A Resource-tag Attribute Graph Clustering Algorithm

Xiaohui Ma*, Juan Chen, Hua Yu

College of Information, Business College of Shanxi University, Taiyuan 030031, Shanxi, China

Abstract
In order to improve the effectiveness of clustering, a resource-tag attribute graph clustering algorithm is proposed in this paper, which makes a two-stage of clustering by the use of attribute tags of nodes in graphs. Firstly tags clusters are generated based on the weight and similarity of tags, which is the preprocessing of tags. Secondly the attribute vectors of nodes in resources are tested and calculated and then according to the pre-programmed threshold value the corresponding clustering sub-graphs join in order to increase the contact edge of nodes. Finally a related comparative test is made between fuzzy clustering algorithm, k-means algorithm and algorithm proposed in this paper. The data is a top250 data set from Douban reading. It turns out that in this algorithm F value can achieve much high and clustering entropy tends to zero. The effectiveness and accuracy of the algorithm are verified.

Key words: Social Tagging System, Tag Co-occurrence, Attributed Graph Clustering

1. INTRODUCTION
Also known as the Social Tagging System or Folksonomy, collaborative Tagging System (CTS) refers to the system allowing users to use any keyword (also called tag) to annotate resources(Golder and Huberman,2006). In CTS users could make personalized descriptions and annotations according to their own understanding of resources without predefined vocabularies and classification approaches. In recent years the application of CTS in Flickr, DeLicio.us, YouTube, Douban et al. has achieved great success, thus CTS has enjoyed much popularity from more and more users.

Collaborative Tagging System helps make effective combinations of users and their cognition on resources. However, some problems such as semantic fuzziness and low accuracy have inevitably arisen since the tags are annotated transparently and freely by users. Moreover, with a growing size of tags, retrieval and classification of resources through vast amounts of tags has become a hot research topic.

As a special kind of data structure, graph is appropriate for describing all kinds of complex system models, because it can depict the complex relationships between entities, such as social network, road network, etc. Graph clustering refers to the process in which the nodes in the graphs will be divided into different groups according to certain rules, for example nodes similarity principle and point distance principle, etc., and then users will get the clustering results of specific sub-graphs.

Clustering research on tagging graphs is in favor of effective classifications and retrievals of resources, thus of mining of users’ interests and making personalized recommendations. The paper firstly makes extraction of tags in certain resources, and then carries out analysis of co-occurrence on these tags; secondly through the method of graph clustering it clusters tags and tagged resources; finally it compares different clustering effects of different clustering methods.

2. RESEARCH REVIEW
Research and applications of tag clustering technology in the field of CTS mainly include: K-Means clustering algorithm and its variations and improvements (Sigurbjornsson and Zwol,2008; Wagstaff,Cardie et al.,2001). As for the improved K-Means algorithm, some literature mentioned certain requirements that tags should distribute throughout every class; if this premise cannot be guaranteed, the clustering effect will not be ideal. Gelbard put forward a kind of hierarchical clustering algorithm, which helped make relative and effective improvement of the accuracy of clustering results(Gelbard, Goldman, Spiegler,2007). This kind of algorithm stored data set to be clustered in the matrix in binary format, so it was also called binary-positive algorithm. After applying the fuzzy set theory to clustering analysis for the first time in 1969, Ruspini came up with fuzzy c-means. Cai et al. raised FGFCM algorithm after a combination of gray scale and local spatial information(Sunet et al.,2008). As we can see, most present clustering algorithms primarily take the distribution of node degree in graphs, structure of graphs, etc. into consideration, as shown in the graph clustering methods based on minimum cuts within the graph(Flakeet al.,2004), the clustering through attribute value of nodes(Pizzuti,2009).
From the existing research results, we can see clearly that the successful clustering of tags helps accelerate orderly organization of tags, thus realize effective classifications of tagged resources. But most clustering algorithm considering only one attribute or condition, this will result in the low efficiency of the algorithm.

The clustering algorithm proposed in this paper not only takes the node attribute into consideration, but also the graph structure. Regarding resources as clustering origin, tags as its property, top250 in Douban reading as a data set, the algorithm makes a two-stage clustering of attribute tags and resource nodes with a combination of some thoughts of graph clustering. It turned out that the effectiveness and accuracy were better than the compared algorithms.

3. PROBLEM DESCRIPTION

In resource-tag attributes graph, nodes are used to describe resources and as attributed information of nodes, tags are used to describe the characteristics of resources. If we’d like to classify resource nodes into different clusters, the first step is to put the resources of highly-similar characteristics into a cluster based on the similarity of attributed group (tag) of resources. The major task of this method is as follows: how to make calculation of the similarity to these extracted tags, and how to make clustering of resource graphs according to the similarity information of tags because clustering of resources is based on clustering of tags’ attributes.

