Personalized Researcher Profiling for Recommendation Using Temporal Pattern

1st Kaichen Tang
School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University
Shanghai, China
tangkc@sjtu.edu.cn

Abstract— **Recommender systems have become increasingly popular in recent years, and are utilized in a variety of areas including movies, music, news and so on. Moreover, recommender system has been adapted to the field of academics in the form of paper recommendation and advisee recommendation, typically based on Collaborative Filtering (CF) or Content-Based filtering (CB). As a result, accurate and diversified user profile would benefit the performance of the recommender system. In this paper, we focus on two main aspects of researcher profiling: a researcher's cooperation pattern, and the variation of his/her research area. We propose a user profiling method based on the temporal pattern, which depends on Hidden Markov Model (HMM) to excavate the transition of hidden states, and use clustering or dimensional reduction to classify the behavior of a researcher. We apply our method on the citation network dataset[*] from AMiner.**

Index Terms—**Research Profiling, Hidden Markov Model, Markov Cluster Algorithm, Spectral Clustering, Latent Dirichlet Allocation, Recommender System.**

I. Introduction

A recommender system or a recommendation system is a subclass of an information filtering system that seeks to predict the "rating" or "preference" a user would give to an item [1].

Recommender systems have become increasingly popular in recent years, and are utilized in a variety of areas including movies, music, news, books, research articles, search queries, social tags, and products in general. There are also a number of websites providing or planning to provide recommender systems for academic researches, like Acemap[†].

Recommender systems typically produce a list of recommendations in one of two ways -- through collaborative filtering or through content-based filtering (also known as the personality-based approach) [2]. Collaborative filtering methods are based on collecting and analyzing a large amount of information on users behaviors, activities or preferences and predicting what users will like based on their similarity to other users [3]. A key advantage of the collaborative filtering approach is that it does not rely on machine analyzable content and therefore it is capable of accurately recommending complex items such as movies without requiring a "understanding" of the item itself [4]. Content-based filtering methods are based on a description of the item and a profile of the users preferences [5]. In a content-based recommender system, keywords are used to describe the items and a user profile is built to indicate the type of item this user likes. In other words, these algorithms try to recommend items that are similar to those that a user liked in the past (or is examining in the present). In particular, various candidate items are compared with items previously rated by the user and the best-matching items are recommended. This approach is embedded in information retrieval and information filtering research.

In the consideration of the performance and accuracy of recommender system, content-based filtering methods and the hybrid recommender systems of CF and CB has been widely used than any other methods [6]. It is essential that the performance and accuracy of these kinds of recommender system depend largely on the description of the item and a profile of the users preferences as we have introduced in the last paragraph. As a result, it is important to build a personalized researcher profiling system that can reveal both the explicit and the implicit preference of users [7] for the recommender system aims for academic purpose.

In the past, most researchers focused on the method to extract explicit information for the user using profile. For example, L. Yao et al. propose a unified approach to perform the task using Conditional Random Fields (CRF) as a method to extract the researcher profile [8]. Their work is based on data collected from the researches' homepage. Using their method, they profile the researcher's name, photo, address, phone and so on from a variety of templates on their homepages. However, they also indicate that only about 40.60% of the researchers have at least one homepage or a Web page that introduces them. In another word, more than a half of researchers' profile cannot be established in this way, because they do not have a centralized display, like a homepage, of their information, or these kind of homepages cannot be found through web crawler or search engine. In addition, many of the researchers may feel troublesome to update their online profile regularly or they prefer other platform to express their progresses. Consequently, their information, including their variation of focus points and their requirement for new cooperative partner, may not be presented on their homepage in time. In short, this deficiency of existing profiling methods stress the importance of using more concrete and up-to-date information sources, like research publication in an academic

[*] http://resource.aminer.org/citation

[†] http://acemap.sjtu.edu.cn/

field. And implicit profiling is predicted to the future trend [7].

We propose using HMM to reveal the implicit behavior pattern of a researcher based on his/her publication sequence. Specifically, to extract a researcher's cooperation pattern, we arrange his/her coauthors in an array as the training data of HMM. After getting the sequence of hidden states, we use Markov Cluster Algorithm (MCL) to merge the similar states and use the distribution of different states of diverse researchers to classify them into a certain number of clusters. What's more, to extract the transition of focus points of a research, we use Latent Dirichlet allocation (LDA) to find the topic distribution of all papers, and use HMM to observe the latent model of the change pattern. We utilize the citation dataset from AMiner to realize our idea.

II. Related Work

A. Existing Researcher Profiling System

L. Yao et al. propose a unified approach to perform the task using Conditional Random Fields (CRF) as a method to extract the researcher profile [8]. The paper shows that with the introduction of a set of tags, most of the annotation tasks can be performed within this approach. We have defined the problem as a task consisting of 19 sub-tasks, that is, photo, position, affiliation and so on. They also adapt face recognition and natural language processing to the data they gathered to establish a well-rounded researcher profile.

M. Lee et al. [10] examines a method of generating comprehensive profiling information for a researcher analysis service. They introduce researcher performance index models for researcher analysis service. The models can that measure qualitative and quantitative performance, researcher influence, and growth potential, which is necessary to analyze the skills of a researcher from multiple perspectives. The quantitative performance index can be evaluated based on a researchers published papers. The Influence index measures the social impact of researchers according to their academic work. The growth potential index determines the speed at which a researcher improves research performance.

