TagClusters: Semantic Aggregation of Collaborative Tags beyond TagClouds

Ya-Xi Chen [1], Rodrigo Santamaría[2], Andreas Butz [1], Roberto Therón[2]

[1] Media Informatics, University of Munich
Amalienstr. 17, 80333 Munich, Germany
[2] Department of Informatics and Automatics, University of Salamanca
Pz. de los Caídos, S/N, CP 37008, Salamanca, Spain
{yaxi.chen, andreas.butz}@ifi.lmu.de, {rodri, theron}@usal.es

Abstract. TagClouds is a popular visualization for the collaborative tags. However it has some instinct problems such as linguistic issues, high semantic density and poor understanding of hierarchical structure and semantic relation between tags. In this paper we investigate the ways to support semantic understanding of collaborative tags and propose an improved visualization named TagClusters. Based on the semantic analysis of the collaborative tags in Last.fm, the semantic similar tags are clustered into different groups and the visual distance represents the semantic similarity between tags, and thus the visualization offers a better semantic understanding of collaborative tags. A comparative evaluation is conducted with TagClouds and TagClusters based on the same tags collection. The results indicate that TagClusters has advantages in supporting efficient browsing, searching, impression formation and matching. In the future work, we will explore the possibilities of supporting tag recommendation and tag-based Music Retrieval based on TagClusters.

Keywords: Improvement of TagClouds, collaborative tagging, visualization of tags, semantic analysis, tag recommendation, music retrieval.

1 Introduction

With the rapid growth of the next-generation Web, many websites allow the normal users to make contributions by tagging the digital items. This collaborative tagging has become a fashion in many websites and the most representative ones are the social bookmarking site Del.icio.us (http://delicious.com/), the photo sharing site Flickr (http://www.flickr.com/) and the music community Last.fm (http://www.last.fm). Their low technical barrier and easy usage of tagging have attracted more than millions of users. The user-contributed tags are not only an effective way to facilitate personal organization but also provide a possibility for the users to search for information or discover new things.

Currently, there are two ways for Music Retrieval based on tags. The first category is the keyword-based search, which is the most common way to seek information on the Web. The system will return all the information related to the keyword. The second

one is a visualization-based method called TagClouds (as figure 1 shows). Due to its easy understandability and aesthetical presentation, TagClouds has become a fashion in many websites. However, it still has some intrinsic disadvantages and many researchers have been dedicated to improve its aesthetical presentation or better semantic understanding. In this paper, we explored how to improve the visualization of tags in Last.fm and to support the semantic understanding of structures and relations between tags which might lead to more success tag recommendation and visualization-based Music Retrieval.

In the remaining sections of this paper, TagClouds and its related works will be first discussed, and then the possibilities to improve semantic understanding of collaborative tags beyond TagClouds will be explored. A prototype named TagClusters and its evaluation will be presented. The discussions on the evaluation results and future work will conclude the paper.

2 TagClouds

TagClouds is a visual presentation of the most popular tags, in which tags are usually displayed in alphabetical order and attributes of the text such as font size, weight or color are used to represent features, for example, font size for prevalence and color brightness for recentness. As a result of collaborative tagging, the popular tags shown in TagClouds have more accurate meaning than that assigned by a single person. TagClouds draws attention to more important items and thus provides an overall impression and reflect the general interests among broad demography [5, 6]. With TagClouds a keyword-based search can be conducted with one selected tag as input.

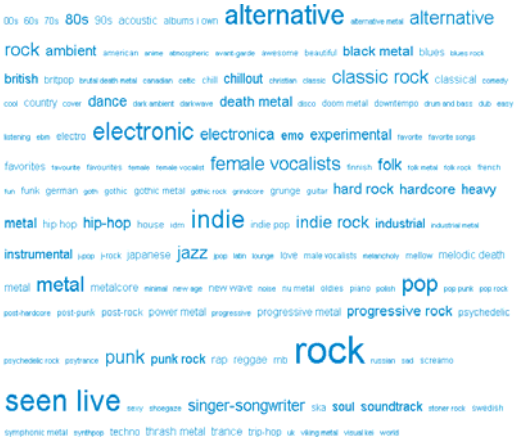

Fig. 1. TagClouds in Last.fm.