Definition 1 is about CTS. It is described as $S= (U, T, R, A)$, in which $U$, $T$, and $R$ respectively stands for three elements in CTS: user, tag and resource. $A \subseteq \{ (u, t, r); u \in U, t \in T, r \in R \}$ is the relational set of the three elements and the element $a=(u,t,r)$ takes a Boolean value. If a user uses tag $t$ to annotate the resource $r$, then a equals 1 else equals 0.

Definition 2 is about attribute graph. Graph $G$ is described as a triple: $G = (V, E, A)$. $V = \{ v_1, v_2, \ldots v_n \}$ is the gather of several ($n$) vertexes; $E = \{ (v_i, v_j) | 1 \leq i, j \leq n, i \neq j \}$ is the gather of edges; $A = \{ a_1, a_2, \ldots, a_m \}$is the gather of several ($m$) attributes of nodes.

Definition 3 is about graph clustering. For graph $G$, $G' = (V', E', A')$ is called a cluster of graph $G$, in which $V' \in V, E' \in E, A' \in AG' \neq \emptyset \forall G_i', G_j' \in G, i \neq j, G_i' \cap G_j' = \emptyset$.

4. CLUSTERING ALGORITHM

In collaborative tagging system the phenomenon will appear that the same resources are annotated with different tags. Michlmayr pointed out that this kind of phenomenon will inevitably lead to semantic relevance between tags (Michlmayr and Cayzer, 2007). In other literaturea clustering algorithm based on tag co-occurrence was proposed, helping generate tag clustering clusters (Wang and Zhang, 2010).

Define 4 is about marked matrix. The matrix $U_{m \times n}$ shown in Table 1 is a marked matrix, in which $m$ is the tag number, $n$ is the resource number and $q_{ij}$ shows how many times the resource $r$ is annotated by tag $t$.

<table>
<thead>
<tr>
<th></th>
<th>$r_1$</th>
<th>$r_2$</th>
<th>$r_j$</th>
<th>...</th>
<th>$r_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>$q_{11}$</td>
<td>$q_{12}$</td>
<td>$q_{1j}$</td>
<td>...</td>
<td>$q_{1n}$</td>
</tr>
<tr>
<td>$t_2$</td>
<td>$q_{21}$</td>
<td>$q_{22}$</td>
<td>$q_{2j}$</td>
<td>...</td>
<td>$q_{2n}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$t_l$</td>
<td>$q_{l1}$</td>
<td>$q_{l2}$</td>
<td>$q_{lj}$</td>
<td>...</td>
<td>$q_{ln}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$t_m$</td>
<td>$q_{m1}$</td>
<td>$q_{m2}$</td>
<td>$q_{mj}$</td>
<td>...</td>
<td>$q_{mn}$</td>
</tr>
</tbody>
</table>

Define 5 is about tag co-occurrence matrix. The matrix shown in Table 2 is such a matrix, in which $f_{ij}$ stands for the co-occurrence frequency of tag $t_i$ and $t_j$. The higher the frequency, the higher similarity of tag $t_i$ and $t_j$.

<table>
<thead>
<tr>
<th></th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$t_j$</th>
<th>...</th>
<th>$t_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>$f_{11}$</td>
<td>$f_{12}$</td>
<td>$f_{1j}$</td>
<td>...</td>
<td>$f_{1m}$</td>
</tr>
<tr>
<td>$t_2$</td>
<td>$f_{21}$</td>
<td>$f_{22}$</td>
<td>$f_{2j}$</td>
<td>...</td>
<td>$f_{2m}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$t_j$</td>
<td>$f_{j1}$</td>
<td>$f_{j2}$</td>
<td>$f_{jj}$</td>
<td>...</td>
<td>$f_{jm}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$t_m$</td>
<td>$f_{m1}$</td>
<td>$f_{m2}$</td>
<td>$f_{mj}$</td>
<td>...</td>
<td>$f_{mm}$</td>
</tr>
</tbody>
</table>

Equation (1) can be used to calculate the co-occurrence frequency of the two tags, $w(t_i, t_j)$ in which stands for the co-occurrence times of tag $t_i$ and $t_j$.  

128
Define 6 is about tag weight matrix. The matrix shown in Table 3 is such a matrix, in which \( w_{ij} \) stands for importance degree of tag \( t_i \) in several (n) resources. Equation (2) is available to calculate the degree.