As we can see from above works, popular researcher profiling mainly focus, or, in another word, restricted to extract personal information from data available on researcher's website or using statistics method to gather the influence of the researcher. However, the actual relation network of researchers can be very complex and there are plenty of underlying information to be discovered. Taking advantage of uprising machine learning methods, we can use more advanced approaches to reveal the hidden behavioral pattern of a researcher to enrich the profile, thus benefit the accuracy and performance of recommender system.

B. Current State and Future Trend of User Profiling

A. Shepitsen et al. present a personalization algorithm for recommendation in folksonomies which relies on hierarchical tag clusters [9]. Their basic recommendation framework does not depend on the clustering method, but they use a context-dependent variant of hierarchical agglomerative clustering which takes into account the users current navigation context in cluster selection. Furthermore, their work demonstrates more utility for recommendation in multi-topic folksonomies than in single-topic folksonomies.

S. Kanoje et al. [11] and D. Godoy et al. [12] conduct surveys of existing user profiling systems. Nevertheless, to the best of our knowledge, all of these methods are a kind of explicit user profiling as it is described in [7]. S. Kanoje et al. categorize user profiling into three sets: explicit user profiling, implicit user profiling, and hybrid user profiling. Implicit user profiling is also known as as behavioral profiling or adaptive profiling. Instead of concentrating on the current information we have about the user this approach, implicit user profiling relies more on actions performed by the user or what we have known about user in the future. Although there are some filtering techniques for implicit user profiling in other areas, no research focus on adapting these ideas into the field of researcher profiling.

C. Application of HMM on Social Networks Analysis

K. Zhang et al. analyze the source of the user behavior audit analysis data [13]. In their paper, cloud environment security issues, cloud storage environment are used as the research object, focusing on the audit mechanism of user behavior in cloud environment. According to the security features of cluster data in cloud storage environment, a user behavior auditing model is proposed, and the data acquisition and preprocessing of user behavior audit are studied. This paper In the process of audit data analysis and processing, a feature vector method is therefore proposed to extract valuable information from audit data. A user behavior modeling method based on Hidden Markov model is proposed in this paper. The user behavior model is utilized to identify the validity of the user's operation, and to ensure the security of the data of the cloud platform.

N. Mohammadifard addresses the problem of finding the influence of advertisements on a user's purchase behavior, by using machine learning methods to analyze purchase data obtained from real online retail systems [14]. His paper is based on a hypothesis that different ads have distinct influences, but also the same ad. can make the user behave differently if she is in different inner states. He replaced the traditional observation model of a Hidden Markov Model with Logistic Regression, which allows us to define an observation model depending not only on the HMM state, but also on external events such as advertising campaigns.

H. Kawazu et al. propose a new analytical method to classify web user behavior based on such latent states of users as intention, interest, or motivation [15]. First, they put the click-stream data of many users into a Hidden Markov Model in which the number of hidden states is large enough to construct a state transition network. Second, they divide each piece of click-stream data into sessions, which they classify using network movement as feature values. They observe the following hidden states that represent the variable latent states of users, such as enthusiasm for the main contents of the service, playing basic content, and daily routines that are well observed by visiting the service.

H. Li et al. discover that reviewers posting rates are bimodal and the transitions between different states can be utilized to differentiate spammers from genuine reviewers [16]. Guided by these findings, they propose a two-mode Labeled Hidden Markov Model to detect spammers. Experimental results show that their model significantly outperforms supervised learning using linguistic and behavioral features in identifying spammers.

These works provide perfect examples of using HMM to extract hidden behavioral pattern of users in social networks like online shopping website or social networking site, which inspired us to observe the hidden state transition network of academic publication data.

III. Citation Network Dataset

The citation data is extracted from DBLP, ACM, MAG (Microsoft Academic Graph), and other sources [17]. The first version contains 629,814 papers and 632,752 citations. Each paper is associated with abstract, authors, year, venue, and title. The dataset has been used for clustering with network and side information, studying influence in the citation network, finding the most influential papers, topic modeling analysis, etc. This dataset is established by J. Tang et al.. Concerning the limited resource, we choice the Citation-network V1 with 629,814 papers and more than 632,752 citation relationships (2010-05-15) to be our dataset. One advantage of this dataset is most papers in it contains abstract, which can be used for topic modeling. We use abstracts and titles to model extract the topic distribution, further modeling a researcher' variation on his/her focus points.

IV. Modeling Researcher Cooperation Pattern

To model a researcher's cooperation pattern, we first use HMM to transform his/her publication sequence into a series of hidden states. Regarding the transition matrix derived from HMM as a network of hidden states, we can merge some states into communities by MCL. Then we use spectral clustering to classify researchers according to their possibility of passing through different communities. Finally, we can sketch the characteristic of each cluster of researchers by observation.

A. Hidden Markov Model

A Hidden Markov Model is a general statistical modeling technique for sequential pattern problems, such as speech recognition. In HMM, the generative model of the data is assumed to form a Markov process with hidden states that are discrete variables [18]. Each hidden state has a probability distribution of observed variables, which are output depending on the present hidden state at each step.

