Tag and Word Clouds as Means of Navigation Support in Social Systems

Martin Leginus and Peter Dolog

Department of Computer Science, Aalborg University,
Selma Lagerlofs Vej 300, 9220 Aalborg-East, Denmark
{mleginus,dolog}@cs.aau.dk
http://iwis.cs.aau.dk/

Abstract. Tag cloud is a visual interface that summarizes an underlying data by depicting the most frequent terms (also called as tags) from the dataset. Tags are linked to documents that contain given tags selection. A majority of tag clouds consists of the most frequent tags from a corpus that are alphabetically sorted. However, it has several drawbacks: frequent tags do not have to be relevant for all users, a vast number of terms are semantically similar hence a cloud contains many redundant depictions, an alphabetical sorting of tag cloud does not allow users to discover relations between terms. The objective of this PhD project is to propose, implement and evaluate novel tags selection methods for more relevant, diverse and novel tag clouds. Enhanced relevance of tag clouds should increase the likelihood that user will accomplish a given information retrieval task. Improved diversity and novelty of tag clouds should result into coverage of the entire spectrum of topics from folksonomy resources. Another objective is to expand a set of well-known synthetic metrics (i.e, Coverage, Overlap and Relevance) with new metrics that will capture diversity and novelty of tag clouds. Next ambition is to develop methods for tags clouds generation on top of social networks such as Twitter or Facebook. The objective is to propose words selection methods that will cover as many diverse subtopics from the underlying set of documents, tweets or statuses. The motivation is to minimize the user effort to skip redundant content.

1 The Scientific Content of the PhD Project

1.1 Background

Collaborative tagging has become an important and popular way of categorizing and retrieving various content within different Web 2.0 sites. Due to the simplicity and massive users engagement, tagging has been incorporated into the leading social web systems such as Facebook, Flickr, Delicious, Youtube, Mendeley, Connotea or Last.fm [16]. A user has possibility to classify and characterize a particular content item by annotating it with an arbitrary term – tag. Aggregated users's tagging activities with corresponding resources create a dataset that is commonly denoted as folksonomy. Folksonomy [16, 14] allows

Q.Z. Sheng and J. Kjeldskov (Eds.): ICWE 2013 Workshops, LNCS 8295, pp. 330–334, 2013.

a convenient retrieval and searching for the content according to a defined tag. Such folksonomy datasets can be visualized and summarized with a navigational interface called tag cloud. A tag cloud is a weighted list of terms - tags which are usually alphabetically sorted where more frequent tags have greater font size. Such visualization interface assists users to navigate through the web site content, discover new unexpected content (serendipitous discoveries) or follow the trends of the most popular and frequent topics of the web site. When no tags are available, words can be extracted from textual documents in order to generate a word cloud. For simplicity, in the rest of this document, we use tag clouds term to refer to word clouds as well.

In the following section, we present the state-of-the-art for this PhD project altogether with corresponding scientific challenges and problems.

Tag clouds usually depict a small subset from the most frequently used tags within the system. Such visual interfaces provide a rough overview about underlying data of the system. Hence, tag clouds are suitable for exploratory retrieval tasks (when user is not completely aware what he/she is looking for) and serve as a starting point for further more specific keyword-based search [1]. Tag clouds require less cognitive and physical workload as depicted tags facilitate an initial phase of the retrieval process [15]. In addition, users might discover unexpected content during a tag cloud exploration – such findings are as well classed as serendipitous [10].

Majority of tag clouds depict tags alphabetically sorted where the most frequent terms have greater font size. However, recent studies show that semantically aggregated tags enhance retrieval process [13, 8]. [8, 6] propose to group semantically related tags and depict them in a tag cloud near by with similar color. Such approach provides better orientation in the tag cloud as related tags can be easier identified by users. Such grouping of tags is usually achieved through a transformation of the underlying folksonomy graph into a co-occurrence graph [11]. We utilize a co-occurrence graph structure for estimating relevant tags for improved tag cloud generation and more detailed related work is presented in our second paper [9]. There are several studies that cluster tags into topics according to the tag pairs co-occurrences utilizing various similarity measures [2, 7, 17].

