User and Document Group Approach of Clustering in Tagging Systems

Rong Pan, Guandong Xu and Peter Dolog
IWIS — Intelligent Web and Information Systems
Department of Computer Science
Aalborg University
{rpan, xu, dolog}@cs.aau.dk

Abstract

In this paper, we propose a spectral clustering approach for users and documents group modeling in order to capture the common preference and relatedness of users and documents, and to reduce the time complexity of similarity calculations. In experiments, we investigate the selection of the optimal amount of clusters. We also show a reduction of the time consuming in calculating the similarity for the recommender systems by selecting a centroid first, and then compare the inside item on behalf of each group.

keywords: User Profile, Document Profile, Spectral Clustering, Group Profile, Modularity Metric

1 Introduction

The success of social tagging resulted in the proliferation of sites like Delicious, CiteUlike, Digg, or Flickr. Such sites contain large amount of user tagged data for information retrieval in social-tagging systems [6; 7; 12; 16; 17], or for the establishment of user profiles and the discovery of topics, among other applications.

[5] uses the tags associated with specified objects to build a single user profile. However, here comes a problem: it is hard to express the entire user profile or the document profile. The traditional user profile expresses the users' preferences depending on collecting users' behaviors information, such as provides many tedious options in their registrations. The disadvantage with such an approach is too much reliance on users who is not able very often to express his entire user profile and interests. The document profile shows the background, categories, and keywords, it also depends on the description when it is added into the system. However, with the increase of the number and types of users, it's hard to express the different emphases for various users with the same document.

In social collaborative (tagging) systems, the common perception or judgment on documents are determined by a group of users rather than a single user. In a similar way, by using a group of documents, rather than one document, it might represent much more specific information during the information search.

Therefore, our assumption is that by utilizing the community views of users and documents, we are able to facilitate the organization of information resources in search and navigation.

In social tagging systems, users express their judgment by annotations or tagging. The tag can endorse their opinions on various web items, which is one of the defining characteristics of Web 2.0 services, allowing them to collectively classify and index information for later search and sharing. With social tagging, a user can express his own perspective on web items, e.g. resources like images, videos, scientific papers, thus allowing other like-minded users to find and use the similar information.

The tagging has been already utilized for organizing the resources. [3] develops a page rank algorithm of resources based on preference tag vectors. [6; 8; 7] investigate social and behavioral aspects of a tag-based recommender system which suggests similar web pages based on the similarity of users' tags. However, there is another problem emerging: not all of the social tagging systems proposed so far maintain high quality and quantity of tag data. It is particularly prominent when a new user enters the system or a new document is added into the system.

If the individual user profile or document profile can be collected and grouped into several groups characterized by the significant tags, it is believed that common tags annotated by the most objects inside the group can reflect the characteristics of user preference or document functionality. Moreover, it will be of benefit for solving the problem of low tag quality of individual user or document. Even when a new user or a new document is added into the system, the tags can be extended to the user by referring to the majority tagging behavior of users on documents.

Regarding to the previous problems, even if the tag is rich enough for the users and documents, the time consuming is still very high when a user wants to get the most appropriate document from a large document database, since the system has to calculate the similarity between users or documents one by one.

We propose the method that calculates the similarity between the target tag vector and the centroids of all clusters to determine the cluster with highest similarity, then calculate the similarity of the target tag vector with the document profiles inside the cluster to rank the whole documents. In such way the time consuming can be reduced. Since we have got the groups of user profile or document profile, how to choose the number of clusters is another problem. The traditional way is to assign the initial clustering number manually. In this paper, we use the modularity metric [13] to evaluate the optimal number for the clusters.

Based on the problems mentioned above, this paper proposes an approach for group modeling by utilizing a clustering algorithm. The group modeling aims at assigning the individual users or document profiles into different groups, which correspond to various user preferences or content relatedness from the large amount of data for tagging.

User Group Profile and *Document Group Profile* can be generated from individual user profiles and document profiles; both of them are expressed by the tags. Group profiling is not constructed based on stereotypes but based on the results of clustering algorithms from transactional data. It can identify the objects inside the community with similar tags, and collect the data for the similar objects. It can expand the tag set for the individual object inside the community which is helpful for the poor tag quality and quantity. Furthermore, for making tag-based recommendation, it will significantly reduce the time consuming in calculating the similarity between the user and document groups.