Let z_n be the hidden state at step n, and the conditional probability is expressed in the following equation because of the assumption of a Markov process:

$$p(z_n|z_1, z_2, \cdots, z_{n-1}) = p(z_n|z_{n-1}). \tag{1}$$

Let x_n be the observed variables. The joint probability of all of the hidden states and the observed variables from steps 1 to N is expressed in the following equation:

$$\begin{aligned} Q(X, Z) &= p(x_1, x_2, \cdots, x_N, z_1, z_2, \cdots, z_N) \\ &= p(z_1)\left[\prod_{n=1}^{N} p(z_n|z_{n-1})\right]\prod_{n=1}^{N} p(x_n|z_n). \end{aligned} \tag{2}$$

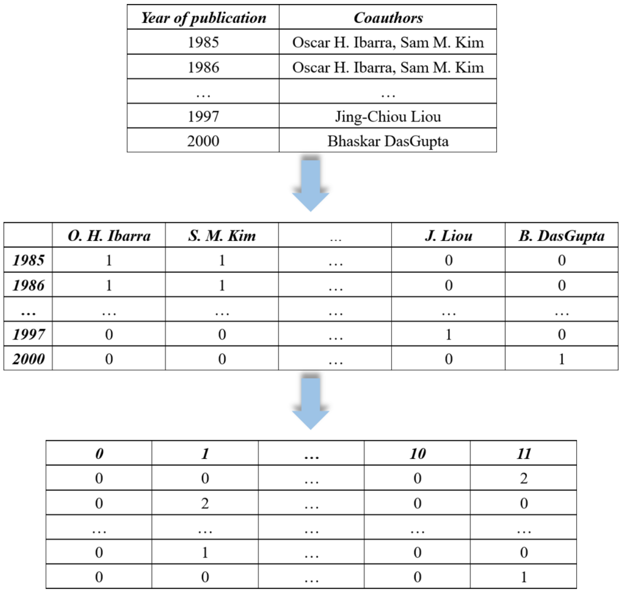
The joint probability has three parameters: initial state distribution π, state transition distribution A, and the probability distribution of the observed variables in each state Σ, which are expressed in the following equations:

$$\pi_i = p(z_i) \tag{3}$$

$$A_{ij} = p(z_j|z_i) \tag{4}$$

$$\Sigma_{ik} = p(x_k|z_i), \tag{5}$$

where $(i, j = 1, 2, \cdots, L.\ k = 1, 2, \cdots, K)$.



Year of publication	*Coauthors*
1985	Oscar H. Ibarra, Sam M. Kim
1986	Oscar H. Ibarra, Sam M. Kim
...	...
1997	Jing-Chiou Liou
2000	Bhaskar DasGupta

	O. H. Ibarra	*S. M. Kim*	...	*J. Liou*	*B. DasGupta*
1985	1	1	...	0	0
1986	1	1	...	0	0
...	...	...	...	...	...
1997	0	0	...	1	0
2000	0	0	...	0	1

0	*1*	...	*10*	*11*
0	0	...	0	2
0	2	...	0	0
...	...	...	...	...
0	1	...	0	0
0	0	...	0	1

Fig. 1. The procedure of constructing the matrix $\mathbf{C}$ of each researcher: firstly, we collect a list of the researcher's coauthor in chronological order; secondly, we do the binaryzation of the list we collected in the first step; finally, we convert the sparse binary matrix into a $l \times 12$ matrix $\mathbf{C}$ as we described above. This example is based on the coauthorship of Michael A. Palis.

In this paper, the training data of HMM is based on researchers' coauthors in time sequence. Firstly, we gather all the coauthors' name of each research in order of publication time. Then we use these information to build a $l \times 12$ matrix $\mathbf{C}$ of each researcher who has $l \geq 10$ publications in the dataset, where $\mathbf{C}_{i,j} (0 \leq j \leq 9)$ stands for the number of the coauthors in the researcher's i^{th} publication that has cooperated with the researcher j years ago. For example, if $\mathbf{C}_{5,2} = 3$, it means that in the researcher's fifth publication in the dataset, there are 3 coauthors that has worked with this researcher 2 years ago in another paper within the dataset. For $\mathbf{C}_{i,10}$ and $\mathbf{C}_{i,11}$, the former stands for the number coauthors who the researcher has worked with exactly 10 years ago or more than 10 years (only counts for about 0.27% in this dataset), and the latter stands for the number of coauthor who work with this researcher for the first time. We only record the data of those who have more than a certain number publications in the dataset to make sure the distribution of hidden states can represent the behavior of that researcher. In practice, we set this threshold to be 10.

After collecting a matrix from each author, we have a list of two-dimension matrices. Then we use this matrix list together as the training set for HMM, where each researcher's sequence is independent but share a same set of parameters including the transition matrix and the covariance matrix. An example of the procedure of constructing the matrix is shown in Fig. 1.

HMM state z_i is characterized by the probability distribution of observed variables Σ_i, as shown in the following equation:

$$\Sigma_i = (p(x_1|z_i), p(x_2|z_i), \cdots, p(x_K|z_i)). \tag{6}$$

Each probability distribution of the observed variables reflected the researcher's choice of coauthor at a certain point of time in a certain situation or a latent state of the researcher, e.g., a state of high interaction motivation with other researchers or participation in cross institute cooperation project. Each hidden state had a different distribution, and so various hidden states represent different latent states of researchers. We set the number of HMM states to 10, which is considered large enough. The line chart diagram of training score of HMM is shown in Fig. 2.

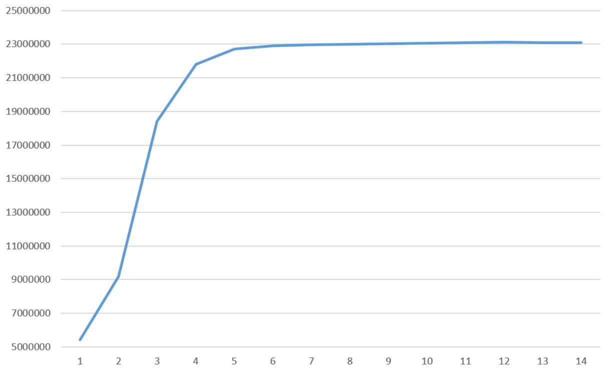


Fig. 2. Line chart diagram of training score of HMM after each iteration. As we can see from the figure, the model converged very fast due to the sparsity of input matrices.