There are a lot of works focusing on the factors influencing the visualization quality of TagClouds. Halvey *et al.* [1] assessed tag presentation by evaluating different factors which ease the finding of tags. They discovered that font size, position and alphabetization can aid to find information more easily and quickly. Similar

conclusions could be found in [2, 15]. Rivadeneira *et al*. [3] also stated the interesting discovery that large fonts are not definitely to be found easily.

2.1 Disadvantages of TagClouds

As revealed in [24], after 100 users indexed the same resource, each tag's frequency is a nearly fixed proportion of the total frequency of all tags assigned to the resource. This stability was also verified in [25, 26]. Although there is a relative stability among collaborative tagging, because of the free nature of tagging, the intrinsic problem with uncontrolled tags is that there are inevitable noise and redundancies [8, 9].

Linguistic problems with free tagging
Nielsen [7] found that different educational and cultural background might lead to the tag inconsistency which was also mentioned in [27]. Specifically, there are two common problems with free tagging systems which are difficult to avoid from the user's side: synonymy and ambiguity. Synonymy is also defined as "inter-indexer inconsistency" [7] and it happens when different indexers use different terms to describe the same item. Ambiguity means one term may have several meanings [10], which will generate noise in retrieval results. Although the social collaborative tagging could alleviate the problems of synonymy and ambiguity, as pointed out in [11], such problems still exist, for example, *hip hop* and *hip-hop* in figure 1.

High semantic density
As discussed in [9, 12], if all the visible tags are selected only by the usage frequency, there might be a problem of high semantic density which means very few topics and related prominent tags tend to dominate the whole cloud and the less important items fade out [6]. Therefore, a more reasonable selection method should be improved.

Poor understandability of structure and relation
Hassan and Herrero [9] claimed that the alphabetical arrangements neither facilitate visual scanning nor infer semantic relation between tags. In the evaluation of Hearst *et al.* [6] a significant proportion of interviewees did not realize that tag clouds are regularly organized alphabetically. They discovered that the users have difficulty to compare tags with small size and to derive semantic relations. There might be wrong relation interpretation with items placed near to each other. Therefore TagClouds is not suitable for understanding of structure and relation.

2.2 Improvements

There are abundant works focusing on the improvements of TagClouds to support better aesthetic visualization and semantic understanding.

Enhancements of TagClouds
Regarding the aesthetic issue, since some factors might influent the effectiveness of visualization, some systems have already allowed the user to adjust these parameters and the representative systems are PubCloud [13] and ZoomClouds [14].
Tight coupling [16] addressed in improving the quality of TagClouds by introduction of spatial algorithms to pack the text in the visualization tighter. Kaser and Lemire [17] used electronic design automation (EDA) to improve the display of tag clouds to avoid large white space. Seifert *et al.* [18] proposed a series of algorithms which can display tags into arbitrary convex polygons with dynamical adaptive font size.
The clustering algorithms were applied to gather semantic similar tags. In [9] the k-means algorithm was applied to group semantic similar tags into different clusters. Similar work can be found in [19]. Li *et al.* [8] supported browsing of large scale social annotations based on analysis of semantic and hierarchical relations. The user profile and time factor can be integrated for personalized or time-related browsing [7].

Different visualizations for tags
Bielenberg and Zacher [20] proposed a circular layout in which the font size and the distance to the center represent the tag importance. Shaw [21] visualized the tags as a graph where edge represents the similarity. TagOrbitals [23] presented tags with relations and summary information in an atom metaphor where each primary tag is placed in the center and other related tags are placed in surrounding bands. The main problem with this visualization is the orientation of texts.
Most the methods discussed above are static visualizations and lack of interactions. Furthermore, the low level sub-structures are still needed to be deeper explored which will help to form a better understanding of hierarchical structure and relations.