The main contributions in this paper are:

1. A group modeling method by utilizing the clustering algorithm.
2. The most appropriate number of clusters to generate the User Group Profile and the Document Profile by using the modularity metric.
3. Reduction in time needed for computation for organizing the documents comparing to the other methods.

The rest of the paper is organized as follows: Section 2 presents the related work in the field of clustering and profiling. In section 3, we describe the preliminaries for the data model. Section 4 discusses the details of user profile and document profile with the introduced mathematical models and how to get the group profile by utilizing the spectral clustering algorithm. The experiment is designed in terms of datasets and evaluation measures in section 5, and experimental results and comparisons are presented in this section as well. We conclude the paper and discuss possible future research directions in section 6.

2 Related Work

The folksonomy in [3] has been defined as a data structure that evolves over time when people annotate resources with freely chosen words. It is user-contributed data aggregated by collaborative tagging systems. In such systems, users are allowed to choose terms freely to describe their favorite web resources. A folksonomy is generally considered to consist of at least three sets of elements, namely users, tags and resources. Although there can be different kinds of resources.

The prerequisite of personalization is to acquire user profile that describes user's interests, preferences and background knowledge about specified domains. Methods are used for modeling user profiles include logic-based representation and inference, Bayesian models, feature-based filtering, Clique-based filtering, and neural networks. However, such user profile is still for individual persons. Our approach is to cluster the similar users in the same communities. [5] proposes to create user profiles from the data available in such folksonomy systems by letting user specify the most relevant objects in the system. Instead of using the objects directly to represent the user profiles, they use the tags associated with the specified objects to build the user profiles.

[20] presents analysis on the personal data in folksonomies, and investigates how accuracy rate user profiles can be generated from this data. They propose an algorithm to generate user profiles which can accurately represent the multiple interests.

F. Durao and P. Dolog in [6; 8; 7] present a tag-based recommender system which suggests similar Web pages based on the similarity of their tags from a Web 2.0 tagging application. They also propose an approach to extend the basic similarity calculus with external factors such as tag popularity, tag representativeness and the affinity between user and tag.

K. R. Bayyapu and P. Dolog in [2] tries to solve the problems of sparse data and low quality of tags from related domains. They suggest using tag neighbors for tag expression expansion. However the tag neighbors are based on the content of documents. We propose another approach to extend the tag set by the group profiling.

[19] uses a framework of User-Profile Modeling based on Transactional data for modeling user group profiles based on the transactional data which can incorporate external information, either by means of an internal knowledge base or on dynamic data supplied by a specific information extraction system. Such user group profiles consist of three types: basic information attributes, synthetic attributes and probability distribution attributes. User profiles are constructed by clustering user transaction data and integrating cluster attributes with domain information extracted from application systems and other external data sources. And Teevan et al. apply group profiles to personalize search by an algorithm to "groupiz" (versus "personalize") in result ranking on group-relevant queries [18] , Abel et al. [1] shows that the quality of search result ranking in folksonomy systems can be significantly improved by introducing and exploiting the grouping of resources and Mei and Church show that group profiles facilitate Web search [11].

Clustering can divide the large amount of data into several groups. Clustering algorithms, specially designed for transactional data, can efficiently partition historic user transactions into clusters [9][15]. Each cluster is a set of transactions representing the interests of a particular user group. It is the assignment of a set of observations into subsets so that observations in the same cluster are similar in some sense. We want to use the clustering for unsupervised learning in the group profiling.

3 Preliminaries for Folksonomy Data Model

The user profile can be used to store the description of the user's characteristics. Such information can be exploited in social tagging systems for taking the persons' characteristics and preferences into account . For example, the social tagging systems usually ask the users to choose their own words as tags to describe the favorite web resource. So the user profiles can justify the benefit and interest for various users.

The document profile is represented by the metadata generated by the community of users tagging the documents. It is the process that refers to the construction of a profile for a specific via the extraction from a set of tagging data.

