INTELLIGENT TAG GROUPING BY USING AN AGLOMERATIVE CLUSTERING ALGORITHM

Andrej Gajduk Gjorgji Madjarov Dejan Gjorgjevikj
Faculty of Computer Science and Engineering
Skopje, Republic of Macedonia

ABSTRACT

Tagging can be defined as a process of assigning short textual descriptions or key-words (called tags) to information objects. It is a simple approach to information organization that was regularly practiced over the last years. Tagging systems usually have relatively flat tags. This means that while one can easily browse by a tag, one cannot as easily see tags that have wider or more specific meaning than a given tag. It is also difficult to get a broad overview of the tags that do exist in the tagging systems, aside from frequency based displays like tag clouds. In this paper we investigate how correlated tags can be grouped by using an agglomerative clustering algorithm considering only the label part (output space) of the data. We have applied this approach on the StackOverflow tag cloud and discuss the obtained results.

I. INTRODUCTION

With the rapid growth of the internet the challenge of organizing the huge amounts of data available is becoming increasingly important. Tagging is a powerful mechanism that can help with information organizing and improving the search options. Tagging usually refers to the action of associating a user generated keyword or phrase with an entity (e.g. a document, image or a video). Tags are different from classical taxonomies, because they rely on an uncontrolled vocabulary where there are no predefined categories, thus producing a tag set that is not finite.

The great flexibility offered by this feature emphasizes on ease of use, but also makes information organization and access more challenging for the research community. The proliferation of Flickr and Delicious engendered a great deal of interest in the tagging service and popularized the concept that is now used by many websites, as YouTube, CiteULike, Technorati, Last.fm and StumbleUpon. Some popular tagging systems that have been studied in depth include Delicious [1], Flickr [2], and Connotea [3].

Despite its popularity, the effectiveness of tagging as a primary organizing mechanism is yet to be shown [4, 5, 6, 7]. Common for the first generation of tagging systems seems the use of flat tags that do not impose a hierarchy or any other relationships to each other. The lack of established relationships between tags is a limiting factor: one cannot easily see the tags with wider or narrower meaning than a concrete tag; it is difficult to get a broad overview of what tags exist in the tagging systems; the same content can be tagged differently by different people; the process of automatic tag suggestion is getting more difficult as a result of the vast number of possible tags.

In this paper we investigate how correlated tags can be grouped by using an agglomerative clustering algorithm considering only the label part (output space) of the data. We have applied this approach on the most popular question and answer site about programming - StackOverflow that contains more than 70,000 flat, non-organized tags.

The rest of this paper is organized as follows. We first briefly review the related work in Section2. Section 3 introduces the concept of agglomerative clustering and the methods we will use for the task of clustering a tag cloud. In section 4, we review the dataset used in the experiments. In section 5, we demonstrate the experimental results on clustering the StackOverflow tag cloud. Finally, Section 6 concludes and indicates several issues for future work.

II. RELATED WORK

Most researchers recognize the need of imposing some kind of connection between the different tags in the tag cloud. Begelman et al. [8] apply the spectral bisection algorithm to the problem of clustering tags using tag-pairs frequency of co-occurrences as a feature vector. Hayes, and P. Avesani [9] explore the possibility of using tags and their correlation for the purpose of identifying topic-relevant blogs using k-means algorithm, while Song et al. [10] are concerned with the problem of visualizing the hierarchy that exists in the tag cloud. They build the tag hierarchy using a custom developed greedy algorithm that forms groups based again on the number of co-occurrences, but this time between a tag and a group of tags. Heymann and Molina [11] suggest building a tag hierarchy using an algorithm that leverages notions of similarity and generality that are implicitly present in the data generated by users as they annotate objects.

III. CLUSTERING

Clustering has been used for improving precision/recall scores for document retrieval systems using topic-driven

language models [12,13], browsing large document collections [14], organizing search engine return sets [15, 16], improving tag recommendation systems [17,18,19,20] and grouping similar user profiles in recommender systems [21,22,23].