### Table 3: Tag important degree matrix

<table>
<thead>
<tr>
<th></th>
<th>( t_1 )</th>
<th>( t_2 )</th>
<th>( \ldots )</th>
<th>( t_i )</th>
<th>( \ldots )</th>
<th>( t_m )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t_1 )</td>
<td>( w_{11} )</td>
<td>( w_{12} )</td>
<td>( \ldots )</td>
<td>( w_{1i} )</td>
<td>( \ldots )</td>
<td>( w_{1m} )</td>
</tr>
<tr>
<td>( t_2 )</td>
<td>( w_{21} )</td>
<td>( w_{22} )</td>
<td>( \ldots )</td>
<td>( w_{2i} )</td>
<td>( \ldots )</td>
<td>( w_{2m} )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td></td>
</tr>
<tr>
<td>( t_i )</td>
<td>( w_{i1} )</td>
<td>( w_{i2} )</td>
<td>( \ldots )</td>
<td>( w_{ii} )</td>
<td>( \ldots )</td>
<td>( w_{im} )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td></td>
</tr>
<tr>
<td>( t_m )</td>
<td>( w_{m1} )</td>
<td>( w_{m2} )</td>
<td>( \ldots )</td>
<td>( w_{mi} )</td>
<td>( \ldots )</td>
<td>( w_{mm} )</td>
</tr>
</tbody>
</table>

\[
   w_{ij} = f_{ij} \times \log\frac{n}{\text{numbers of Resources tagged by } t_i}
\]

If the same resources are annotated by two tags with a high frequency, then there is a much high similarity between the two tags. The cosine similarity here is adopted to evaluate the similarity of tag. Let's set vector A and vector B as \( A = (a_1, a_2, \ldots, a_n), B = (b_1, b_2, \ldots, b_n) \), and Equation (3) is available to calculate the cosine similarity of A and B.

\[
   \text{Sim}(A, B) = \frac{A \cdot B}{|A| \times |B|} = \frac{\sum_{i=1}^{n} a_i \times b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} b_i^2}}
\]

If in the same cluster node attribute appears with high similarity and tight structure, and in different clusters there is a major difference between nodes attribute value, then this kind of effect is called an ideal clustering effect (Wu, Zhong and Xiong, et al. 2013). A graph clustering algorithm based on attribute tag clustering is proposed in this paper, in which in the process of node clustering two partition principles are followed according to tag propagation algorithm thoughts (Raghavan, Albert and Kumara, 2007). Principle 1: if the descriptions of attribute tags of two nodes are consistent with each other and the two tags are in the same tag cluster, then the two nodes should be divided into the same cluster. Principle 2: if most neighbor nodes of the two nodes are the same, then the two nodes should also be divided into the same cluster.

#### Algorithm 1 Resource Tag Graph Clustering Algorithm

**Input:** resource node set with attribute tags; \( G = (V, E, A) \) \( V = \{v_1, v_2, \ldots, v_n\} \) is a collection of several (n) vertexes; \( A = \{a_1, a_2, \ldots, a_m\} \) is a collection of several (m) attribute tags of node \( v_i \); \( E = \{\} \) is a collection of edges.

**Output:** several (k) clustering result of graph \( G, R = G_1, G_2, \ldots, G_k \) in which \( \forall G_i, G_j \subseteq G, i \neq j, G_i \cap G_j = \emptyset \).

1. **Step 1:** For all the tag and resource, set \( R = \emptyset \).
2. **Step 2:** Data preparation; tag pretreatment; determining several (M) attribute tag as division tags of a graph clustering.
3. **Step 3:** According to the Equation (1) and Equation (2) get the tag weight matrix.
4. **Step 4:** According to the Equation (3) calculate the similarity of the tags.
5. **Step 5:** According to calculation results, put the tags into several (N) tag clusters.
6. **Step 6:** According to tag cluster choose several (K) resource nodes as the initial clustering center, and number several (K) nodes sequentially, written as \((u_1, u_2, \ldots, u_k)\).
7. **Step 7:** Suppose nodes, \( v \in V \), \( y(v) = (a_1(v), a_2(v) \ldots a_i(v)) \) as the attribute vector of nodes; calculate the vector similarity of each node in \( y(v) \) and the initial several (K) nodes, thus we get \( \text{Sim}(y(u_1), y(v)), \text{Sim}(y(u_2), y(v)), \ldots, \text{Sim}(y(u_k), y(v)) \). Add node \( v \) into node in order to get the biggest Sim values, and increase the non-directional edges between nodes.
8. **Step 8:** Update the attribute tag clustering of nodes in the cluster graph.
9. **Step 9:** Repeat step 7 and 8 until all resources are clustered into corresponding clustering sub-graphs and get \( R = G_1, G_2, \ldots, G_k \).
10. **Step 10:** End
5. EXPERIMENT AND RESULT ANALYSIS