After this step, we get a sequence of hidden states of each researcher who has more than 10 publications. These sequences represent latent cooperation behavior of researches.

B. Markov Cluster Algorithm

The state transition distribution of the HMM is a network whose nodes are the hidden states and whose directed links are the transition probabilities, as it is shown in Fig. 3. Each copartnership of a publication can be expressed in the networks corresponding movement. Network movement is the transitions of the researcher states, which represent his/her behavior. To acquire more abstract factors than the hidden states, we detected communities on this network. The nodes in the same community have links of high weight, which means that researchers tended to stay within one community. The communities represented the latent states of researchers more abstractly than the hidden states. We adapt Markov Cluster Algorithm (MCL) [19] to cluster hidden states into communities.

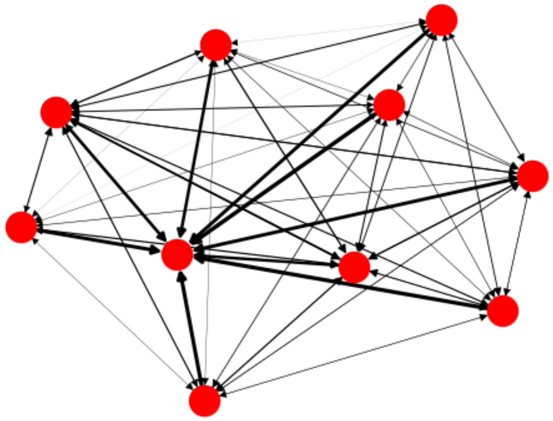


Fig. 3. The transition network of the 10 hidden states before clustering. The width and darkness of the edge is proportional to the possibility of corresponding movement.

The MCL algorithm is a fast and scalable unsupervised cluster algorithm for graphs based on simulation of stochastic flow in graphs. The algorithm simulates flow alternating two simple algebraic operations on matrices. Its formulation is simple and elegant. There are no high-level procedural instructions for assembling, joining, or splitting of groups-cluster structure is bootstrapped via a flow process that is inherently affected by any cluster structure present. To the most important, MCL is scalable and fast, because its worst-case complexity is $O(Nk^2)$, where N is the number of nodes of the input graph, and where k is a threshold for the number of resources allocated per node [20].

We use MCL as a method of weighted and directed network clustering. The MCL algorithm simulates random walks within a graph by an alternation of two operators called expansion and inflation. Expansion coincides with taking the power of a stochastic matrix using the normal matrix product (i.e. matrix squaring). Inflation corresponds with taking the Hadamard power of a matrix (taking powers entrywise), followed by a scaling step, such that the resulting matrix is stochastic again, that is the matrix elements (on each column) correspond to probability values.

With the help of MCL, we can merge 10 hidden states into 4 communities, with expand factor of 3, inflate factor of 2.2. With advanced abstraction of latent states sequence of researchers, following clustering will be accelerated and become more accurate.

C. Spectral Clustering

The sequences of movements on the state transition network represent user behaviors. Next, we cluster and label researcher's behavior based on these sequences. In this paper, we use spectral clustering [21].

Spectral clustering does a low-dimension embedding of the affinity matrix between samples, followed by a KMeans in the low dimensional space. In multivariate statistics and the clustering of data, spectral clustering techniques make use of the spectrum (eigenvalues) of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions [22].

The basic steps [23] of spectral clustering go as follows, when given a set of points $S = s_1, s_2, \cdots, s_n$ in $\mathbb{R}^l$ that we want to cluster into k subsets:

- Form the affinity matrix $A \in \mathbb{R}^{n\times n}$ defined by $A_{ij} = exp(\|s_i - s_j\|^2 / 2\sigma^2)$ if $i \neq j$, and $A_{ii} = 0$.
- Define D to be the diagonal matrix whose (i, i)-element is the sum of A's i^{th} row, and construct the matrix $L = D^{-1/2}AD^{-1/2}$.
- Find $x_1, x_2, \cdots, x_k$, the k largest eigenvectors of L (chosen to be orthogonal to each other in the case of repeated eigenvalues), and form the matrix $X = [x_1 x_2 \cdots x_k] \in \mathbb{R}^{n\times k}$ by stacking the eigenvectors in columns.
- Form the matrix Y from X by renormalizing each of X's rows to have unit length (i.e. $Y_{ij} = X_{ij}/(\Sigma_j X_{ij}^2)^{(}1/2))$.
- Treating each row of Y as a point in $\mathbb{R}^k$, cluster them into k clusters via K-means or any other algorithm (that attempts to minimize distortion).
- Finally, assign the original point S_i to cluster j if and only if row i of the matrix Y was assigned to cluster j.

We use a vector to represent the states sequence of a research to be the training set of spectral clustering. A feature vector of researcher r_i is defined as

$$\mathbf{F_i} = (\frac{n_{i1}}{l_i}, \frac{n_{i2}}{l_i}, \cdots, \frac{n_{ij}}{l_i}, \cdots), \tag{7}$$

where l_i is the number of publications of r_i in the dataset and n_{ij} is the number of occurrences of community j in r_i's sequence.

