Hui-zong Li, Xue-gang Hu, Yao-jin Lin, Wei He, Jian-han Pan, 2016. A social tag clustering method based on common co-occurrence group similarity. *Frontiers of Information Technology & Electronic Engineering*, **17**(2):122-134. http://dx.doi.org/10.1631/FITEE.1500187

A social tag clustering method based on common co-occurrence group similarity

Key words: Social tagging systems, Tag co-occurrence, Spectral clustering, Group similarity

Contact: Hui-zong Li
E-mail: lihz_aust@sina.com
ORCID: http://orcid.org/0000-0002-1459-989X

Motivation/Main ideas

- **Motivation**

  Many ambiguous and uncontrolled tags produced by social tagging systems not only worsen users' experience, but also restrict resources' retrieval efficiency. Tag clustering could aggregate tags with similar semantics together, and help mitigate the above problems.

- **Main ideas**
  - Analyze the co-occurrence relationship of tags from the ternary annotation relation and discover the most valuable tag co-occurrence status.
  - Present a common co-occurrence group similarity to measure the semantic relevance between tags.
  - Use a spectral clustering algorithm to aggregate the tags.

Method (I)

1. Tag co-occurrence

Definition 1 (Tag co-occurrence) Let $S = (U, R, T)$ be a social annotating data set, where $U = \{u_1, u_2, \ldots, u_l\}$ is a set of users, $R = \{r_1, r_2, \ldots, r_m\}$ a set of resources, and $T = \{t_1, t_2, \ldots, t_n\}$ a set of tags. The binary annotating relation set based on users and resources is decided by A_{t_i}, $A_{t_i} = \{< u_p, r_q > | u_p \in U, r_q \in R, t_i \in T, < u_p, r_q, t_i > \in S\}$, and $A_{t_i} \subseteq U \times R$. If A_{t_i} and A_{t_j} have the same element, namely $(A_{t_i} \cap A_{t_j}) \neq \varnothing$, we call tag t_i a co-occurrence with t_j, signed as $\mathrm{Co}(t_i, t_j)$.

Method (II)

2. Tag's individual co-occurrence similarity

Definition 2 (Tag individual co-occurrence similarity) Given two tags t_i and t_j, their binary annotating relation sets A_{t_i} and A_{t_j} have the same element. In this situation, we call the two tags t_i and t_j individual co-occurrence. The symbol $\mathrm{sim}_{\mathrm{indiv}}(t_i, t_j)$, used to represent the individual co-occurrence similarity between two tags, can be measured by the Jaccard coefficient. Therefore, the tag individual co-occurrence similarity is defined as follows:

$$\mathrm{sim}_{\mathrm{indiv}}(t_i, t_j) = \frac{|A_{t_i} \cap A_{t_j}|}{|A_{t_i} \cup A_{t_j}|}, \tag{3}$$

where $|A_{t_i} \cap A_{t_j}|$ is the number of elements in the intersection set of A_{t_i} and A_{t_j}, representing the individual co-occurrence frequency between tags t_i and t_j, and $|A_{t_i} \cup A_{t_j}|$ is the number of elements in the union set of A_{t_i} and A_{t_j}, representing the total using frequency of tags t_i and t_j.

Method (III)

3. Tag's common co-occurrence group similarity

Definition 3 (Tag common co-occurrence group similarity) Given two tags t_i and t_j, and $\mathrm{Co}(t_j, t_j)$, let $C = \{t_y \,|t_y \in T\}$ be a common co-occurrence group tag set of t_i and t_j, and $C \subseteq T$. We use t_c^k to represent the kth tag in C, and symbol $\mathrm{sim}_{\mathrm{group}}(t_i, t_j)$ to represent the common co-occurrence group similarity between t_i and t_j. The common co-occurrence group similarity between the two tags can be measured by an improved Jaccard coefficient. The tag common co-occurrence group similarity is defined as follows:

$$\mathrm{sim}_{\mathrm{group}}(t_i, t_j) = \frac{1}{|C|} \sum_{k=1}^{|C|} \frac{|(A_{t_i} \cup A_{t_c^k}) \cap (A_{t_j} \cup A_{t_c^k})|}{|(A_{t_i} \cup A_{t_c^k}) \cup (A_{t_j} \cup A_{t_c^k})|}. \tag{4}$$

