Proceedings of I-KNOW '08 and I-MEDIA '08
Graz, Austria, September 3-5, 2008

Harnessing Wikipedia for Smart Tags Clustering

Maria Grineva, Maxim Grinev
Denis Turdakov, Pavel Velikhov, Alexander Boldakov
(Institute for System Programming of Russian Academy of Sciences
Moscow, Russia
rekouts@ispras.ru, maxim@grinev.net
turdakov@ispras.ru, velikhov@ispras.ru, boldakov@ispras.ru)

Abstract: The quality of the current tagging services can be greatly improved if the service is able to cluster tags by their meaning. Tag clouds clustered by higher level topics enable the users to explore their tag space, which is especially needed when tag clouds become large. We demonstrate TagCluster - a tool for automated tag clustering that harnesses knowledge from Wikipedia about semantic relatedness between tags and names of categories to achieve smart clustering. Our approach shows much better quality of clusters compared to the existing techniques that rely on tag co-occurrence analysis in the tagging service.

Keywords: Wikipedia, User-generated content analysis, network analysis, tagging
Categories: H.3.3, H.3.5

1 Introduction

Tagging services that assign relevant keywords to documents or objects have become quite popular in the past few years. Currently tagging is an essential part of Web 2.0 applications such as social bookmarking services (Furl, del.icio.us), blogging (Technorati) and photo-sharing services (Flickr) [O'Rreilly, 2005].

Tagging services provide their users with a repository of tagged objects - a *tag space* - that can be explored via *tag cloud*. Tag cloud is some kind of a visual depiction of a set of tags. Tags in a tag cloud are either listed alphabetically, or the size of tags in a tag cloud is proportional to their popularity. In practice, such tag clouds fail to help in exploration of the tag space when the number of tags becomes more a less significant (approximately more than 100). Improving search and exploration in tag spaces has been studied from different perspectives, among a variety of approaches we distinguish a simple yet promising idea that automatically dividing tag cloud into a number of semantically cohesive clusters would make it much more helpful for tag space exploration [Begelman, 2006], [Hassan-Monteroa, 2006].

In existing methods [Begelman, 2006], [Hassan-Monteroa, 2006], relatedness between tags is inferred by means of tags co-occurrence analysis of the tagging service repository: tags are considered related if they are assigned to a common object. However, this assumption is quite uncertain and as a consequence the methods often produce dirty clusters [Begelman, 2006]. In contrast, the key point of our approach is using Wikipedia to compute semantic relatedness between tags and to pick the names for the clusters and we demonstrate that this results in higher quality tag clusters.

2 Key Techniques

TagCluster processes tag cloud in three steps. The following key techniques are used in these steps:

• For each tag in the tag cloud we find its corresponding concept in Wikipedia. We use our Wikipedia-based disambiguation tool for proper handling of homonym tags. The disambiguation tool is described in [Velikhov, 2008].

• We create a weighted graph for the tags, where each vertex is the corresponding Wikipedia concept and each edge is the relatedness between concepts with the corresponding relatedness weight. Relatedness measure between Wikipedia concepts is computed as described in [Velikhov, 2008], [Lizorkin, 2008].

• We use Girvan-Newman community detection algorithm [Newman, 2004] to partition the graph into semantically cohesive subgraphs. For each subgraph we derive its topic: we compute centrality measures for its vertices, then collect the Wikipedia categories of the concepts (vertices) and rank them according to the concept's centrality measure. Categories with the highest rank constitute the community subgraph topic.

3 Demonstration

We demonstrate TagCluster on the del.icio.us tag clouds. We collect the tag cloud of any given del.icio.us user and perform the clustering. We show that our approach produces smarter clusters compared to existing approaches.

Acknowledgements

The work is partially supported by grants of RFBR NN 08-07-00195 and 08-07-12010.

References

[Begelman, 2006] G. Begelman, P. Keller, F. Smadja. Automated Tag Clustering: Improving search and exploration in the tag space. In *Proc. of the WWW2006*

[Hassan-Monteroa, 2006] Y. Hassan-Monteroa and V. Herrero-Solana. Improving Tag-Clouds as Visual Information Retrieval Interfaces *InSciT2006,* 2006

[Lizorkin, 2008] D. Lizorkin, P. Velikhov, M. Grinev. Accuracy Estimate and Optimization Techniques for SimRank Computation. *To appear in VLDB '08*.

[Newman, 2004] Girvan, M. and Newman, M. E. Finding and evaluating community structure in networks. *Phys. Rev. E 69, 026113*, 2004

[O'Rreilly, 2005] O'Reilly, T.: What Is Web 2.0. Design Patterns and Business Models for the Next Generation of Software. O'Reilly Radar report, 2005

[Velikhov, 2008] P. Velikhov et al.. Efficient Ranking and Computation of Semantic Relatedness and its Application to Word Sense Disambiguation. *ISPRAS MODIS Technical report,* 2008