When users want to annotate web resource for better organization and use the relevant information to their needs later, they will tag such information with free-text keywords. The tags, which are given by the users, reflect the navigational preference and interest of them. On the other hand, with the increase of documents number that the user visited and annotated, each user has his own tag set which characterizes the interest or preference. Likewise, each tagged document also has its own tag set which expresses the content relatedness and subject of the document. In the context of social tagging systems, the user profiles and document profiles thus are expected to be represented by the representative tags. Therefore the process of user and document modeling is to capture the significant tags from a large volume of tagging data in a social collaborative environment.

There are a number of studies on user and document profiling (see for example [20; 19]). Amongst them, the basic idea of such approaches is originated from the introduction of a specific mathematical modeling of folksonomy. The folksonomy is a three-dimensional data model of social tagging behaviors of users. In social tagging systems, both the user profiles and document profiles are formulated starting from the folksonomy model. In the following section, in order to well reveal the mutual relationships between these three-fold entities, i.e. user, item and tag, we firstly briefly discuss the data model used in the following group profiling processes.

A folksonomy F according to [10] is a tuple $F = (U, T, D, A)$, where U is a set of users, T is a set of tags, D is a set of Web documents, and $A \subseteq U \times T \times D$ is a set of annotations. The relationship is shown as Fig1.

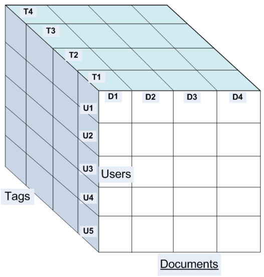

Figure 1: Relationship of users, tags, resources in folksonomy

We can construct the folksonomy data model from the tagging data by such following steps: collecting the data of users, tags and resources from the explicit information and implicit information. And then represent them in the three-dimensional vector space. Based on this we can define the documents' data as the X coordinate, the users' data as the Y coordinate, and tags' data as the Z coordinate. The relationship in folksonomy is $R_{tagging} = U \times T \times D$, $R_{tagging} \in A$, where $U = \{U_1, U_2, ..., U_m\}$ is the set of users and $T = \{T_1, T_2, ..., T_k\}$ is the set of tags,$D = \{D_1, D_2, ..., D_n\}$ is the set of documents. Shown in Fig1, for each point in the three-dimensional vector space, it can be defined as user $u \in U$ has tagged document $d \in D$ with tag $t \in T$.

Upon the folksonomy data model, we can derive the user and document profile by utilizing the relationship among the users, tags and documents in the tagging procedures, which will be discussed in the following section.

4 User Group Profiling and Document Group Profiling by Clustering

As mentioned in the introduction section, the poor quality and quantity of tag data would be a problem. Meanwhile, the time complexity is also a big concern when calculating the recommendation rank for the objects based on the large amounts of data. In the following parts, this paper will focus on solving such problems.

4.1 User Profile and Document Profile

In the social tagging systems, we can get the user profile and document profile by utilizing and analyzing the relationships among the users, tags and documents modeled in folksonomy.

First of all, we discuss the user profiling. For a given user, if we want to study his interests, only the tags, associated with documents, need to be concentrated on. In the folksonomy data model, we can use a user vector assgined with a unique id, for example, the user $U_i \in U, i$= 1,..., M.

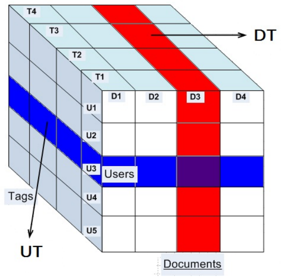

Figure 2: Matrix UT and Matrix DT in folksonomy

As shown in Fig2, a two-dimensional matrix UT_i is extracted from the relationship between the documents and tags for a particular user. In UT_i , each column is corresponding to the documents $D_n \in D, n$= 1,..., N that used by user U_i , and each row is corresponding to the tags $T_k \in T, k = 1, ..., K$.

$$UT_i = \begin{bmatrix} u_{11}, & u_{12}, & \cdots, & u_{1n} \\ u_{21}, & u_{22}, & \cdots, & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{k1}, & u_{k2}, & \cdots, & u_{kn} \end{bmatrix}, u_{kn} \in \{0, 1\}$$

Here u_{kn} means that if there exists an association between tag T_k and document D_n, annotated by user U_i, the u_{kn} sets to 1, otherwise it is 0.