Method (IV)

4. Tag spectral clustering algorithm

Algorithm 1 Tag spectral clustering based on common co-occurrence group similarity

Require: Dataset A_T, the number of clusters k

Ensure: Clusters $T_1, T_2, \ldots, T_k$

1: Generate the tag similarity matrix $\boldsymbol{W} \in \boldsymbol{\Phi}^{n \times n}$ from A_T based on Eq. (4) and set $W_{ii} = 0$.

2: Compute the unnormalized Laplacian matrix $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, where $\boldsymbol{D}$ is a diagonal matrix satisfying $D_{ii} = \sum_{ij} w_{ij}$.

3: Compute the normalized Laplacian matrix $\boldsymbol{L}' = \boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{1/2}$.

4: Compute the first k eigenvectors of $\boldsymbol{L}'$, i.e., $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k$.

5: Let $\boldsymbol{E} \in \boldsymbol{\Phi}^{n \times k}$ be the matrix containing vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k$ of $\boldsymbol{L}'$ as columns.

6: Form the matrix $\boldsymbol{R} \in \boldsymbol{\Phi}^{n \times k}$ from $\boldsymbol{E}$ by normalizing the rows to norm 1, which is set as $r_{ij} = e_{ij} / (\sum_k e_{ik}^2)^{1/2}$.

7: For $i = 1, 2, \ldots, n$, let $\boldsymbol{t}_i \in \boldsymbol{\Phi}^k$ be the vector corresponding to the ith row of $\boldsymbol{R}$.

8: Cluster the points $(t_i)_{i=1,2,\ldots,n}$ with the K-means algorithm into clusters $T_1, T_2, \ldots, T_k$.

9: Return $T_1, T_2, \ldots, T_k$.

Major results

Table 2 Comparisons of results under different clustering methods

Clustering algorithm	Similarity measure method	SC/Dunn				
		$k = 40$	$k = 60$	$k = 80$	$k = 100$	$k = 120$
K-means algorithm	Tag's resource VSM	0.357/0.504	0.432/0.504	0.521/0.389	0.634/0.504	0.591/0.504
	Tag's user VSM	0.160/0.376	0.240/0.376	0.292/0.337	0.299/0.337	0.385/0.244
	Tag's resource co-occurrence	0.334/0.358	0.424/0.525	0.580/0.358	0.446/0.525	0.651/0.525
	Tag's user co-occurrence	0.145/0.287	0.278/0.382	0.262/0.382	0.316/0.343	0.369/0.287
Agglomerative hierarchical clustering algorithm	Tag's resource VSM	0.328/0.504	0.415/0.504	0.551/0.458	0.548/0.389	0.645/0.458
	Tag's user VSM	0.120/0.376	0.255/0.376	0.198/0.244	0.233/0.337	0.266/0.244
	Tag's resource co-occurrence	0.312/0.358	0.358/0.525	0.510/0.431	0.557/0.358	0.628/0.525
	Tag's user co-occurrence	0.152/0.382	0.175/0.382	0.218/0.343	0.322/0.287	0.287/0.382
Spectral algorithm	Tag's common co-occurrence group	**0.380/0.668**	**0.438/0.668**	**0.407/0.585**	**0.689/0.668**	**0.782/0.668**

Conclusions

- Use the common co-occurrence group similarity to measure the relevance of tags based on tag co-occurrence and use a spectral clustering algorithm to aggregate the tags
  - An improved Jaccard coefficient based on the common co-occurrence group is proposed to calculate the similarity of tags.
  - A spectral clustering method is used for tag clustering.

- Implement tag relevance measure on a real world dataset from CiteULike, and compare the tag spectral clustering method with other different clustering approaches. Experimental results show that the proposed method achieves good performance in tag clustering.