3 TagClusters: support semantic understanding of tags

As we discussed above, TagClouds has instinct linguistic problems and it is difficult to understand the relation and structure between tags. In this paper, we have choosen the tags in Last.fm as the experimental source. We explore the problems with TagClouds in Last.fm and investigate the possibilities to improve the semantic understanding of tags.

3.1 Research issues

The key issues in our research are the semantic aggregation to support efficient hierarchical browsing and relation understanding. We believe that if all the tags are organized in a more understandable semantic way, it will be more helpful for tag recommendation and tag-based Music Retrieval.

Semantic aggregation
Based on the text analysis, the synonym issue can be controlled by grouping semantic similar tags into one cluster, for example, *favorite* and *favorites*, *rock and roll* and

rock n roll. The semantic aggregation also helps to alleviate the problem of ambiguity. For example, fewer users know that “electronic” and “IDM” present roughly the same genre. Within the visualization of TagClusters, the user can see these two tags are grouped into the same cluster which means they have same meaning. This is also an efficient way to help the users gain some music knowledge.

Hierarchy exploration and relation visualization
We explore the implicit hierarchical structure hidden inside the free input tags. With such structure, the user can have a better understanding of tags, especially genre-related categories. Following the top-down fashion, the user can search more specifically with less ambiguity problems. With the highlighting of the overlapped part, the user can tell the relation between tags in a high semantic level.

Possible applications based on TagClusters
Based on the hierarchical structure of TagClusters, there might be potential usages such as tag recommendation and tag-based music retrieval. Once the hierarchical structure of tags is derived, the user can get useful tag suggestions while avoiding spelling error and redundancy effectively. Genre is one of the most common criteria for music organization and retrieval [29], however, there is no standard definition of genre. For the tag recommendation, the system should offer possible suggestions for the user instead of generating them automatically. For example, if the user types in “electronic”, the system should prompt possible tags such as “IDM”. When the semantic similar tags are grouped hierarchically, it facilitates tag-based searching and the system can return richer relevant retrieval results.

3.2 TagClusters user interface

TagClusters is implemented based on Overlapper [28], a visualization tool focused on the connections and overlappings in data. TagClusters is an interactive interface where tags are drawn as labels with different sizes like in TagClouds, and tag groups are drawn as transparent colored areas (see figure 2). TagClusters uses the underlying structure of Overlapper, based on a Force-Directed Layout that does not use a typical node-link approach, but a Venn-diagram approach in order to represent groups and group relationships. The initial placement and label size of tags are not random, but coherent with tag co-occurrence (see section 4.2). Therefore, TagClusters can be described as a TagClouds where position is semantic relevant and the visualization is reinforced by group wrappers. These characteristics exploit the user perception abilities for traceability and group detection, improving the visual analysis.
In addition, the interface offers several options to the user: pan and zoom without losing context, hide/show tags and groups, modify label sizes, search tags by text, modify underlying forces, export the visualization, etc. The system also offers multiple options for tag selection which might facilitate the tag-based Music Retrieval. For example, the user can choose multiple tags or groups by combination of keyboard and mouse. He/she can also draw a shape manually and all the tags included in this shape will be selected.

Since most of the popular tags are genre-related and the relevant groups are overlapped with each other, the genre-related groups are placed in the center and other less semantic related tags and groups are scattered in the non-central space. Within a group, related tags will be further grouped into sub-groups. Groups are represented as transparent colored areas, so the overlapped groups (groups that share one or more tags) will have intersecting, more-opaque areas, thus highlighting overlapping tags. With such visualization, the user can have a better understanding of structure and relation between tags. For example, in Figure 2, we can see that *rock* is the most relevant category and related to several others, such as *pop* and *punk*. Also we find several genres that relate to *metal* and *rock* at the bottom of the figure.