5.1. Experimental data Set

The data in this experiment comes from a typical collaborative tagging system—Douban Reading. Actually the data set is a book list of top250 extracted from pages in April of the year 2015 on Douban (website is: https://book.douban.com/top250?icn=index-book250-all). The book list is produced based on a comprehensive evaluation on these books including how many times every book was read and what kinds of evaluations did every book get. This paper chooses the top250 of the book list and their corresponding 2500 tag information as experimental data set. In order to have a more objective evaluation on the clustering algorithm proposed in this paper, a comparison about cluster quality was made between the K-Means clustering algorithm, SNAP clustering algorithm and the above one.

5.2. Evaluation Index

When it comes to evaluation on clustering results, this paper adopts F-Measure and Clustering Entropy as evaluation indexes. Calculation of F-Measure includes the Precision Rate as shown in Equation (4) and Recall Rate in Equation (5). Actually a harmonic mean of the above two, the F-Measure can be calculated with Equation (6).

\[
\text{Precision} = \frac{\text{apriority i and clustered to } i}{\text{apriority i and clustered to } i + \text{apriority j and clustered to } i}
\]

(4)

\[
\text{Recall} = \frac{\text{apriority i and clustered to } i}{\text{apriority i and clustered to } i + \text{apriority i and clustered to } j}
\]

(5)

\[
F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

(6)

Equation (7) is a definition of the probability about node members of cluster i belong to cluster j, in which \( m_i \) is the number of members in clustering i, and \( m_{ij} \) is the number of members in cluster i belonging to the cluster j. The calculation of every cluster entropy can be achieved by Equation (8), and the calculation of the whole cluster partition entropy by Equation (9). K stands for clustering number.

\[
P_{ij} = \frac{m_{ij}}{m_i}
\]

(7)

\[
e_i = -\sum_{j=1}^{K} P_{ij} \log_2 P_{ij}
\]

(8)

\[
e = \sum_{i=1}^{K} \frac{m_i}{m} e_i
\]

(9)

5.3. Analysis of Experimental Results

The algorithm clusters top250 books derived from Douban reading based on attribute tag. In order to enhance strong ties within the tag cluster, we set the threshold of tag dividing as 0.6. In addition, every tag cluster contains 10 tags of strong ties and the number of tag set is 50. Considering the execution efficiency of the algorithm, we define the K value of node clustering as 5 in the process of clustering of resource nodes.

Figure 1, 2 and 3 respectively shows us a comparison of recall ratio, precision and F value of three kinds of clustering algorithm. Node value takes an average value of the three test results.

![Figure 1. Recall ratio comparison of algorithms](image-url)
It can be seen that with the increase of sample data sets, the accuracy, recall rate and F value (calculated by the former two) of the three algorithms also increase a lot. What’s more, the clustering effect of algorithm proposed in this paper is better, because not only the process of clustering is divided into two stages, but the algorithm also takes both the node attribute and graph structure into consideration, which is significantly superior to the clustering algorithm considering only one attribute.

An ideal clustering sub-graph should include a similar attribute set and a relatively balanced graph structure. Besides, the vertexes in generated cluster have strong cohesion and uniform distribution in clusters. As for five clusters eventually produced by algorithm in this paper, tag attribute value of every type is basically similar and entropy value is close to zero because of the tag pre-classification. However, the entropy values of the other two algorithms are higher. With the existence of different attribute value, some nodes with similar tag values are put into different classes in the process of clustering. And Figure 4 show that, as the clustering number increase, tag attributes joining the sub-classes have the tendency of convergence and nodes in this class have more tense ties, thus the overall entropy value accordingly decrease.

5. CONCLUSIONS

This paper has proposed a two-stage clustering algorithm based on graph. By the algorithm we dug tags with strong dependency from attribute tag set to form tag clusters, and then we regarded tag cluster as a set of
attributes from which we extracted initial clustering nodes of attribute graphs. In order to improve the clustering accuracy, we have set certain threshold in the clustering of attribute tags and nodes. A comparative experiment between the algorithm in this paper and the other two algorithms has been made on a data set of top250 books in Douban reading. The results have showed that the algorithm in this paper is effective because it has obvious advantages in multiple indexes of evaluation. The next stage mainly focuses on the study of cleaning rules of tag choosing in order to improve the accuracy of clustering and the study of balanced structure of sub-graph in clustering graph.

Acknowledgements

This work was supported by fundamental research project of the Shanxi Science and Technology Department. (No. 2014011018-1)

REFERENCES


