Getting Clusters from Structure Data and Attribute Data

David Combe, Christine Largeron, Előd Egyed-Zsigmond, Mathias Géry

Getting clusters from structure data and attribute data

David Combe*, Christine Largeron*, Előd Egyed-Zsigmond†, Mathias Géry*
*Université de Lyon, F-42023, Saint-Étienne, France,
CNRS, UMR 5516, Laboratoire Hubert Curien, F-42000, Saint-Étienne, France
Université de Saint-Étienne, Jean-Monnet, F-42000, Saint-Étienne, France
Email: {david.combe, christine.largeron, mathias.gery}@univ-st-etienne.fr
†Université de Lyon
UMR 5205 CNRS, LIRIS
7 av J. Capelle, F-69100 Villeurbanne, France
Email: elod.egyed-zsigmond@insa-lyon.fr

***Abstract*—If the clustering task is widely studied both in graph clustering and in non supervised learning, combined clustering which exploits simultaneously the relationships between the vertices and attributes describing them, is quite new. In this paper, we present different scenarios for this task and, we evaluate their performances and their results on a dataset, with ground truth, built from several sources and containing a scientific social network in which textual data is associated to each vertex and the classes are known. We argue that, depending on the kind of data we have and the type of results we want, the choice of the clustering method is important and we present some concrete examples for underlining this.**

I. Introduction

Among the clustering methods, one can distinguish on the one hand the non supervised learning techniques, also called vector-based clustering, which exploit the attributes describing the objects, like hierarchical clustering or k-means and on the other hand those which consider the relationships between the different objects as it is usually the case in graph clustering. The goal of graph node clustering, related to community detection within social networks, is to create a partition of the vertices, taking into account the topological structure of the graph, such that the clusters are composed of vertices strongly connected [1], [2], [3], [4]. Among the core methods proposed in the literature, we can mention those that optimize a quality function to evaluate the goodness of a given partition, like the modularity, the ratio cut, the min-max cut or the normalized cut, hierarchical techniques like divisive algorithms based on the minimum cut, spectral methods or the Markov Clustering algorithm and its extensions.

Graph clustering techniques are very useful for detecting strongly connected groups in a graph but many of them mainly focus on the topological structure, ignoring the properties of the vertices. Nowadays, various data sources can be seen as graphs where vertices have attributes and a new challenge in graph clustering consists in combining structure data corresponding to the network and attribute data describing the vertices.

In this article, we focus on the clustering of scientist networks, mainly based on the publications and the participation in scientific events. We have textual data (publication titles, abstracts, full text, ...) and relationship data (co-authorship, co-participation in a same event). In order to detect strongly connected clusters containing persons with similar research interests, we have to exploit both attributes associated to each people and relationships between the members of the network. Depending on the weight allowed to each type of data, textual or structural, and of the way to combine them during the clustering, the results can be very different. In Section II we formally define the problem while we propose several approaches which consider simultaneously structure data and attribute data in section III. In section IV, we present an experimental study whose results confirm that clustering, based on structure and attribute data, provides more meaningful clusters than methods that take into account one type of data (text or structure).

II. Problem statement

We consider a graph $G = (V, E)$ where $V = \{v_1, \ldots, v_i, \ldots, v_{|V|}\}$ is the set of vertices and $E \subset V \times V$ is the set of unlabeled edges. The clustering process consists in partitioning the set V of vertices into r clusters $\mathscr{P} = \{C_1, \ldots, C_r\}$ such that:

- $\bigcup_{k \in \{1,..,r\}} C_k = V$
- $C_k \cap C_l = \emptyset$, $\forall\ 1 \leq k < l \leq r$
- $C_k \neq \emptyset$, $\forall k \in \{1, \ldots, r\}$

Moreover, we suppose that each vertex $v_i \in V$ is associated to a document represented by a vector $d_i = (w_{i1}, \ldots, w_{ij}, \ldots w_{iT})$ where w_{ij} is the weight of the term t_j in the document d_i. These documents can been seen as vertice attributes and G defined as an attributed graph [5].

In an attributed graph clustering problem, the structural links and the attributes are both considered, in such a way that:

- firstly, there should be many edges within each cluster and relatively few between the clusters;
- secondly, two vertices belonging to the same cluster are more similar in terms of attributes, than two vertices belonging to two different clusters.

Thus, the clusters should be well separated and, the vertices belonging to the same cluster should be connected and homogeneous on attribute data.