By accumulating the row of matrix UT_i, the frequency of tag is defined as $t_{ik} = \sum_{n=1}^{N} u_{kn}$ which reveals the user's preference and interest. Then we can obtain the full set of the pairs of tags and their frequency weights. So the profile of userU_i in the form of tag set can be defined as UP_i =

$\{(T_1, t_{i1}), (T_2, t_{i2}) \cdots (T_k, t_{ik})\}, k = 1, \cdots, K$, where $t_{ik} = \sum_{n=1}^{N} u_{kn}$, $T_k \in T, k = 1, \cdots, K$.

Similarly, given a document D_n, we can obtain another two-dimensional matrix DT_i, where each column denotes the user U_i and row is the related tags that the userU_i is used to annotate the documentD_n. The size of matrix DT_i is M users by K tags.

$$DT_i = \begin{bmatrix} v_{11}, & v_{12}, & \cdots, & v_{1m} \\ v_{21}, & v_{22}, & \cdots, & v_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ v_{k1}, & v_{k2}, & \cdots, & v_{km} \end{bmatrix}, v_{km} \in \{0, 1\}$$

The elementv_{km} in DT_i is corresponding to the user U_m and tag T_k. The value of v_{km} is defined that, if there exists an annotation between tag T_k and user U_m, that means T_k is associated with the U_m, the v_{km} sets to 1, otherwise it is 0.

By accumulating the row of matrix DT_i, the frequency of tag is defined as$t_{ik} = \sum_{m=1}^{M} u_{km}$ Then we can obtain the full set of the pairs of tags and their frequency weights. So the profile of document D_n in the form of tag set can be defined as $DP_i = \{(T_1, t_{i1}), (T_2, t_{i2}) \cdots (T_m, t_{im})\}, m = 1, \cdots, M$, where $t_{ik} = \sum_{m=1}^{M} u_{km}$, $T_k \in T, k = 1, \cdots, K$.

From the steps mentioned above, the user profiles and document profiles are defined as a single user or document respectively rather than a group of users or documents. However, in social tagging systems, the group profiles of users or documents are more likely to reflect the common preference or relatedness of like-minded users or documents with similar functionality. In the following section, we will discuss the group profiling approach by using clustering.

4.2 Similarity Matrixes for the Users and Documents

The relationship among all of the users is to calculate the similarity. The similarity is quantity that reflects the strength of relationship between two objects. In the last part, each user profile can be represented by the pair of tags and frequencies. We utilized the cosine distance between users. Its value ranges from 0 to 1, the higher value of the similarity, the more similar the objects are. The similarity matrix $SM(U_i, U_j)$is given by,

$$SM(U_i, U_j) = \frac{UP_i \cdot UP}{|UP_i| \times |UP_j|}$$

The users' relationship can be represented in the form of bipartite graph model. Given a graph G = *(U,E)*, where U is a set of users as $U_i = \{U_1, U_2, \cdots, U_m\}$, and E is a set of edges which entry $SM\left(U_i, U_j\right)$ reflects the similarity between users, the similarity matrix $SM(u_i, u_j)$.

Similarly with the document profiles, we can define a graph with N users G = *(D,E)*, where D is a set of documents as $D_i = \{D_1, D_2, \cdots, D_n\}$, and E is a set of edges which entry $SM\left(D_i, D_j\right)$ reflects the similarity between documents. The similarity matrix $SM(D_i, D_j)$is given by:

$$SM\left(D_i, D_j\right) = \frac{DP_i \cdot DP_j}{|DP_i| \times |DP_j|}.$$

4.3 Group Profiling via Clustering Algorithm

To accomplish the group profiling, one of the approaches is to group the user profile and document profile into several groups based on the similarity so that the objects in the same groups can share tag set. The clusters of users or documents reveal the common user preference or relatedness of documents. It can benefit the user in the same group to share the similar interests or documents.

Clustering algorithm aims to assign a set of observations into subsets so that observations in the same cluster are similar in some sense. It is specially designed for transactional data and can efficiently partition historic user transactions into clusters [8; 14; 12]. Each cluster is a set of transactions representing the interests of a particular user group. So it can find the potential groups from the user profile and document profile.

The object of clustering is: similar objects have high similarity, the similarity is low between objects of different clusters. Clustering is the process to adjust the ranks of the similarity matrix, by a number of matrix blocks to meet the similarity value larger among the inside elements, while the similarity value is small between the clusters.

