend more time looking on smaller tags for identifying the topic. This explanation is consistent with participants feedback, however further research possibly using eye tracking technology is needed to understand the visual search strategies in detail.

Users Rating of Difficulty

An ANOVA[3] of difficulty rating of the users with regard to the different test conditions showed a similar pattern as the analysis for search time. We found a significant main effect ($F_{3,54}$=9.80, p<0.001), with the alphabetic layout rated better than all other layouts, and folksonomy-based layout significantly better than random.

EXPERIMENT TWO: FINDING TAGS RELATED TO A SPECIFIC TOPIC

In experiment two we studied the effects of tag cloud layout for more general searches. The main task for the participants (same as in experiment one) was to find a tag that belongs to a specific topic. Participants had to search alternately for three different topics within the same tag cloud. Each participant made 3 trials resulting in 9 search tasks per tag cloud. Test participants were instructed to select the same tag only once. For each topic 5 words (2 small, one of every other size class) were present.

Effects of tag cloud layout and trial on search time

A two-way repeated measures ANOVA with search time as dependent variable and the factors tag cloud layout and trial (first, second or third search for one topic) showed a significant effect for trial ($F_{2,44}$=6.87, p=0.003). Tag cloud layout ($F_{3,66}$=0.70, p=0.555) and interaction between factors ($F_{6,132}$=0.81, p=0.558) did not effect search time.

This result is surprising, as we expected semantic tag layout to provide - if any - benefit for general search tasks and not specific searches (experiment 1). We think that efforts of users to comprehend the semantic structure may counteract possible performance gains. However, more research is needed to clarify the involved mechanisms in detail.

Effects of tag cloud layout on size of selected tags

Analysis of Variance showed no effect for tag cloud layout ($F_{3,66}$=0.56, p=0.642) on the size of selected targets and a significant influence for trial ($F_{2,44}$=8.41, p=0.001): Tags in big fonts were generally found earlier than tags with smaller fonts. However, there is also a significant interaction ($F_{6,132}$=2.7, p=0.017). Figure 2 (right) illustrates the differences of the size of the selected tag for trial 1, 2 and 3. In contrast to the other three conditions in the folksonomy-based layouts the selected tags font size is not continously decreasing. Font size of the identified targets is decreasing from the first to the second trial, and then increases again for the third. The increased focus on smaller tags is consistent with the findings of experiment one.

[3] Due to an error in the test software ratings from the first 5 participants were not logged and therefore are not available for the analysis.

EXPERIMENT THREE: RECALLING TAGS

The third experiment was designed to evaluate whether variations in the tag arrangement have significant influence on the ability to recall tags or not.

16 test persons (7 male, 9 female) participated in experiment three. The main task of participants was to remember as many tags as possible. Tag clouds with different tag arrangements where shown for 30 seconds to the test subjects, and immediately afterward they were asked to enter all words they could remember into a form.

Effects of tag cloud layout on amount of recalled tags

On average people were able to recall 8.24 (StdD. 2.89) words correctly, with an average of 0.68 (StdD. 1.24) wrong words. One-way repeated measures ANOVA showed no influence of tag arrangement on amount of recalled tags ($F_{3,45}$=0.071, p=0.975).

In the interviews participants stated that the layout didn't make any difference whatsoever. This and the almost complete overlapping confidence intervals for the four test conditions support the interpretation that there is no or only an extremely small effect of tag cloud layout on recall rate.

Influence of tag size on recall rate

When comparing the distribution of remembered words across the different size categories to the expected distribution based on actual distribution of tag sizes in the original tag clouds a CHI-Square test shows a highly significant (Chi-Square=257.685, df=3, p<0.000) difference in the distributions. Not surprisingly the bigger the font of a word the higher the relative chance of being remembered is.

USER PREFERENCES

After finishing the study the participants of experiment one and two were shown printed tag clouds in the four different layouts and asked which one they would prefer for specific searches, general searches and on a web page in general.

	Specific Search	General Search	Web Page
Alphabetic	21	9	11
Random	0	1	1
Folksonomy	2	11	12
Linguistic	1	3	0

Table 1. Participant preferences (cells show number of participants who choose the layout).

DISCUSSION

Topically layouts can improve search performance for specific search tasks compared to random arrangements, but they still perform worse than alphabetic layouts. Considering that we used very simple clustering and arrangement algorithms we expect further advancement on semantic arrangements with more elaborate procedures.

The differences in results for the folksonomy-based and linguistics-based approaches indicate that the quality of the used clustering algorithms is of major importance for the resulting interaction. The semantic arrangement must be good enough otherwise users will not be able to distinguish it from random layouts. Semantic layouts therefore should only be used when the quality of the arrangement can be assured. Test participants also commented that it was difficult to identify clusters and relations beyond single lines. In future we plan to work on improved layout algorithms that take care of these limitations.

CONCLUSION

In this paper we investigated whether topically tag cloud layouts are good representations for finding specific tags, finding tags related to a category and for recalling tags. Our results include encouraging evidence for continuing semantic presentation approaches but also show that many details of such approaches are not yet fully understood.

REFERENCES

1. Banerjee, S. and Pedersen, T. Extended gloss overlaps as a measure of semantic relatedness. In *Proc. Artificial Intelligence. 2003*. 805-810
2. Bateman, S., Gutwin, C., and Nacenta, M. Seeing things in the clouds: the effect of visual features on tag cloud selections. In *Proc. Hypertext and Hypermedia 2008*. ACM Press (2008), 193-202.
3. Berlocher, I., Lee, K., and Kim, K. TopicRank: bringing insight to users. In *Proc. SIGIR 2008*. ACM Press (2008), 703-704.
4. Budanitsky, A. and Hirst, G. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Comput. Linguist.* 32, 1 (Mar. 2006), 13-47.
5. Fellbaum, C. (Ed.) *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA, USA. 1998
6. Fujimura, K., Fujimura, S., Matsubayashi, T., Yamada, T., and Okuda, H. Topigraphy: visualization for large-scale tag clouds. In *Proc. WWW 2008*. ACM Press (2008), 1087-1088.
7. Halvey, M. J. and Keane, M. T. An assessment of tag presentation techniques. In *Proc. WWW 2007*. ACM Press (2007), 1313-1314.
8. Hassan-Montero, Y., & Herrero-Solana, V. Improving tagclouds as visual information retrieval interfaces. *Proc. InfoSciT 2006*.
9. Hearst, M.A., and Rosner, D., Tag Clouds: Data Analysis Tool or Social Signaller? In *Proc. HICSS 2008*
10. Kuo, B. Y., Hentrich, T., Good, B. M., and Wilkinson, M. D. 2007. Tag clouds for summarizing web search results. In *Proc. World Wide Web 2007*. ACM Press (2007), 1203-1204.
11. 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 Press (2007), 995-998.
12. Sinclair, J. and Cardew-Hall, M. The Folksonomy Tag Cloud: When is it Useful? *Journal of Information Science.* 34, 1 (Feb. 2008), 15-29.