We evaluate the number of clusters using Calinski-Harabaz index [24], which is an evaluation function of clustering. The score is defined as the ratio between the within-cluster dispersion and the between-cluster dispersion. The Calinski-Harabaz index is calculated as:

$$CH = \frac{tr(B_k)}{tr(W_k)} \frac{m-k}{k-1}, \tag{8}$$

where m is the size of training set, k is the number of clusters, B_k is the inter-cluster covariance matrix, and W_k is the intra-cluster covariance matrix.

Fig. 4 shows the line chart diagram of Calinski-Harabaz index using different number of cluster ($1 \sim 8$) and using different kernel coefficient γ (0.01, 0.1, 1, 10) for radial basis function.

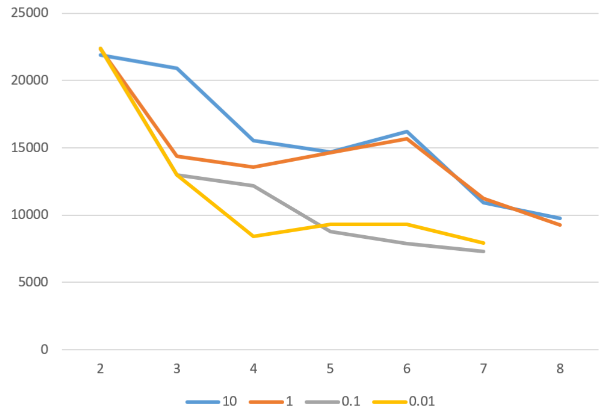


Fig. 4. Line chart diagram of Calinski-Harabaz index to evaluate the result of clustering with different coefficient.

As we can see from Fig. 4, the clustering achieve a maximum score when the number of clusters is set to be 2, which is quite common in many clustering process that less cluster number achieve better score. On the other hand, we can find that clustering result with cluster number equal to 6 often lead to a "local optimal". When $\gamma = 1$, it even surpasses any other results except the global optimal. In the consideration of usefulness of the classification, we would like to set the number of clusters to be 6. We could predict the more number of clusters lead to finer classification of characteristic of researchers.

D. Evaluation

Setting the number of clusters to be 6, we will get the distribution of each the number of researchers in each cluster as it is shown in Fig. 5.

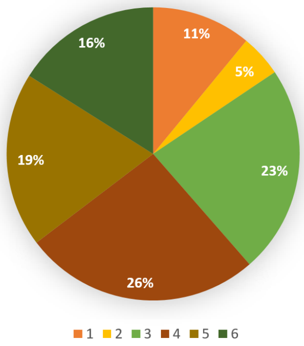


Fig. 5. Sector diagram of the number of researchers in each cluster.

By observation, we can give a definition to the characteristic of the researchers in each clusters as follows:

- Cluster 1: "Freewheeling" or "unfettered" to pick partner, open to make new friends;
- Cluster 2: "Faithful" to only one person, complete nearly all publications with him/her;
- Cluster 3: Do not enjoy steady partnership (reluctant to work with same person again), eager to make new friends;
- Cluster 4: Prefer to work alone;
- Cluster 5: Open to new partners and keep relation with some of them;

- Cluster 6: Partnership evolve over time, sometimes active interaction suddenly break up.

V. Modeling Researcher Research Field Variation

To model a researcher's variation on his/her research field, we first adapt the topic model–LDA–to the title and abstract of researchers' publications to get their probability distribution. Due to the dependence of each topic, we use PCA to do the dimensional reduction, thus making the data orthogonal. Again, we use HMM to model the sequence of topic distribution of researchers. By calculating the discrepancy of two continuous publications of one researcher and its corresponding hidden states, we can evaluate the hidden states and thus do the prediction.

A. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words [25].

Topic models provide an interpretable low-dimensional representation of the documents. LDA assumes the following generative process for each document $\mathbf{w}$ in a corpus D:

1) Choose $N \sim Poisson(\xi)$.
2) Choose $\theta \sim Dir(\alpha)$.
3) For each of the N words w_n:
   (a) Choose a topic $z_n \sim Multinomial(\theta)$.
   (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

This process reveals how the words of each document are assumed to come from a mixture of topics: the topic proportions are document-specific, but the set of topics is shared by the corpus. Our intention is to use topical modeling to give a content based representation of the research domain of the paper in the researcher's sequence of publications.

In practice, ahead of using LDA, we remove the stop-words from all the abstracts and titles in the dataset, thus building a corpus. After this step, we use the corpus and the word frequency to train the LDA with 10 topics. At last, we use the to predict the topic distribution of each paper whose author has more than 10 publications in the dataset. (The reason has been discussed in the last section of this paper.) The top three keywords in each topic is provided in TABLE I.

B. Principal Component Analysis

From TABLE I, we can find that, although some keywords seems to be unique, there are some overlapping between keywords sets of different topics. Actually, since there are hundreds of keywords in one topic of LDA, the overlapping is likely to be inevitable, which would hurt the orthogonality and independency of the training data of following HMM. Consequently, we would like to use Principal Component Analysis (PCA) to project the topic distribution onto low and orthogonal dimensions.

TABLE I
Top three keywords in each topic.

Index	First	Second	Third
1	data	query	database
2	algorithm	problem	algorithms
3	model	language	system
4	system	applications	web
5	game	games	windows
6	information	users	web
7	software	system	systems
8	performance	memory	test
9	network	networks	routing
10	method	based	model

PCA [26] is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components $[v_1, v_2, \cdots, v_d]$, such that $XX^{\mathrm{T}}v = \lambda v$, where X is the input matrix and λ is eigenvalue. This transformation is defined in such a way that the first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set [27].