There are a lot of clustering algorithms such as k-means, fuzzy c-means, single linkage and so on. In clustering analysis, almost all approaches are based on the similarity between subjects to partition the data points. Various clustering approaches have different advantages and drawbacks. Among the traditional clustering algorithms, spectral clustering has the superior capability of effectively group data by leveraging the statistical property of similarity matrix of data.

In this paper, we will introduce the Spectral Clustering Algorithm. Spectral clustering refers to a class of techniques which rely on the eigenvalues of the adjacency similarity matrix; it can partition all of the points into disjoint clusters, the points that have high similarity will be classified under the same cluster. One cluster's points have low similarity with other clusters' points. The spectral clustering is based on the graph partition. We have explained how to get the similarity matrix from the graph in 4.2. It maps the original inherent relationships onto a new spectral space, on which the user or document profile is projected. After the projection, the whole user or document profiles are simultaneously partitioned into disjoint clusters with minimum cut optimization.

Compared to those algorithms, spectral clustering has many fundamental advantages. Results obtained by spectral clustering often outperform the traditional approaches, it is easy to implement and can be solved efficiently by standard linear algebra methods. Spectral clustering techniques make use of the spectrum of the similarity matrix of the data to perform dimensionality reduction for clustering in fewer dimensions.

The original formula for the spectral clustering is:

$$L = I - D^{-1/2} S D^{-1/2}$$

According to the spectral graph theory in [4], the k singular vectors of the reformed matrix $RM_{User} = D^{-1/2} SM_{User} D^{-1/2}$ present a best approximation to the projection of user-tag vectors on the new spectral space.

And the $RM_{Document} = D^{-1/2} SM_{Document} D^{-1/2}$ presents the document-tag vector on the new spectral space.

The D_u and D_d is the diagonal matrix of user similarity matrix and documents similarity matrix, which are defined as:

$$D_u(i,i) = \sum_{j=1}^{N} SM\left(U_i, U_j\right), i = 1, \cdots, M$$

$$D_d(i,i) = \sum_{j=1}^{N} SM\left(D_i, D_j\right), i = 1, \cdots, N$$

Let's take the documents set for example.

In this case we assume that the first K singular eigenvectors represent the best approximation of original profile space. Let L_s the $m \times k$ matrix of the k singular vectors of $RM_{Document}$. As our aim is to conduct a clustering on the document profile attributes, we create a new $m \times k$ matrix RV to reflect the projection of the row and column vectors on the new spectral space in lower dimension as: $RV = [D_d^{-1/2} L_s]$.

The clustering results in the group profiling. The full steps of group profiling via clustering algorithm is summarized in the below Algorithm.

Input: The N document profile collection $DP = \{DP_i \,|i = 1, 2 \cdots N\}$, $DP_j = \{(T_1, t_1), (T_2, t_2) \cdots (T_K, t_K)\}$.

Output: A set of k clusters $DGP = \{DGP_i \,|i = 1, 2 \cdots k\}$such that the cut of k-partitioning of the bipartite graph is minimized.

1. Construct the usage similarity matrix $SM_{Document}$ from the document profile, whose element is determined by the distribution of tags of all users;
2. Calculate the diagonal matrixes D_d;
3. Form a new matrix $RM_{Document} = D^{-1/2} SM_{Document} D^{-1/2}$;
4. Perform SVD (Singular value decomposition) operation on $RM_{Document}$, and obtain k singular vectors L_s create a new projection matrix RV;
5. Execute a clustering algorithm on RV and return clusters of documents:$DGP = \{DGP_i \,|i = 1, 2 \cdots k\}$.

From the above steps, the N documents are divided into t clusters, the document group profile for each cluster is: centerline$DGP_i = \{UP_{i1}, UP_{i2}, \cdots UP_{it}\} = \{(T_1, w_{i1}), (T_2, w_{i2}) \cdots (T_i, w_{it})\}$

Where $T_s \in T, s = 1, \cdots, t$ and $(w_{i1}, w_{i2}, \cdots w_{it})$ is the centroid of the document cluster DGP_i.