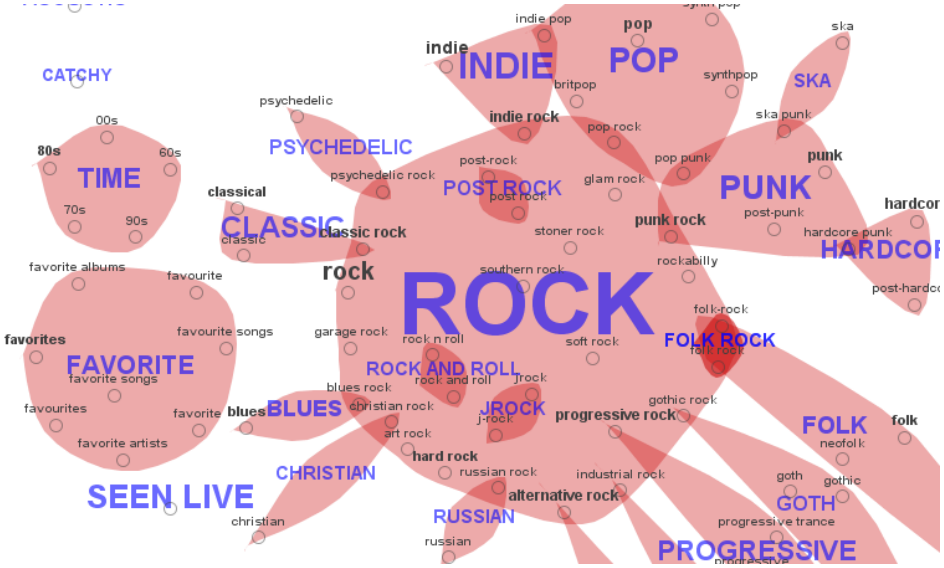

Fig. 2. TagClusters visualization with the same tag collections in figure 1.

4 Underground semantic analysis

The organization of tags in TagClusters is based on a semantic analysis which determines the structure and position for each tag. Firstly we applied text analysis to create the hierarchical structure while excluding redundancy. Then we determine the initial position for each tag based on the semantic similarity calculation.

4.1 Text analysis based clustering

After observation of TagClouds shown in figure 1, we found that synonymy is the most prominent problem with the texts which might be caused by single/plural such as “female vocalist” and “female vocalists”. The different spelling habits always lead to redundancy, for example, “favorite” and “favourite”. Besides, people tend to add

different separations between the same words, for example, “post-rock” and “post rock”, “rock and roll” and “rock n roll”.
In Last.fm, all the tags especially the genre-related tags have a characteristic feature: the tag in lower semantic level always contains the tag in the higher level and the length of tag is proportional with its semantic level, for example, “death metal” and “brutal death metal”. This feature helps to derive the hierarchical structure.
In our system, after removal of different separators, such as “_” and “&”, the Porter algorithm [30] is applied to detect the stem for each tag. Tags with the same stemmed words will be clustered in the same group. Within one group, tags with similar semantic meanings will be further clustered into sub-groups. For example, all the tags containing “metal” will be grouped in the Metal group and related tags such as “death metal” and “brutal death metal” will be placed in a sub-group (see figure 3); all the tags related to gender will be clustered in a vocal group and similar for the time-related tags, such as “80s” and “00s” (see left part of figure 2). After the text analysis, most of the tags can be effectively grouped into the relevant cluster. We also found that this basic technique should be further enhanced to solve the literal similar but actually irrelevant tags such as *classic* and *classic rock*.