We recommend using an agglomerative clustering as a method to group similar tags considering only the label part (output space) of the data. An agglomerative clustering algorithm starts with N groups, each initially containing one instance, merging similar groups to form larger groups, until there is only one single group left. At each iteration of the agglomerative algorithm, the two closest groups are merged. In order to apply this algorithm we need to define a feature vector representation for the tags, a similarity measure and a linkage criterion.

The feature vector for a tag is composed of boolean values that correspond to the occurrence of that tag in a given question. Formally, for each tag T_i we have the following feature vector

$$\boldsymbol{x}_i = \begin{bmatrix} f(T_i, Q_1) \\ f(T_i, Q_2) \\ \vdots \\ f(T_i, Q_k) \end{bmatrix} \quad (1)$$

In (1) k stands for the total number of questions Q_j thus, determining the size of the vector $\boldsymbol{x}_i$ while the function f is defined as follows

$$f(T_i, Q_j) = \begin{cases} 1: if\ the\ question\ Q_j\ is\ tagged\ with\ T_i \\ 0: otherwise \end{cases} \quad (2)$$

The resulting vector is sparse so a suitable representation to improve memory and computational performance was used. We define the distance between two tags as the cosine similarity measure for two vectors

$$d(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{(\boldsymbol{x} \cdot \boldsymbol{i})(\boldsymbol{y} \cdot \boldsymbol{i})} \quad (3)$$

In (3) $\boldsymbol{x} \cdot \boldsymbol{y}$ is the standard dot product operation on two vectors $\boldsymbol{x} \cdot \boldsymbol{y} = \sum_{i=1}^{\|\boldsymbol{x}\|} x_i\, y_i$, and $\boldsymbol{i}$ is a vector of ones. As a linkage criterion we use mean linkage which defines the distance between two clusters A and B

$$D(A, B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a, b) \quad (4)$$

The algorithm stops when the distance between the two closest clusters is greater than some threshold value θ or when there is only one cluster left. By using different threshold values we can obtain different groupings. Imposing a lower threshold value will produce smaller, strongly connected clusters. On the opposite hand, if we use a higher threshold value we will end up with larger clusters containing more loosely coupled tags.

IV. Dataset

StackOverflow is a question and answer site for programmers and enthusiasts with close to 800,000 questions, each tagged with at least one or at most five tags. Tagging is not limited to the user posting the question, but instead the entire community can suggest tags for any question. On average a question has 3.14 tags. The tag clouds contains up to 70,000 tags. However, most of these tags are rarely used, so in our experiment we decided to exclude all the tags that occur less than 1000 times to make the problem less complex. These tags, 1300 in total, will be referred to as relevant as they account for 87% of all the tags on every question in the system. The most frequent tag is "c#" with over 400,000 occurrences whilst the least frequent tag with more than 1000 occurrences is "reverse-engineering" with just 1003 occurrences. These statistics were computed based on the StackOverflow full data dump made in September 2011.

V. Results

By applying the method described above to the set of relevant tags on StackOverflow we get outstanding results. Using a threshold value of $\theta = 0.9996$ a total number of 45 clusters were derived. The largest cluster contains 56 tags which occurrences when summed up account for 10% of all the tag occurrences, making this a huge cluster. The smallest cluster contains only 3 tags and as such accounts for only 0.3% of all tag occurrences. Some of the clusters included the following tags:

- *javascript, jquery, html, css, ajax, web-development, forms, internet-explorer, json, validation, firefox, web-applications* (the largest one)
- *oop, class, function, reflection, inheritance object, interface, parameters, enums, lambda, attributes, methods, static, properties, constructor, struct, naming-conventions*
- *exception, exception-handling, error-handling, error-message, try-catch*
- *performance, debugging, optimization, memory-leaks, memory-management*
- *Delphi, delphi-2009, delphi-2010* (the smallest one)

Manual inspection reveals that many of the clusters reflect real-life correlation between the different tags which represent tools, languages and concepts from the programming world.

VI. Conclusion

In this paper we recommend using an agglomerative clustering algorithm as a method for grouping correlated tags considering only the label part (output space) of the data. We apply this method on the StackOverflow tag cloud producing reasonably sized and meaningful clusters. The results show great promise. Several directions for further work can be followed. First, we will focus on building a more complete and complex tag hierarchy structure. Then we plan to use the similarity between certain tags expressed through the tag clusters to help in the task of automatic tag suggestion for the users.

