A Cluster-Based Approach for Search and Exploration of Tag Spaces

Joni Radelaar Aart-Jan Boor Damir Vandic Jan-Willem van Dam
Frederik Hogenboom Flavius Frasincar

Erasmus School of Economics, Erasmus University Rotterdam
P.O. Box 1738, 3000 DR Rotterdam, the Netherlands

The full version of this paper, entitled IMPROVING THE EXPLORATION OF TAG SPACES USING AUTOMATED TAG CLUSTERING appeared in: International Conference on Web Engineering (ICWE 2011), volume 6757 of Lecture Notes in Computer Science, pages 274-288. Springer, 2011.

Abstract

Although Semantic Web technology is increasingly becoming more and more important, tagging remains a popular method to describe Web resources. Therefore it is important to address the issues that are found in current tagging search engines, such as Flickr. We find that the free nature of tagging results in many issues for tag search engines, such as synonyms, homonyms, syntactic variations, etc. The Semantic Tag Clustering Search (STCS) framework deals with these issues by detecting syntactic variations of tags and by clustering semantically related tags. We evaluate our framework using Flickr data from 2009 and compare the STCS framework to two previously introduced tag clustering techniques.

1 Introduction

Tagging is very popular among users because it provides a flexible way of describing resources on the Web. This flexibility stems from the fact that there are no restrictions on the tags users can use. However, this freedom that tagging brings comes with a price, as there are some problems associated with retrieving resources using tag-based search engines. These problems are often caused by different tags having the same or closely related meaning. This can be the result of the use of synonyms, but it could also be caused by syntactic variations. Users may also use different levels of specificity while describing a resource, which is identified as the basic level variation problem [1]. The usage of homonyms, i.e., words with multiple unrelated meanings, is another problem associated with tagging.

In this paper we propose the STCS framework as a solution to these problems. The main idea is that we create clusters of syntactically and semantically related tags. Search algorithms can then use these clusters to significantly improve the recall of the search results. Tags occurring in multiple semantic clusters can be used to identify tags with multiple meanings. If a tag occurs in multiple clusters it most likely also has multiple meanings, e.g., “turkey” can refer to both the country and the animal.

2 STCS framework

For the syntactic variations, we implement and evaluate the use of the Levenshtein similarity measure [4] and a combination of the Levenshtein similarity and the cosine similarity measure (the cosine of the angle between two tag co-occurrence vectors), as similarity measures for syntactically related tags. We introduce the cosine similarity in order to address the issue with short tags. For tags like ‘car’ and ‘cat’, the Levenshtein similarity is not a reliable syntactic variation measure. The final syntactic similarity measure is a weighted average of the normalized Levenshtein similarity and the cosine similarity between the two tags.

The weight is determined by the maximum length of the two tags. For longer tags, we give more weight to the normalized Levenshtein similarity and for shorter tags the cosine similarity gets more weight. In order to obtain syntactic clusters, we first construct a graph, where the tags are the vertices and the edges are weighted by the syntactic similarity measure. Next, we cut all edges that are below a certain threshold. The final step is to find all connected components, which are the syntactic clusters.

For identifying semantic clusters we implement and evaluate the semantic clustering algorithm proposed by Specia and Motta [6] and a clustering algorithm proposed by Lancichinetti et al. [2]. Additionally, we propose a modification to the Specia and Motta approach. The original algorithm of Specia and Motta relies on cluster merging heuristics. Our modification relates to the fact that we use a dynamic threshold instead of a constant threshold for the number of missing elements in the process of merging two clusters. Also, we introduce a semantic merging heuristic, which considers how 'semantically' related two clusters are, instead of just considering the number of tags that are in one cluster and not in the other.

For the evaluation of the clustering algorithms, we use the average precision measure [3] and the purity measure [5] for a sample of 100 clusters, in combination with a three user pool where we apply majority voting. For the STCS version of semantic clustering, the average precision is 0.86 and the purity is 0.89, while the average precision and purity for the original clustering algorithm are 0.80 and 0.87, respectively. Our algorithm significantly outperforms the original algorithm with respect to the precision, at a 5% significance level. The average precision of Lancichinetti's method is 0.81 and the cluster purity is 0.77.

3 Conclusions

Our framework addresses the search issues in tag spaces by using two algorithms, one for creating syntactic clusters and one for creating semantic clusters. For the syntactic clustering, we make use of a combined measure of the Levenshtein distance and the cosine similarity. We compared the results of clustering with the combined STCS measure with clustering using the Levenshtein distance. Our conclusion is that the combined STCS measure performs significantly better in terms on precision. The clustering method as proposed in our framework was able to effectively filter out syntactic variations from the data set. For semantic clustering, the framework uses an adaptation of the approach proposed by Specia and Motta [6]. Our experiments show that our proposed method significantly outperforms the original method by Specia and Motta and outperforms on average the method of Lancichinetti in terms of precision. Finally, we have shown that our results are valid on a significantly larger data set than was used before in the existing body of literature.

References

[1] S.A. Golder and B.A. Huberman. The Structure of Collaborative Tagging Systems. Technical report, Information Dynamics Lab, HP Labs, 2005. From: `http://www.citebase.org/abstract?id=oai:arXiv.org:cs/0508082`.

[2] Andrea Lancichinetti, Santo Fortunato, and Janos Kertesz. Detecting the Overlapping and Hierarchical Community Structure in Complex Networks. *New Journal of Physics*, 11(3):1–19, 2009.

[3] B. Larsen and C. Aone. Fast and Effective Text Mining using Linear-Time Document Clustering. In *5th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 1999)*, pages 16–22. ACM, 1999.

[4] V. I. Levenshtein. Binary Codes Capable of Correction Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[5] C.D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[6] L. Specia and E. Motta. Integrating Folksonomies with the Semantic Web. In E. Franconi, M. Kifer, and W. May, editors, *4th European Semantic Web Conference (ESWC 2007)*, volume 4519 of *Lecture Notes in Computer Science*, pages 503–517. Springer, 2007.