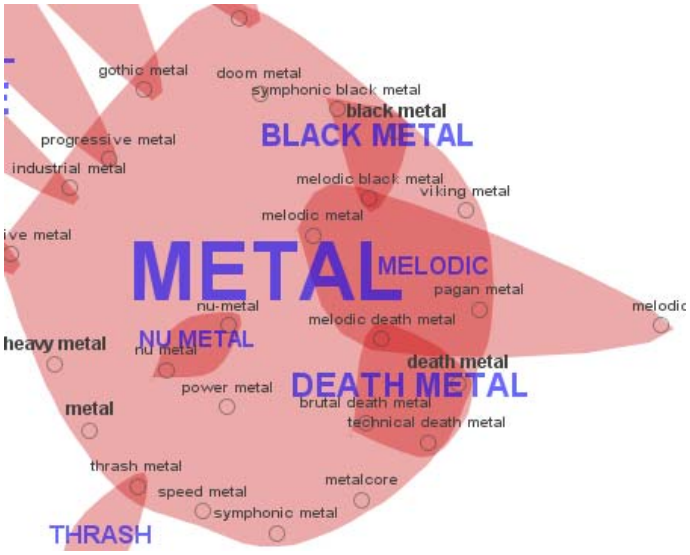

Fig. 3. Examples of text analysis results.

4.2 Calculation of semantic similarity

After deriving the hierarchical structure of tags, the semantic similarity is calculated based on the co-occurrence. Co-occurrence is widely used in the field of Music Retrieval to determine the semantic relation between information items [7, 27]. In the tag case, this semantic similarity equals to the division between the number of resources in which tags co-occur and the number of resources in which any of the two tags appears [7, 9], as equation 1 shows.

$$RC(A,B) = |A \cap B| / |A \cup B| \quad (1)$$

With the semantic analysis all the tags will be well organized: the initial location of each tag is assigned by means of a 2D projection based on a multidimensional scaling of co-occurrences. The genre-related tags, which might be the most useful category for tag-based searching, become prominent in the visualization. Other categories such as time- or emotional-related categories are scattered because of the less semantic relationship with the genre category. Instead of exclude them from the visualization, these categories still remain in the visualization and can be inspected by browsing or keyword-based searching.

5 Evaluation

To evaluate TagClusters, we recruited 12 participants at the University of Munich, 7 German from the Media Informatics Group and 5 foreigners from other groups, 4 female and 8 male with a mean age of 27 years. All participants are generally experienced with computers. We conducted a comparative evaluation of TagClouds and TagClusters. The evaluation was task-oriented and the participants were required to conduct 6 tasks concerning searching, browsing, impression formation and matching. Each task is consisted of two similar sub-tasks and the following is a brief description of the representative tasks.

Task 1: locate one single item: Find a tag named “german”.
Task 2: tag sorting: list the top 5 popular tags.
Task 3: tag comparison and filtering: list the top 5 genre-related tags.
Task 4: derive group structure: give a hierarchical structure for the Metal-related tags.
Task 5: find relation between tags: is there overlap between Indie and Classic?
Task 6: judge the tag similarity: is Alternative more similar to Rock or Electro?

5.1 Quantitative analysis

The complete time and the answer precision for both systems are shown in figure 4.

For task 1, although the tags in TagClouds can be located by the alphabetical order, locating the first character still needed some time. Furthermore, 25% of the participants did not realize that TagClouds are ordered alphabetically thus they spend more time to locate one tag. The participants claimed that the searching functionality in TagClusters helped to locate the item quickly.

TagClouds performed better with task 2 and 3. It contains all the tags in a small graph and it is easier to scan and locate tags without panning or zooming. To present all the tags as the concept of group and describe the similarity between tags as spatial distance, TagClusers needs more space and thus creates a larger graph in which the participant had to keep panning and zooming to get a complete overall impression. To form the correct impression, the participant also needed to mentally compare and memorize the relevant information which might slow down the response time and answered with lower precision.