III. Attributed graph clustering approaches

We introduce different approaches to partition the graph using both the structural data and attribute data. The methods differ on the manner in which the relational and attribute data are combined.

A. Structure-based clustering on attribute weighted graph

Model TS1: we define a textual attribute-based distance dis_T, for instance the euclidean distance or the cosine distance, well suited for textual attributes. The value $dis_T(d_i, d_j)$ is associated to each edge (v_i, v_j) of E. Then, any weighted graph clustering algorithm can be used to partition the set of the vertices V. In our experiments, the cosine distance on the tf-idf vectors and Blondel algorithm were used [6].

B. Attribute-based clustering on structural distance

Model TS2: structural information is used, together with vertex attribute similarity to obtain a distance matrix (between each pair of vertices), which can then be processed by any classical unsupervised clustering algorithm. In our experiments, the cosine distance computed on the TF-IDF vectors is associated to each edge in order to obtain a weighted graph. Then, the geodesic distance between two vertices is defined as the smallest sum of the weights of the path edges between these vertices. Finally, a hierarchical agglomerative clustering is applied on the geodesic distance matrix, using usual distance between clusters: single link, complete link, average link and center of gravity.

C. Hybrid clustering

Model TS3: attributes and structure are considered separately in order to compute a distance on each type of data. These distances are then combined into a global distance that can be exploited by any unsupervised clustering algorithm or used to obtain a valued graph which can be processed by any weighted graph clustering algorithm. In our experiments, we used the cosine distance on textual information and geodesic distance on the graph G. Then a hierarchical agglomerative clustering is applied with the global distance matrix defined as a linear combination of the previous distances.

IV. Experimental study

A. Network data

In order to evaluate the different methods presented previously, we have built a data set with a ground truth. So we are able to compare the community of each vertex with its cluster provided by the methods. We concentrated on two conferences: SAC 2009 and IJCAI 2009. A co-participation network was generated from the well-known DBLP[1] dataset and the abstracts, titles and research areas were extracted from the websites of the selected conferences.

[1]http://www.informatik.uni-trier.de/~ley/db/

1) Authors and research areas: Three research areas, corresponding to conference sessions, were selected: Robotics, Bioinformatics and Constraint Programming. In both conferences there is a Robotics session, while only SAC 2009 has a session on Bioinformatics and IJCAI 2009 on Constraint Programming. There are 99 authors in the four sessions. Each of these authors corresponds to one vertex of V and its research area membership is used during the evaluation step.

The abstracts and the titles of the articles published by the authors at IJCAI 2009 and SAC 2009 are represented in the vector space model introduced by Salton *et al.* [7]. After a preprocessing of the text with stemming and stopword removal, an attribute vector d_i, in which the components are computed with the tf-idf formula, is attached to each author of V.

2) Social Network: We consider an event e as a journal or a conference referenced in DBLP between 2007 and 2009. A co-participation network is built on the set V, using the DBLP database, as follows. Let v_i and v_j be two authors belonging to V, if there exists at least one event e such that v_i and v_j are authors for articles published in e (even if they are not co-authors), then (v_i,v_j) $\in E$.

3) Graph: We obtain the attributed graph $G = (V, E)$ having the vertices created with the authors and the edges given by the co-participation relations. Moreover, each vertex (*i.e.* author), is described by textual attributes corresponding to the tf-idf vector associated to his articles and, its true class is the research area (*i.e.* the session (A, B, C or D) in SAC 2009 or IJCAI 2009)) of this author.

B. Hypotheses

We enumerate here our clustering scenarios and hypothesis and present the foreseen results. We consider 4 vertex subsets, given by the authors publishing in the 4 extracted sessions:

- A: Bioinformatics (SAC): 24 authors
- B: Robotics (SAC): 16 authors
- C: Robotics (IJCAI): 38 authors
- D: Constraint Programming (IJCAI): 21 authors

1) Text: 3 research areas / 3 clusters (P_T): Considering only textual vertex attributes, the hypothesis underlying our experiments is that this information should permit to retrieve the three research areas: Robotics, Bioinformatics and Constraint Programming, giving the partition into three clusters containing the authors of the three research areas: $P_T = \{A, B \cup C, D\}$.

2) Structure: 2 conferences / 2 clusters (P_S): On the other hand, we suppose that taking into account only structural data should allow to identify two groups corresponding to authors participating to each conference: SAC2009 and IJCAI2009, which define the partition into two clusters $P_S = \{A \cup B, C \cup D\}$.