[18] introduces various synthetic metrics that measure a quality of tag clouds. The coverage, overlap, balance etc., are introduced. The advantage of these metrics is an ability to capture various aspects of a tag cloud without a need for an expensive user evaluation. On the other hand, these metrics do not capture important characteristics of tag clouds such as diversity or novelty of depicted tags and consequently diversity and novelty of their hyperlinked resources. Therefore, the outcome of this PhD project will be an extension of the existing synthetic metrics such that properties as diversity and novelty will be captured when measuring qualities of a tag cloud. Further, [18] proposes 4 tags selection algorithms for tag cloud generation. The first method selects the most popular tags within a system. The other two algorithms choose tags based on tf-idf (tags, documents). The last and most promising algorithm maximizes the coverage of selected tags

in the tag cloud. The proposed tag selection methods do not optimize a tag cloud structure towards relevance, novelty and diversity. Therefore, another objective of this PhD project will be a proposal of tag selection algorithms that will generate more relevant, novel and diverse tag clouds. It is important to mention, that proposed tag selection methods should not optimize only towards one particular tag cloud property. Instead, the methods should consider different tag cloud properties altogether during the generation process. The following paragraph presents related work about diversity and novelty from information retrieval perspective.

Diversity and Novelty. As there is no work about diversity and novelty for tag cloud generation, we present briefly related work from information retrieval perspective. [5] describes a brief history of diversity and novelty research. The initial idea claimed that a relevance of a document must be determined with respect to the documents appearing before it. [4] propose a marginal maximal relevance measure which attempts to maximize the relevance of the current document. At the same time, it strives for minimal similarity between the current document and previously selected documents. [19] introduces sub-topic retrieval methods which intuition is to include documents that cover many sub-topics early in the ranking and minimize the number of documents that redudantly cover the same subtopic. [5] proposes evaluation framework which distinguishes between novelty (minimisation of redudancy) and diversity (resolving of ambiquity).

Diversity and novelty are important aspects for developing information retrieval systems. Therefore, we plan to propose tag selection methods that will optimize tag cloud structure in terms of diversity and novelty. Moreover, we plan to propose a set of synthetic metrics that will appropriatelly capture diversity and novelty properties of tag clouds. These contributions will be based on the findings from the state-of-the-art of information retrieval diversity and novelty research studies.

Tag Clouds for Social Networks. [12] points out that microblog search needs a certain way of summarization. On the other hand trending topics presented at Twitter interface are characterized only by the single keyword or phrase. This results into more difficult understanding what a given topic is about. [12] proposes a system with tokenization module, where all unigrams, bigrams and trigrams are considered as possible sub-topics phrases for the final faceted search interface. Each phrase candidate is scored by the ratio of occurrences within query condition tweets and the entire tweets corpus. When candidate phrases are linking to the similar set of tweets, these topics are merged and represented by the more specific phrase. The final interface presents the top 40 sub-topic phrases related to the given query. This work is very similar to the proposed PhD objective, however, there are several drawbacks such as: it is difficult to quantify what is the real number of sub-topics and the effectiveness of the system was not evaluated by the users. It is important to know exact number of sub-topics to get a detailed understanding of the trending topic and to avoid reading redundant tweets, statuses or other resources. The problem with redudancy is even more evident within the trending topics which are available at Twitter interface. The reason is that thousands of users publish tweets (most of them are

semantically similar) about the same event within a short time interval. Therefore, the objective of this PhD project is to propose sub-topics detection algorithms which will simplify exploration of trending topics on Twitter. Another goal will be to investigate a possibility to apply these sub-topics detection algorithms for enhanced tag cloud generation for Facebook, where users are overwhelmed with a large number of statuses.

Further, we briefly present related work about tag clouds used for social networks. [3] proposes a novel topic-based browsing interface called Eddi. The interface clusters user's feeds according to their topics. Each tweet is transformed into a search query which is passed to a search engine. The retrieved documents are used for derivation of the tweet topic. The tag cloud interface is utilized for the depiction of the most trending topics within user's tweets. The main limitation of this work is the complete dependence of topics derivation through the usage of external search engine. Moreover, the tag cloud interface does not depict relations between topics and sub-topics. Above presented work should be possible to extend with the outcome of this PhD project. In other words, it should be possible to integrate sub-topics detection algorithms for enhanced tag cloud generation within Eddi interface.