After using PCA, we convert the topic distribution to a 9-dimension vector, which keeps more than 97% information, as the value of rest dimension is comparatively negligible.

C. Hidden Markov Model

Again, we use HMM to capture the hidden variable of each researcher. This time, we use a matrix in the shape of $l \times 9$ to represent the transition of the research field of each researcher with more than 10 publications in the dataset, where $l \geq 10$ represents the number of publications. We set the number of hidden states to be 20 to ensure that all states can be separate with less error. The training process of HMM after PCA is shown in Fig. 6, while that of HMM without PCA is shown in Fig. 7.

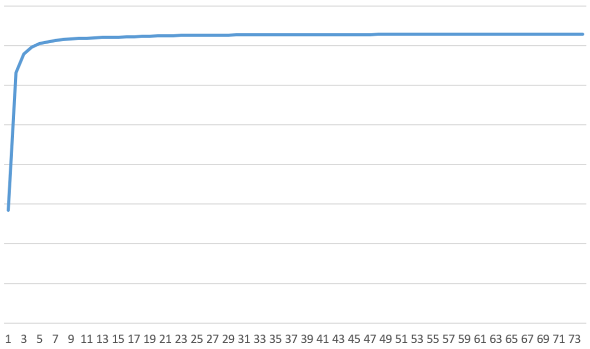


Fig. 6. Line chart diagram of training process of HMM after PCA.

Comparing Fig. 6 and Fig. 7, we can see that the training of HMM after PCA converges quite faster with much less iterations than HMM without PCA, where the former achieve

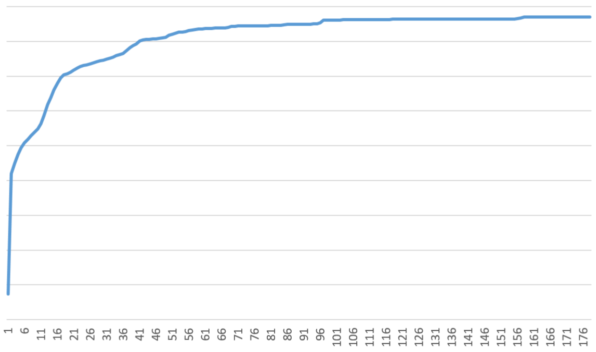


Fig. 7. Line chart diagram of training process of HMM without PCA.

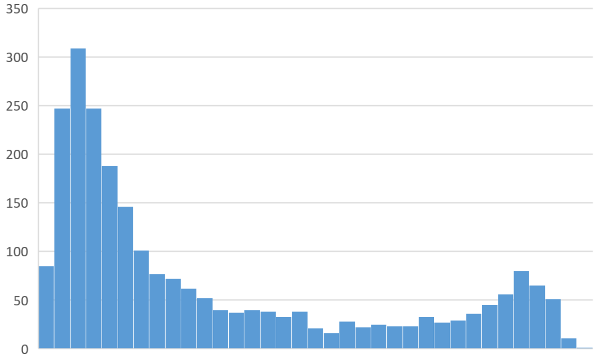


Fig. 8. Histogram of one hidden state's distance. (5th state)

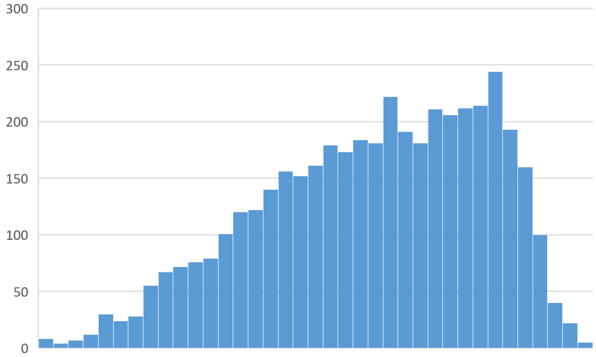


Fig. 9. Histogram of one hidden state's distance. (12th state)

90% of final score by only 2 iteration and the later achieve the same percentage of score by 40 iteration.

The trained HMM can also be used to predict a researcher's future focus point. More details will be introduced in next sub-section.

D. *Evaluation*

To evaluation the learned model, thus predicting the future focus trend of a researcher, we first observe the characteristic of each hidden states.

The system being modeled is assumed to be a Markov process with unobserved states. A Markov chain is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. Consequently, we can use the "distance" between two directly connected states as a measurement. Specifically, for a hidden state z_i with observed variable x_i and its subsequent z_{i+1} and x_{i+1}, where x_i and x_{i+1} are 9-dimension vector as we described in last sub-section. Now, we can use the distance between x_i and x_{i+1} as one measurement of z_i–the very hidden states lead to this transition. Precisely, we use the Euclidean distance:

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{9}$$

to define the distance between two vectors x_i and x_{i+1} after normalization.

After collecting all the distances of each hidden states, we calculate their average and put these distances into histogram, as in Fig. 8 and Fig. 9. These two histograms both have a box size of 0.05. From the two figures, we can see the great discrepancy of the distribution of the data. In Fig. 8, the data concentrates to the left part, which indicates that the following states have less distance from the current states. However, in Fig. 9, the situation is the opposite.

For states like Fig. 8, we can predict that the researcher's research field will not change in short term, because the distances of topic distribution after dimensional reduction tend to be small after this state. Nevertheless, when we find that a researcher is in the state like Fig. 9, it is probable that he/she is going to change the focus point of research. In the 20 hidden states we examined, there are 10 states (less in total frequency) have the distribution like Fig. 9, other 8 states (greater in total frequency) similar to Fig. 8, while the rest 2 states do not exhibit a clear trend in the future.