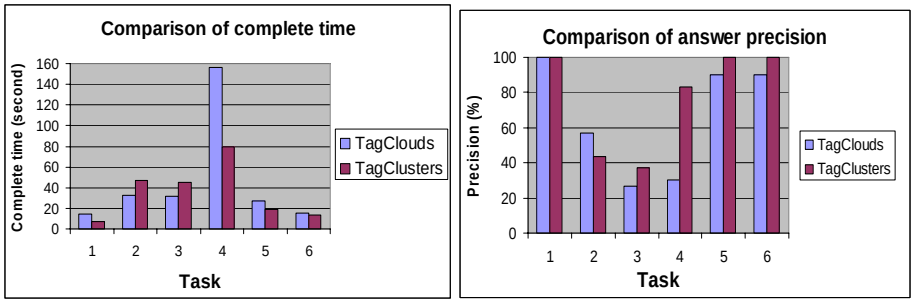

Fig. 4. Comparison of complete time and correctness.

TagClusters worked significantly better with task 4. Since the tags are hierarchically organized and semantic similar groups are placed near to each other, it is easy to find the structure. With TagClouds, the semantic similar tags might be placed scattered all over the graph, the participants had to scan all the tags and to form a structure mentally which spend much more time and led to lower precision. To derive the complex structure for Metal-related tags (as figure 4 shows), the participants spent much more time with TagClouds while received lower precision with the answers.
Since the semantic similar tags are hierarchically grouped and the overlapped part is visually highlighted, it is easier to determine the relation between tags. TagClusters worked better with task 5 and 6 which need understanding of semantic relation of tags.
After completed each task, the participants were asked to score the easiness of each task and the usefulness of both systems. The result is shown in figure 5.

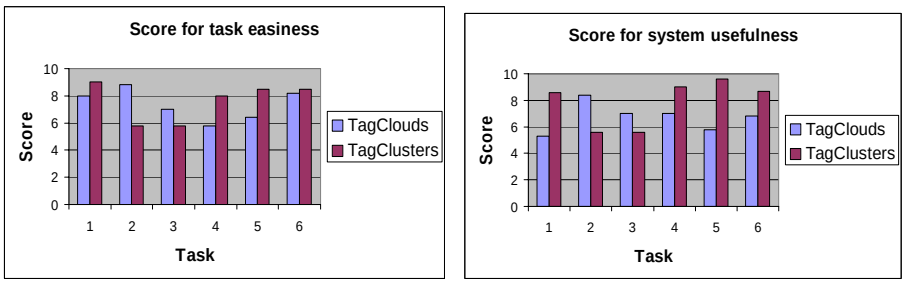

Fig. 5. Comparison of task easiness and system usefulness.

The participants scored higher for TagClusters except for task 2 and 3, which is consistent with the result in figure 5. It implies that we should take better usage of the space in TagCluster in order to create a smaller and more efficient visualization.
After completing all the tasks, the participants filled out a post-questionnaire which concerns the overall impression of both systems in the aspects of enjoyment, understandability, helpfulness and aesthetical beauty. TagClusters was scored better with all the criteria (as figure 6 shows).

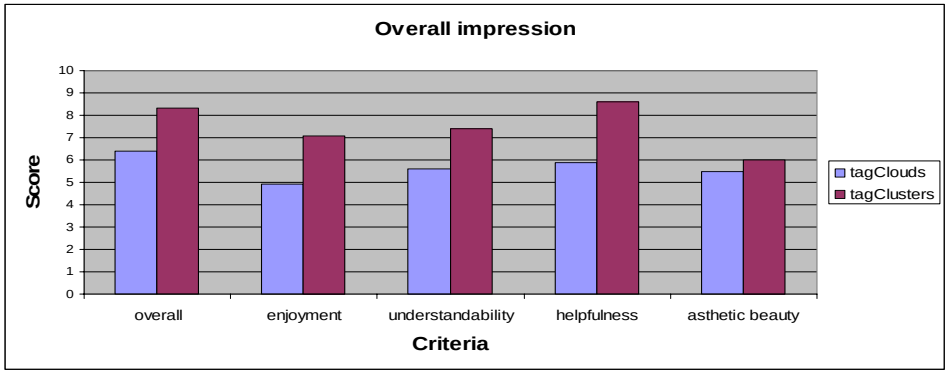

Fig. 6. Overall impression of both systems.