3) Text and structure: 4 sessions / 4 clusters (P_{TS}): However, if we want to discover each session separately, both textual and structural information have to be used. In this case the partition will be into four clusters $P_{TS} = \{A, B, C, D\}$.

TABLE I
RESULTS SYNTHESIS: MODELS T, S, TS_1, TS_2 AND TS_3

Model	Accuracy considering: P_T	P_S	P_{TS}
T	87%	-	69%
S	-	100%	63%
TS_1	-	-	76%
TS_2	-	-	73%
TS_3	-	-	47-69%

C. Evaluation

In order to check these hypotheses, we evaluate several methods combining text and structure (models TS_1, TS_2, TS_3), corresponding to the different approaches detailed in Section III. We compare also our models against two baselines: clustering based on text only (model T) and clustering based on structure only (model S).

The different methods were evaluated using the accuracy of the obtained clusters, compared to the ground truth considered: research areas (P_T), conferences (P_S) or sessions (P_{TS}). The results are synthetized in Table I.

Text-only based clustering (Model T): As expected, the accuracy is higher for the partition in three clusters P_T ($\frac{(11+16+38+21)}{99} \times 100 = 87\%$) than for the partition in four clusters P_{TS} (69%). This result confirms our hypothesis according which the textual data allows to identify the different research areas but fails to detect correctly the four sessions.

Structure-only based clustering (Model S): The identification of the two conferences using structural data is perfectly achieved. However, the accuracy is only equal to 63% if we consider the four sessions as the ground truth (P_{TS}).

Structure-based clustering on attribute weighted graph (Model TS_1): Taking into account structural and attribute data improves the accuracy which reaches 76% for the partition in four clusters (P_{TS}), when it is only equal to 69% without attribute data. This result confirms our hypothesis according which the two types of information are useful to improve the classification accuracy.

Attribute-based clustering on structural distance (Model TS_2): With a classification accuracy of 73% for the partition in four clusters (P_{TS}), the results are similar to those obtained with the modularity based algorithm and higher than those obtained using only one type of information (textual or structural).

Hybrid clustering (Model TS_3): Even if this method appears as a simple solution for exploiting simultaneously the two types of data, it is not so easy to use since it requires to set the parameter α in the linear function. Moreover, in our experiments, the accuracy for the partition in four clusters (P_{TS}) varies in function of α between 47% (α set to 0.85, 0.96) and 69% (α set to 1).

Thus, the best accuracy corresponds to those obtained with a text-based clustering and it is not so good than those obtained with the other methods combining structural data and attribute data.

V. CONCLUSION

As shown in the previous section, we obtain very different results according to the clustering method and the data taken into account when partitioning an attributed graph.

In our experiments, text attribute based clustering enables quite well to retrieve the research interests, structure based clustering allows to identify the conferences and, finally, the structural information and textual information are useful to retrieve the four sessions corresponding to groups of participants in one conference who share a common interest. This result confirms the fact that the accuracy of the clustering can be improved by taking into account the vertex properties and the relationships of the network to detect groups of vertices strongly connected and similar in terms of attributes. Moreover, it seems that the combination of the different data types is not obvious. More particularly, hybrid method based on linear combination is not the best suited. The other hybrid methods presented in this article give better results and they are easier to apply since they do not need to set a parameter.

ACKNOWLEDGMENT

This work was partially supported by St-Etienne Metropole (http://www.agglo-st-etienne.fr/) and by the Région Rhône-Alpes.

REFERENCES

[1] P.-O. Fjällström, "Algorithms for graph partitioning: A survey," *Science*, vol. 3, no. 10, 1998.
[2] M. Newman, "Detecting community structure in networks," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 38, no. 2, pp. 321–330, 2004.
[3] S. Schaeffer, "Graph clustering," *Computer Science Review*, vol. 1, no. 1, pp. 27–64, 2007.
[4] A. Lancichinetti and S. Fortunato, "Community detection algorithms: a comparative analysis," *Physical review E*, vol. 80, no. 5, p. 056117, 2009.
[5] Y. Zhou, H. Cheng, and J. Yu, "Graph clustering based on structural/attribute similarities," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 718–729, 2009.
[6] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, 2008.
[7] G. Salton and M. J. McGill, *Introduction to modern Information Retrieval*. McGraw-Hill, 1983.