Efficiency and Stability of Clustering Algorithms for Linked Data

Isabel Drost and Tobias Scheffer
Humboldt-Universität zu Berlin
Department of Computer Science
Unter den Linden 6, 10099 Berlin, Germany
drost@informatik.hu-berlin.de, scheffer@informatik.hu-berlin.de

Abstract

We are interested in finding clusters ("communities") in networks of linked data, such as citation networks or web pages. Hierarchical clustering for networks is reviewed and an algorithmic improvement that leads to a significant performance increase is introduced.

Our main focus is on the development of partitioning clustering algorithms that can deal with data represented only by link information (*e.g.,* documents represented only by their citations) and the development of an EM algorithm for such data. A desirable property of clustering is stability; that is, small changes to the data should not lead to dramatically different clusterings. In our experiments with citation data we compare the hierarchical and partitioning clustering algorithms in terms of efficiency, stability and intra-cluster similarity.

The problem of mining linked data (*e.g.,* [4]) has become quite important as more and more information - such as scientific publications or simple web pages - is made available online. The most popular link mining tasks concentrate on finding communities in citation data [9] or in web pages from link topology [5]. Identification of terrorist networks [1; 8] and of fraud in telecommunication networks [2] are among the relevant applications which motivate research in this field.

Clustering is an elementary data analysis step that is well examined for traditional machine learning settings and is now also being applied to linked data. When analysing linked data, it seems obvious to represent each node of this network either by its inbound, by its outbound or by both kinds of links. One can distinguish hierarchical agglomerative [6] and flat, partitioning clustering algorithms (*e.g.,* [3]). Hierarchical clustering algorithms require a distance metric between pairs of instances to be defined, whereas k-means and EM with mixture models require the instances to be represented as a vector in feature space.

Up until now, in most cases an agglomerative clustering algorithm was employed when clustering linked data. Yet this kind of algorithm is known to be very time consuming when clustering large numbers of objects and has shown to be sensitive to perturbations of the data to cluster [7]. We examine the applicability of partitioning algorithms to linked data and compare their performance in terms of efficiency, stability and intra-cluster similarity to the agglomerative algorithm.

Our contribution is threefold. Firstly, we propose a caching strategy for hierarchical agglomerative clustering that improves its performance for clustering link data. Secondly, we derive an EM clustering algorithm for link data. Thirdly, we compare the clustering algorithms in terms of efficiency, stability and intra-cluster similarity for a publication database.

References

[1] W. Baker and R Faulkner. The social organization of conspiracy: illegal networks in the heavy electrical equipment industry. *Am. Social. Rev.*, 58:837–860, 1993.

[2] C. Cortes, D. Pregibon, and C. Volinski. Communities of interest. In *Proceedings of the International Symposium on Intelligent Data Analysis*, 2001.

[3] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39, 1977.

[4] Lise Getoor. Link mining: A new data mining challenge. In *SIGKDD Exploration 5*, 2003.

[5] David Gibson, Jon M. Kleinberg, and Prabhakar Raghavan. Inferring Web Communities from Link Topology. In *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, pages 225–234, Pittsburgh, Pennsylvania, June 1998.

[6] A. Griffiths, L. Robinson, and P. Willett. Hierarchical agglomerative clustering methods for automatic document classification. *Journal of Documentation*, 40(3):175–205, 1984.

[7] J. Hopcroft, O. Khan, and B. Selman. Tracking evolving communities in large linked networks. In *Proceedings of the SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[8] V. Krebs. Mapping networks of terrorist cells. *Connections*, 24(3):43–52, 2002.

[9] H. White and K. McCain. Bibliometrics. *Annual Review of Information Science and Technology*, 24:119–186, 1989.