We also use the histograms to measure the result without PCA based on either Euclidean distance or Jensen-Shannon divergence [28], they also show similar trend that there are two different kinds of states indicating that the focus point of an researcher is going to change or not. But it is not as clear as the result of our proposed method with PCA.

Besides that, HMM also can use the probability distribution of current state and transition matrix to predict the observed variable of the future, which would be the next focus point of the researcher in this case. Practically, we use the HMM to predict the observed data of the next stage. As the input data were treated with PCA, we should re-transform the predicted vector to the LDA topic distribution by the variables stored in trained PCA model. Finally, we can choose several topics with the highest probability to be the prediction results.

To evaluate the accuracy of prediction, we use all the data of researchers except his/her latest publication as the training data for LDA and HMM to get the prediction result. Now, we have to find an approach to measure the discrepancy between the predicted topic and one of the researcher's publication. Since LDA has been used in the proposed method, we have to find another topic representation for the measurement.

In this sense, we involve term frequencyinverse document frequency (tfidf) in our evaluation. In information retrieval, tfidf is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling. The tf-idf value increases proportionally to the number of times a word appears in the document and is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. Tf-idf is one of the most popular term-weighting schemes today; 83% of text-based recommender systems in digital libraries use tf-idf [30].

In our experiment, we use the tf-idf to extract all the keywords in each researcher's latest publication with

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D) > 0, \tag{10}$$

where $tf(t, d)$ equals to the number of times that term t occurs in document d and $idf(t, D)$ is the logarithmically scaled inverse fraction of the documents that contain the word [29]. Then, we collect the first n prediction from the HMM and m words with highest weight of the predicted topic from LDA. Let's denote the keywords list of tf-idf as t and the n keywords lists of prediction as $B = \{b_1, b_2, \cdots, b_n\}$. We define the top k accuracy of prediction as

$$top\, k\, accuracy = \frac{|t \bigcap \{b_1, b2, \cdots, b_k\}|}{|t|}, \tag{11}$$

The result are shown in Fig 10.

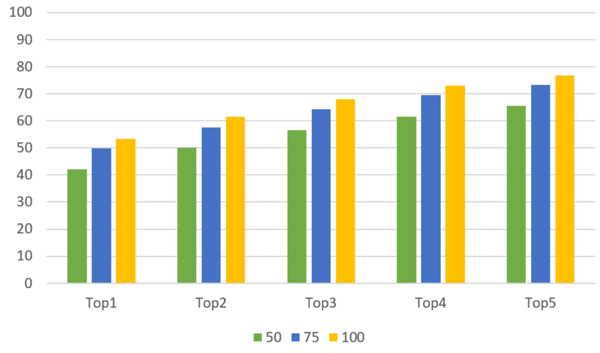


Fig. 10. The bar diagram of top k accuracy of our method, with $m = 50, 75, 100$. The highest accuracy is 76.84%.

As we can see in the Fig 10, our method achieves good accuracy, even though the m words from our model are picked from thousands of papers.

VI. Discussion

A. Limit and Remedy

We may acknowledge that there are some potential drawbacks of our proposed method and we have some possible remedies for them.

First and foremost, researcher profiling has been regarded as a tough problem partly due to its sparsity. Our existing work is based on the AMiner's citation network dataset's first version–Citation-network V1–with 629,814 papers. However, the newest version of the citation network–DBLP-Citation-network V10–has 3,079,007 papers, which is about 4 times than of the first version. More papers lead to more researchers with a number of publications over the threshold as well as comprehensive publication information of on the researcher. It is predictable that there are some miss points in the researcher's publication sequence we used in our work, some of which may be crucial to the classification of the researcher and may alter its subsequent node's hidden states. On the other hand, since, in our proposed method, we have used an algorithm like MCL to average the difference of different states. This kind of deficiency has been reduced. Moreover, clustering is based on the frequency of each states, instead of their appearance at some definite point of time. The risk result from the sparsity of the data would be likely to be controllable. Additionally, as there are a variety of dataset available, using a larger dataset would definitely lead to a better solution.

Secondly, considering the prediction result, it could also be influenced by the problem discussed in last paragraph. One solution to overcome the inaccuracy individual states, we can use the predicted states is a period of time for prediction. For example, if we want to predict the researcher's future focus points, we can use his/her hidden states sequence in the last five years to reduce the risk, instead of using just the variable of his/her last publication, which is could be unreliable.

Thirdly, in the dataset, the specific date of the publication is not provided. As a result, we may put some of the publication from one author in the wrong order. According to our count, there are approximately 3% of the papers suffers from this condition. Except from using a similar method, as we described in last paragraph, we can find a dataset with more details for future research. What's more, we only use the abstract and the title of the paper in this paper, but the text of the paper may improve the quality of the result of LDA.

Last but not least, in the first part of our work, we do not distinguish the first author of the paper, who typically finishes a great proportion job in the group. We may use a diminishing weight to represent each coauthor's contribution to the paper in the future to describe their relationship more accurately.

B. System Design

To implement such researcher profiling system in practice, it would be impossible to train a new model whenever a new batch of publication is available. So we recommend to use the old model to do the clustering and prediction for a certain period of time and retrain the model periodically.

To adopt the proposed model to enhance the performance of recommender system, we can use the feature vector or the index of clustering as a representation of the researcher. Furthermore, for the mentor recommendation system, we should directly use our prediction and classification in the researcher's profile, thus giving students a chance of finding a tutor that can get on well with and share the same research field with him/her.