Meanwhile, the selection of cluster number k is another concern in the context of clustering, which is commonly encountered. The selection of k value has a straight impact on the performance of clustering: the bigger number of k results in the over-separation of users and documents, while the smaller number of it prevents the data from being sufficiently partitioned. Thus it is necessary before performing the clustering to select an appropriate value of k to achieve a better clustering performance. In the experimental part, we will investigate the study of k selection.

In similar way, the user group profile of k users can be generated in the same way: $UGP_i = \{DP_{i1}, DP_{i2}, \cdots DP_{ij}\} = \{(T_1, x_{i1}), (T_2, x_{i2}) \cdots (T_K, x_{iK})\}$

Where $T_s \in T, s = 1, \cdots, t$ and $(x_{i1}, x_{i2}, \cdots x_{it})$ is the centroid of the user cluster UGP_i.

5 Experimental evaluations

In order to evaluate the proposed group profiling, we performed experiments on the "MovieLens" dataset. Our experiments focus on the cluster number selection; demonstration of the group profiles; and the computational cost reduction.

5.1 Dataset and Modularity Metric

As for experiment dataset, we utilize the part of the "MovieLens" data, which contains tags provided by users on movies. It includes 521 users, 1399 documents and 1956 tags. The differences of results are shown when choosing different numbers of clusters in order to get the optimal number of clusters. The data is based on the average result of executing the same experiment ten times over the same dataset.

The modularity metric is one of the standard quantitative measures for the evaluation of "goodness" of the clusters. The modularity of a particular division of a network is calculated based on the differences between the actual number of edges within a community in the division and the expected number of such edges if they were placed randomly. Good divisions, which have high values of the modularity, are those dense connections between the nodes within modules but sparse connections between different modules. It will help to evaluate the quality of the cluster; i.e. the similarity of each cluster.

After clustering, we can get several clusters. Consider a particular division of a network into k communities. We can define a $k \times k$ symmetric matrix SM whose element sm_{ij} is the fraction of all edges in the network that link vertices in community p to vertices in community q. The similarity of smC_{pq} between the two clusters C_p and C_q,is defined as,[13]

$$smC_{pq} = \frac{\sum\limits_{c_p \in C_p} \sum\limits_{c_p \in C_q} c_{pq}}{\sum\limits_{c_p \in C} \sum\limits_{c_q \in C} c_{pq}}, p, q = 1, 2 \cdots m$$

where c_{pq} is the element in the similarity matrix SM. When p=q, the smC_{pq} is the similarity between the elements inside the clusters, while $p \neq q$, the smC_{pq} is the similarity between the cluster C_p and the cluster C_q. So the condition of a high quality cluster is $\arg\max(\sum\limits_p smC_{pp})$ and $\arg\min(\sum\limits_{p,q} smc_{pq}), p \neq q, p, q = 1, 2, \cdots m$.

Summing over all pairs of vertices in the same group, the modularity, denoted Q, is given by:

$$Q = \sum_{p=1}^{m} [smc_{pp} - (\sum_{q=1}^{m} smc_{pq})^2] = TrSM - \left\|SM^2\right\|$$

where the m is the amount of clusters. The trace of this matrix $TrSM = \sum\limits_{p=1}^{m} smC_{pp}$ gives the fraction of edges in the network that connect vertices in the same community Clearly a good division into communities should have a high value of this trace. If we place all vertices in a single community, the value of would get the maximal value of 1 because there's no information about community structure at all.

This quantity measures the fraction of the edges in the network that connects vertices of the same type minus the expected value of the same quantity in a network with the same community divisions. Utilizing the value Q to evaluate the clusters [13] is a common method: the values approaching Q=1, which is the maximum value, indicate the networks with strong community structure. In practice, the values of such networks typically range from 0 to 1. The higher value of Q, the better quality for the cluster corresponding to a predefined cluster number k. So examining the Q value allows us get the optimal number of clusters.

5.2 Experimental Results

Optimal Cluster Number Selection

Here we compare the result of Q values by using Spectral Clustering, Single Linkage Clustering and Random Clustering.