Acknowledgments

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss.Cyril and Methodius: University.

References

[1] S. Golder and B. A. Huberman, *The Structure of Collaborative Tagging Systems*, Journal of Information Science 32(2), pp. 198–208, 2006

[2] C. Marlow, M. Naaman, D. Boyd, and M. Davis, *Position Paper, Tagging, Taxonomy , Flickr, Article, ToRead*, in Proceedings of the seventeenth conference on Hypertext and hypermedia, pp. 31-40, 2005

[3] B. Lund , T. Hammond , M. Flack and T. Hannay*, Social bookmarking tools: A case study – connote,* D-Lib Magazine, Volume II Number 4, April 2005

[4] C. H. Brooks and N. Montanez, *An analysis of the effectiveness of tagging in blogs,* In Proceedings of AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, pp. 9-14. 2006

[5] C. Hayes, P. Avesani, and S. Veeramachaneni, *An analysis of the use of tags in a blog recommender system*, in Proceedings of the 20th international joint conference on Artificial intelligence, pp. 2772-2777, 2007

[6] P.Heymann, G. Koutrika and H. Garcia-Molina, *Can social bookmarking improve web search?*, in Proceedings of the 2008 International Conference on Web Search and Data Mining , pp. 195-206, 2008

[7] N. Zheng and Q. Li, *A recommender system based on tag and time information for social tagging systems*, in Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 195-206 2011

[8] G. Begelman, P. Keller, and F. Smadja, *Automated Tag Clustering: Improving search and exploration in the tag space*, in Proceedings of the Collaborative Web Tagging Workshop at WWW'06, 2006

[9] C. Hayes, and P. Avesani, *Using Tags and Clustering to Identify Topic-Relevant Blogs*, in Proceedings of the 1st International Conference on Weblogs and Social Media (ICWSM 07), pp. 67-75, 2007

[10] Y. Song, B. Qiu, and U. Farooq, *Hierarchical Tag Visualization and Application for Tag Recommendations*, in Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 1331-1340, 2011

[11] P. Heymann, and G. H. Molina, *Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems*, Stanford Info Lab Technical Report, No. 2006-10, April 2006

[12] X. Liu and W.B. Croft, *Cluster-based retrieval using language models*, in Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 186-193, 2004

[13] X. Wei and W.B. Croft, *LDA-based document models for ad-hoc retrieval*, in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 178 – 185, 2006

[14] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey, *Scatter/gather: a cluster-based approach to browsing large document collections*, in Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, 1992

[15] O. Etzioni and O. Zamir, *Grouper: A dynamic clustering interface to web search results*, in Proceedings of the 8th International World Wide Web Conference, pp. 283-296, May 1999

[16] H.J. Zeng, Q.C. He, Z. Chen, W.Y. Ma, and J. Ma, *Learning to cluster web search results*, in Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 210-217, 2004

[17] Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W. C. Lee, and C. L. Giles, *Real-time automatic tag recommendation*, in Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 2008

[18] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme, *Evaluating similarity measures for emergent semantics of social tagging*, in Proceedings of the 18th international conference on World wide web, pp. 641-650, 2009

[19] S. A. Golder and B. A. Huberman, *Usage patterns of collaborative tagging systems*, Journal of Information Science*,* vol. 32 no. 2 198-20, April 2006

[20] R. J□Aschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme, *Tag recommendations in folksonomies,* in Proceedings of the 11th European conference on Principles and Practice of Knowledge, pp. 506 – 514, 2007

[21] B. M. Sarwar, G. Karypis, J. Konstan, and J. Riedl, *Recommender systems for large-scale e-commerce:* Scalable Neighborhood Formation Using Clustering, In Proceedings of the 5th International Conference on Computer and Information Technology, 2002

[22] J. Kelleher and D. Bridge, *An accurate and scalable collaborative recommender*, Journal Artificial Intelligence Review, Volume 21 Issue 3-4, pp. 193 – 213, June 2004

[23] M. O'Connor and J. Herlocker, *Clustering items for collaborative filtering*, in Proceedings of the 22nd International Conference on Research and Development in Information Retrieval, 1999