VII. Conclusion

In this paper, we propose a researcher profiling method based on HMM to reveal the hidden states sequence of a researcher. We present two aspects of researcher profiling by our method: researcher's cooperation pattern and researcher's focus field. According to our experiment, the two models both have a clear result and good performance. In the future, we may try more innovative approaches and larger dataset to modify our method. And more experiment may be conducted to evaluate our models and their hyper-parameters.

Acknowledgment

The author would like to thank Pf. Wang and Dr. Fu for the wonderful course they designed and instruction they gave.

References

[1] F. Ricci, L. Rokach, and B. Shapira, "Introduction to Recommender Systems Handbook," ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, Bc, Canada, DBLP, pp. 1373–1374, June 2011.

[2] H. Jafarkarimi, "A naive recommendation model for large databases," Qa75 Electronic Computers Computer Science, 2012.

[3] A. Oldale, J. Oldale, J. V. Reenen, and M. Campbell, "COLLABORATIVE FILTERING," US, WO/2002/010954, 2002.

[4] Y. Ma, "Collaborative Filtering," Computer Science 57(4), pp. 189-189, 2017.

[5] C. C.Aggarwal, "Recommender Systems: The Textbook," Springer Publishing Company, Incorporated, 2016.

[6] J. Basilico, and T. Hofmann, "Unifying collaborative and content-based filtering," International Conference on Machine Learning. ACM, pp. 9, 2004.

[7] S. Kanoje, S. Girase, and D. Mukhopadhyay, "User Profiling Trends, Techniques and Applications," Computer Science 1(11), pp. 2348-4853, 2015.

[8] L. Yao, J. Tang, and J. Li, "A Unified Approach to Researcher Profiling," Web Intelligence, IEEE/WIC/ACM International Conference o. IEEE, pp. 359-366, 2007.

[9] A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke, "Personalized recommendation in social tagging systems using hierarchical clustering," ACM Conference on Recommend System, pp. 259-266, 2008.

[10] M. Lee, M. Cho, C. Jeong, and H. Jung, "Researcher Profiling for Researcher Analysis Service," SWCIB2014 workshop, collocated with JIST2014 conference, November 2014.

[11] S. Kanoje, S. Girase, and D. Mukhopadhyay, "User Profiling for Recommendation System," Computer Science 29(3), pp. 1005-1007, 2015.

[12] D. Godoy, and A. Amandi, "User profiling in personal information agents: a survey," Knowledge Engineering Review 20(4), pp. 329-361, 2005.

[13] K. Zhang, C. Jiang, Y. Yang, Y. Wang, and G. Zhang, "Research on the Application of User Behavior Auditing Based on Hidden Markov Model in Cloud Environment," 3rd International Conference on Materials Science and Mechanical Engineering (ICMSME 2016), 2016.

[14] N. Mohammadifard, "Modeling User Behavior from E-Commerce Data with Hidden Markov Models and Logistic Regression," unpulished.

[15] H. Kawazu, F. Toriumi, M. Takano, K. Wada, and I. Fukuda, "Analytical method of web user behavior using Hidden Markov Model," IEEE International Conference on Big Data. IEEE, pp. 2518-2524, 2017.

[16] H. Li, G. Fei, S. Wang, B. Liu, W. Shao, A. Mukherjee, and J. Shao, "Modeling Review Spam Using Temporal Patterns and Co-bursting Behaviors," unpublished, 2016.

[17] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "ArnetMiner:extraction and mining of academic social networks," ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. DBLP, pp. 990-998, 2008.

[18] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257286, 1989.

[19] S. v. Dongen, "Graph Clustering by Flow Simulation," Phd Thesis University of Utrecht, 2000.

[20] S. v. Dongen, "A cluster algorithm for graphs," Information Systems [INS], pp. 1-40, 2000.

[21] W. E. Donath, and A. J. Hoffman, "Lower Bounds for the Partitioning of Graphs," Ibm J.res.decelop 17(5), pp. 420-425, 1973.

[22] Tutorial on Spectral Clustering. Available at: http://www.kyb.mpg.de/fileadmin/user_upload/files/publications/attachments/Luxburg07_tutorial_4488%5b0%5d.pdf.

[23] A. Y. Ng, and M. I. Jordan, Y. Weiss, "On spectral clustering: analysis and an algorithm," Proc Nips 14, pp. 849–856, 2001.

[24] T CalinSki, and J. Harabasz, "A dendrite method for cluster analysis," Communications in Statistics 3(1), pp. 1-27, 1974.

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J Machine Learning Research Archive 3, pp. 993-1022, 2003.

[26] K. Pearson, "On lines and planes of closest fit to systems of points in space," Philosophical Magazine 2(11), pp. 559-572, 1901.

[27] S. M. Holland, "PRINCIPAL COMPONENTS ANALYSIS (PCA)," Computers & Geosciences 19(3), pp. 303-342, 1993.

[28] I Grosse, P Bernaolagalvn, P Carpena, R Romnroldn, J Oliver, and H. Eugene Stanley, "Analysis of symbolic sequences using the Jensen-Shannon divergence," Physical Review E Statistical Nonlinear & Soft Matter Physics 65(1), pp. 041905, 2002.

[29] H. P. Luhn, "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," IBM Corp, 1957.

[30] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "Research-paper recommender systems: a literature survey," International Journal on Digital Libraries 17(4), pp. 1-34, 2015.