5.3 Qualitative analysis

Visualization issue

For TagClouds, the alphabetical order is useful when the user has an specific tag in mind. However, the user who is unfamiliar with this visualization tends to ignore this feature. Although the tags with bigger font size are easier to be noticed which is helpful to get the information of popular tags, the tags with smaller size might be ignored. The position is also a crucial factor to draw the user's visual attention. Some user claimed that top half part of TagClouds seems to be more prominent and they tend to ignore the bottom half. In order to get a compact view and take usage of space effectively, the system truncated the tag with long length into separate lines which might lead to misunderstanding. For example, in the first line of figure 1, alternative rock is placed into 2 lines and some participants were confused.

By grouping semantic similar tags, TagClusters helps to discover tags with small font size. For example, the rock-related tags with small font size might be ignored in figure 1 while still remaining relevance in figure 2 since they are clustered into the same group with the prominent Rock tag.

Semantic understanding

Without indication of semantic relation in TagClouds, some participants wrongly interpreted the semantic similarity as near position or similar font size. There is no semantic organization and sometimes the user has to scan all the tags line by line and might have problems with locating multiple tags in one time. Some participants even used mouse to locate viewed tag or staring at the screen while writing down the answers. Another problem is that the user who has less music knowledge might meet difficulties with judging the relation between some uncommon tags and it was a prominent problem with majority of the foreigner participants. With illustration of semantic structures in TagClusters they could conduct the same tasks easier.

For TagClusers, the participants also came up with some suggestions for the aesthetical issues in TagClusters, such as stronger highlighted effect, color coding for different tag categories and the desire for a more compact graph.

6 Conclusion

In this paper we investigate the ways to support semantic understanding of collaborative tags and propose a new visualization named TagClusters. We compare the performance of traditional TagClouds and our system and the results imply that our system has advantage in supporting semantic browsing and better understanding of hierarchical structure and relation between tags. The semantic organization of tags can exclude redundancy effectively and might facilitate the tag recommendation and tag-based music retrieval which will be explored in our future work.

7 Acknowledgement

This research was funded by the Chinese Scholarship Council (CSC), the German state of Bavaria and the Spanish Ministerio de Educación y Ciencia (project CSD 2007-00067). We would like to thank the participants of our user study.

References

1. Halvey, M. J. and Keane, M. T. An assessment of tag presentation techniques. Proc. of the 16th international conference on World Wide Web (New York, USA, 2007). WWW '07.
2. Golder, S. and Huberman, B. A. The structure of collaborative tagging systems. Journal of Information Science, 32(2), 198–208.2006.
3. Rivadeneira, A. W., Gruen, D. M., Muller, M. J., & Millen, D. R. 2007. Getting our head in the clouds: toward evaluation studies of tagclouds. Proc. of the SIGCHI conference on Human factors in computing systems(New York, USA). CHI '07.
4. Viégas F. B., Wattenberg M. Tag Clouds and the Case for Vernacular Visualization. Interactions. 15(4):49-52, 2008.
5. Hearst M. A. What's up with Tag Clouds? May 2008. Accessed December 30, 2008. http://www.perceptualedge.com/articles/guests/whats_up_with_tag_clouds.pdf.
6. Hearst, M. A. and Rosner, D. Tag Clouds: Data Analysis Tool or Social Signaller? Proc. of the International Conference on System Sciences (Waikoloa, HI, 2008).
7. Nielsen M. Functionality in a second generation tag cloud. Master thesis, Department of Computer Science and Media Technology, Gjøvik University College, 2007.
8. Li R., Bao S., Yu Y., Fei B., Su Z. Towards effective browsing of large scale social annotations. Proc. of the 16th international conference on World Wide Web (New York, USA, 2007).WWW '07.
9. Hassan M. Y. and Herrero S. V. Improving tag-clouds as visual Music Retrieval interfaces. Proc. of the International Conference on Multidisciplinary Information Sciences & Technologies (Merida, Mexico, 2006). INSCIT '06.