Of the entire 521 user profiles constructed, we employ various clustering algorithms to build up the group user profiles. We generate the cluster from 2 to 260 and utilize the modularity method to evaluate the results. We close the number until 260 because it is half of the total amount. When the number of cluster is higher than 260, the average number of members in each cluster is lower than 2, which will not provide reasonable clustering information. The results are shown in Fig3. that the value of Q for Spectral Clustering Algorithm is consistently higher than the other two algorithms. When the number is 25 the Q gets the maximal as 0.381. With the growth of the number of clusters, the value of Q is gradually decreasing to 0.19. It is concluded that, for this dataset, 25 clusters is the best choice.

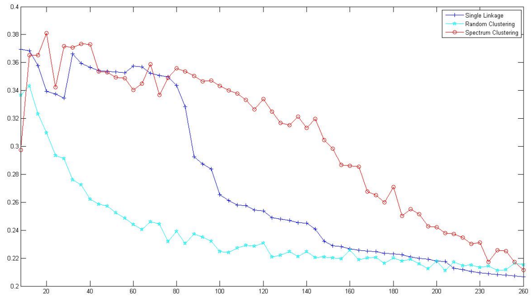

Figure 3: Comparison of the three algorithms on 521 users

Of the entire 1399 document profile models constructed, we generate the cluster from 5 to 700 and utilize the modularity method to evaluate the results. Similarly as the user profile, we close the number until 700. As shown in Fig4, when the number is 20 the Q gets to the maximum at 0.277. With the growth of number of clusters, the value of Q is decreasing to 0.026. We then found that, for this dataset, 20 clusters is the best choice.

Demonstration of the Group Profiles

It is shown that the 20 clusters is the optimal number for the 1399 documents, we take 2 clusters of them for analysis. One of the cluster contents 51 documents, with 239 tags. The main tag in it is about the "classic", "based on a book", "black and white", "National Film Registry", "breakthroughs", "Disney" and so on. The movies in such cluster seem related to the movies about life.

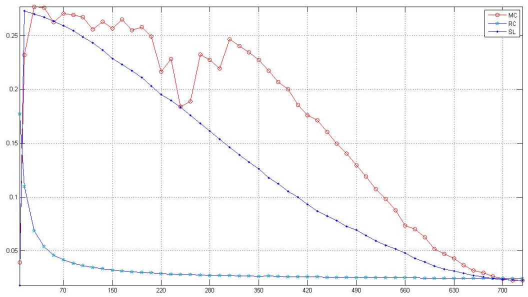

Figure 4: Comparison of the three algorithms on 1399 documents

And another cluster has 79 documents with 397 tags, the dominant tags are "action", "organized crime", "guns", "hysterical", "USA film registry", "afternoon section", "Oscar (Best Actor)", "Oscar (Best Cinematography)", "Oscar (Best Director)", "Oscar (Best Picture)" and so on. Such movies tend to the Oscar movies with some breathtaking content.

Comparison between the Time Consuming

When the user or document profiles are used in tagging systems for further applications, similarity calculation is a major operation involved. An advantage of group profiling is the possibility of reducing the computational complexity. For example, cosine similarity is often executed to determine the ranking of candidates. The traditional way is to calculate the similarity between its tags and each document's tags. In such way the time complexity is the $O(n)$. It will cost much time when the system has large dataset.

After clustering for all of the documents, each cluster will have its own centroid as the representation of the group profile, which means the "center point" in the cluster. Centroid can be generated by the average frequency of tags inside the cluster. The similarity between the centroid and the items inside the same cluster should be the highest; the similarity between the centroid and the items inside the other clusters should be the lowest. So the centroid is the representative of the cluster. If N documents have clustered into m communities, the process of similarity calculation is divided into two steps: firstly, calculate the similarity between the target tag vector and the centroids of all clusters to determine the cluster with highest similarity score; secondly, calculate the similarity of the target tag vector with the document profiles inside the cluster to rank the whole documents. Since the number of centroids, m, is equal to the number of communities, which is highly lower than the number of documents, N, the time consuming of calculating similarity is dramatically reduced from $O(N)$ to $O(m+\frac{N}{m})$.

In our experiment, the time consuming computing the similarity for all 1399 documents respectively is 152.83 seconds, however, it just needs 0.045 seconds by our proposed approach to get the final ranking of documents.