1.2 Project Objectives

The objectives of this PhD project are trying to address the above-mentioned drawbacks and challenges of the state-of-the-art of navigation algorithms for social systems. The main objectives are following:

- To propose and evaluate new tags and words selection algorithms for tag cloud generation. The aim is to enhance tag cloud's structure where only the most relevant, diverse and novel tags with corresponding documents should be presented. The proposed tags selection methods should compositely consider as many tag cloud properties as possible during the generation process. In other words, generation of more relevant tag clouds should not result into miserable diversity or novelty of the tag clouds and vice versa.
- To enhance the set of synthetic metrics for tag clouds generation such that these metrics will capture diversity and novelty of tag clouds.
- To propose methodologies for tag cloud generation on top of social networks. These methods should provide sub-topics detection algorithms and consequent words selection algorithms so that users will obtain an appropriate understanding of the underlying data from Twitter or Facebook.

References

1. Aras, H., Siegel, S., Malaka, R.: Semantic cloud: an enhanced browsing interface for exploring resources in folksonomy systems. In: Workshop on Visual Interfaces to the Social and Semantic Web (VISSW 2010), IUI 2010, Hong Kong, China, February 7 (2009)
2. Begelman, G., Keller, P., Smadja, F.: Automated tag clustering: Improving search and exploration in the tag space. In: Collaborative Web Tagging Workshop at WWW 2006, Edinburgh, Scotland, pp. 15–33. Citeseer (2006)

3. Bernstein, M.S., Suh, B., Hong, L., Chen, J., Kairam, S., Chi, E.H.: Eddi: interactive topic-based browsing of social status streams. In: Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology, pp. 303–312. ACM (2010)
4. Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1998, pp. 335–336. ACM, New York (1998)
5. Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, pp. 659–666. ACM, New York (2008)
6. Durao, F., Dolog, P., Leginus, M., Lage, R.: SimSpectrum: A similarity based spectral clustering approach to generate a tag cloud. In: Harth, A., Koch, N. (eds.) ICWE 2011. LNCS, vol. 7059, pp. 145–154. Springer, Heidelberg (2012)
7. Grahl, M., Hotho, A., Stumme, G.: Conceptual clustering of social bookmarking sites. In: Proceedings of I-KNOW, vol. 7, pp. 5–7 (2007)
8. Hassan-Montero, Y., Herrero-Solana, V.: Improving tag-clouds as visual information retrieval interfaces. In: International Conference on Multidisciplinary Information Sciences and Technologies, pp. 25–28. Citeseer (2006)
9. Leginus, M., Dolog, P., Lage, R.: Graph based techniques for tag cloud generation. In: Proceedings of the 24th ACM Conference on Hypertext and Social Media. ACM (2013)
10. Mathes, A.: Folksonomies-cooperative classification and communication through shared metadata. Computer Mediated Communication 47(10) (2004)
11. Mika, P.: Ontologies are us: A unified model of social networks and semantics. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 522–536. Springer, Heidelberg (2005)
12. O'Connor, B., Krieger, M., Ahn, D.: Tweetmotif: Exploratory search and topic summarization for twitter. In: Proceedings of ICWSM, pp. 2–3 (2010)
13. Schrammel, J., Leitner, M., Tscheligi, M.: Semantically structured tag clouds: an empirical evaluation of clustered presentation approaches. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems, pp. 2037–2040. ACM (2009)
14. Sinclair, J., Cardew-Hall, M.: The folksonomy tag cloud: when is it useful? J. Inf. Sci. 34, 15–29 (2008)
15. Sinclair, J., Cardew-Hall, M.: The folksonomy tag cloud: when is it useful? Journal of Information Science 34(1), 15–29 (2008)
16. Smith, G.: Tagging: people-powered metadata for the social web. New Rider Pr. (2008)
17. Specia, L., Motta, E.: Integrating folksonomies with the semantic web. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 624–639. Springer, Heidelberg (2007)
18. Venetis, P., Koutrika, G., Garcia-Molina, H.: On the selection of tags for tag clouds. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM 2011, pp. 835–844. ACM Press, New York (2011)
19. Zhai, C.X., Cohen, W.W., Lafferty, J.: Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, SIGIR 2003, pp. 10–17. ACM, New York (2003)