10. Mathes A. Folksonomies - cooperative classification and communication through shared metadata.http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html. Accessed December 30, 2008.
11. Wu X., Zhang L., Yu Y. Exploring social annotations for the semantic web. Proc. of the international conference on World Wide Web (Edinburgh, Scotland, 2006). WWW '06.
12. Begelman G., Keller P., Smadja F. Automatic tag clustering: improving search and exploration in the tag space. Proc. of the 15th international conference on World Wide Web (Edinburgh, Scotland, 2006). WWW '06.
13. Kuo B. Y.-L., Hentrich T., Good B. M. ., Wilkinson M. D. Tag clouds for summarizing web search results. Proc. of the 16th international conference on World Wide Web (New York, USA, 2007). WWW '07.
14. Zoomclouds. http://zoomclouds.egrupos.net/. Accessed December 30, 2008.
15. Bateman S., Gutwin C., Nacenta M. Seeing things in the clouds: the effect of visual features on tag cloud selections. Proc. of the nineteenth ACM conference on Hypertext and hypermedia (Pittsburgh, USA, 2008). HT '08.
16. Ahlberg C. and Shneiderman B. Visual information seeking: tight coupling of dynamic query filters with starfield displays. Proc. of the SIGCHI conference on Human factors in computing systems (New York, USA, 1994). CHI '94.
17. Kaser O. and Lemire D. Tag-cloud drawing: algorithms for cloud visualization. Proc. of the international conference on World Wide Web (New York, USA, 2007). WWW '07.
18. Seifert C., Kump B. Kienreich W. On the beauty and usability of tag clouds. Proc. of the 12th international conference on Information Visualization (London, UK, 2008). IV '08.
19. Provost J. Improved document summarization and tag clouds via singular value decomposition. Master thesis, Queen's University, Kingston, Canada. September 2008.
20. Bielenberg K. and Zacher M. Groups in social software: Utilizing tagging to integrate individual contexts for social navigation. Master's thesis, Program of Digital Media, University of Bremen, 2005.
21. Shaw B. Utilizing folksonomy: Similarity metadata from the del.icio.us system. http://www.metablake.com/webfolk/web-project.pdf. Accessed December 30, 2008.
22. Westerman S. J. and Cribbin T. Mapping semantic information in virtual space: dimensions, variance and individual differences. International Journal of Human-Computer Studies, 53(5), 765–787.2000.
23. Kerr B. TagOrbitals: tag index visualization. IBM research report. http://domino.research.ibm.com/library/cyberdig.nsf/1e4115aea78b6e7c85256b360066f0d4/8e1905dc8b000673852571d800558654?OpenDocument. Accessed December 30, 2008.
24. Golder B. and Huberman B. A. Usage Patterns of Collaborative Tagging Systems. Journal of Information Science, 32(2): 198-208.
25. Golder S. and Huberman B. A. The structure of collaborative tagging systems. Journal of Information Science, 32(2), 198–208. 2006.
26. Halpin, H., Robu, V., & Shepherd, H. 2007. The complex dynamics of collaborative tagging. Proc. of the 16th international conference on World Wide Web (New York, USA, 2007). WWW '07.
27. Begelman G., Keller P., and Smadja F. 2006. Automated tag clustering: Improving search and exploration in the tag space. Proc. of the 16th international conference on World Wide Web (Edinburgh, Scotland, 2006). WWW '06.
28. Rodrigo Santamaría and Roberto Therón. Overlapping Clustered Graphs: Co-authorship Networks visualization. Proc. of the 9th international Symposium on Smart Graphics (Rennes, France, 2008). SG '08.
29. Cunningham SJ, Bainbridge D., Falconer A. 'More of an Art than a Science': Supporting the Creation of Playlists and Mixes. Proc. of the 7th International Conference on Music Music Retrieval (Victoria, Canada, 2006). ISMIR '06.
30. Porter M. F. An algorithm for suffix stripping. Program, vol. 14, no. 3, pp 130-137, 1980.