6 Conclusion and future work

In this paper, we discuss an algorithm for user group profile and document group profile in social tagging systems. Utilizing the clustering algorithm, group profiling can be processed by the user profile and document profiles. We implement experiments on real tagging dataset to validate the proposed approach, investigate the modularity method to compare the optimal number of clusters, and demonstrate the content of clusters. At last, we compare the time consuming for the similarity calculation involved in real applications. It is shown that the group profiling can be dealt with the tasks outlined in the paper effectively.

For the future work, we intend to conduct research on the optimization for the algorithm, and explore the deployment of group profiling in tag-based recommender system. We will investigate clustering for tags which we believe should help in tag recommendation and representing user interests.

References

[1] F. Abel, N. Henze, D. Krause, and M. Kriesell. On the effect of group structures on ranking strategies in folksonomies. *Weaving Services and People on the World Wide Web*, pages 275–300, 2009.

[2] K. R. Bayyapu and P. Dolog. Tag and Neighbour Based Recommender System for Medical Events. In *Proceedings of MEDEX 2010: The First International Workshop on Web Science and Information Exchange in the Medical Web colocated with WWW 2010 conference*, 2010.

[3] J. Davies, D. Fensel, C. Bussler, and R. Studer. The Semantic Web: Research and Applications. In *Proceedings of the First European Semantic Web Symposium*, 2004.

[4] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274. ACM, 2001.

[5] J. Diederich and T. Iofciu. Finding communities of practice from user profiles based on folksonomies. In *Proceedings of the 1st International Workshop on Building Technology Enhanced Learning solutions for Communities of Practice*. Citeseer, 2006.

[6] F. Durao and P. Dolog. A personalized tag-based recommendation in social web systems. *Adaptation and Personalization for Web 2.0*, page 40, 2009.

[7] F. Durao and P. Dolog. Social and Behavioral Aspects of a Tag-Based Recommender System. In *Ninth International Conference on Intelligent Systems Design and Applications, 2009. ISDA'09*, pages 294–299, 2009.

[8] F. Durao and P. Dolog. Extending a hybrid tag-based recommender system with personalization. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 1723–1727. ACM, 2010.

[9] S. Guha, R. Rastogi, and K. Shim. Rock: A robust clustering algorithm for categorical attributes* 1. *Information Systems*, 25(5):345–366, 2000.

[10] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Folkrank : A ranking algorithm for folksonomies. In K.-D. Althoff and M. Schaaf, editors, *LWA*, volume 1/2006 of *Hildesheimer Informatik-Berichte*, pages 111–114. University of Hildesheim, Institute of Computer Science, 2006.

[11] Q. Mei and K. Church. Entropy of search logs: how hard is search? with personalization? with backoff? In *Proceedings of the international conference on Web search and web data mining*, pages 45–54. ACM, 2008.

[12] A. Nanopoulos, H. H. Gabriel, and M. Spiliopoulou. Spectral Clustering in Social-Tagging Systems. *Web Information Systems Engineering-WISE 2009*, pages 87–100, 2009.

[13] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):26113, 2004.

[14] M. G. Noll and C. Meinel. Web search personalization via social bookmarking and tagging. In *Proceedings of the 6th international semantic web conference and 2nd Asian conference on Asian semantic web*, pages 367–380. Springer-Verlag, 2007.

[15] S. Rendle, L. B. Marinho, A. Nanopoulos, and L. Schmidt-Thieme. Learning optimal ranking with tensor factorization for tag recommendation. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 727–736. ACM, 2009.

[16] J. Stoyanovich, S. Amer-Yahia, C. Marlow, and C. Yu. Leveraging tagging to model user interests in del. icio. us. *AAAI SIP*, 2008.

[17] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos. A unified framework for providing recommendations in social tagging systems based on ternary semantic analysis. *IEEE Transactions on Knowledge and Data Engineering*, 2009.

[18] J. Teevan, M. R. Morris, and S. Bush. Discovering and using groups to improve personalized search. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 15–24. ACM, 2009.

[19] Y. Yang and N. Marques. User group profile modeling based on user transactional data for personalized systems. *Progress in Artificial Intelligence*, pages 337–347, 2005.

[20] C. M. A. Yeung, N. Gibbins, and N. Shadbolt. A study of user profile generation from folksonomies. In *Social Web and Knowledge Management, Social Web 2008 Workshop at WWW2008*. Citeseer, 2008.