Journal of Computers Vol. 28, No. 3, 2017, pp. 120-131
doi:10.3966/199115592017062803010

A Chinese Expert Name Disambiguation Approach Based on Hypergraph Partitioning

Ze-Quan Fan[1], Hua Lai[1], Zheng-Tao Yu[2,*], and Xu-Dong Hong[1]

[1] School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650051, China
{1273418658, 405904235, 459102534}@qq.com

[2] Intelligent Information Processing Key Laboratory, Kunming University of Science and Technology, Kunming 650500, China
ztyu@hotmail.com

Received 29 May 2016; Revised 06 December 2016; Accepted 24 December 2016

Abstract. There is always a complex relationship between expert pages, while these relationships is the foundation of expert name disambiguation, traditional graph clustering method only consider the simple binary relation between expert pages, but this method always ignore multi relation between pages which are more complex, In order to utilize the relationship between expert pages efficiently, a Chinese expert name disambiguation approach based on hypergraph partitioning is proposed. Firstly, extract the characteristic attribute of expert in expert page. Secondly, construct a similarity matrix between the documents on different expert pages with the utilization of the attributes features and the associated relationship of the expert pages. Finally, set two kinds of constraints which are strong connection and negative connection, construct an expert page hypergraph model and use the method based on hypergraph partitioning to achieve expert name disambiguation. Through the contrast experiment in the Chinese expert disambiguation, it turns out that the disambiguation effect is much better with the adoption of hypergraph partitioning method.

Keywords: Chinese expert name disambiguation, clustering method, hypergraph partitioning, page-associated relationships

1 Introduction

With the rapid development of Internet technology, huge amounts of information are stored on the network. People often use search engines to search for person information what they are interested in. However, in real life, different people share the same name is widespread, this phenomenon known as the name ambiguity [1]. For example, an online search query for "Li Jie" may retrieve pages of University Dr., the pages of artist, the pages of cellist, and the pages of other persons having that name. When users want to get information of a particular "Li Jie", they need to browse a large number of unrelated web pages, and which has a serious impact on the efficiency of the user's review.

In that case, a method must be used to distinguish between person have the same name. This is often referred to as the personal name disambiguation problem.

Name disambiguation [2-3] is the process of determining whether the same name string refers to the same entity in reality, which has numerous applications in social network analysis [4-5], information retrieval [6] and population knowledge database. It is also applied to automatic question answering, multiple text summarization, hot spot tracking and so on. Names disambiguation has great practical value.

* Corresponding Author

At present, the main idea to solve the problem is to use clustering method [7]. The result of clustering requires that the pages of the same cluster are closely connected, and the connection between different clusters is very sparse. However, there is always a complex relationship between expert pages, traditional clustering methods only extract some features of expert pages for similarity calculation. These methods ignore the association relationships between multiple expert pages. For example, gender feature in expert pages that refer to the same expert must be the same. Association relationships have an important impact on the results of clustering, and how to express association relationships between multiple expert pages still need to be further studied.

Hypergraph model is able to express complex relationships between multiple objects. In recent years, hypergraph has numerous applications in natural language processing, methods based on hypergraph have achieved good results in abstract extraction and keyword extraction [8]. However, hypergraph has not been used in name disambiguation.

Here, we propose to use a Chinese expert name disambiguation approach based on hypergraph partitioning. In this method, expert pages are regarded as the vertexes of hypergraph, the similarity between pages is calculated by expert attribute features which are extracted from expert pages, then we set "strong connection" and "negative connection" as constraints to optimize and generate hypergraph model. In the generated hypergraph, each hyperedge can represent an expert, finally, we use hMeTis algorithm for clustering for disambiguation. Compared with the traditional method, hypergraph model can better represent the relationship between multiple pages, and the experiment proves the feasibility of applying hypergraph in name disambiguation.

2 Related Work

Several approaches have been proposed for name disambiguation using different algorithms. Long Chong et al. [9] used named entities and common nouns as features, and weighted the features by the sentence distance between the features and the names which to be disambiguated; Guerreiro et al. [10] tested the effect of different features on name disambiguation results. He et al. [11] propose a deep learning approach that automatically learns context-entity similarity measure for entity disambiguation. Ferreira et al. [12] extract basic information as a feature.

With the rapid development of Internet technology, Network information is widely used in name disambiguation. Vu et al. [13] using Web Directories to enrich the text feature; Zhou et al. [14] present a method base on exclusive and non-exclusive character attributes; Yang et al. [15] put forward a method that can get more features related to documents through search engine extension with the help of rich Internet resources.

The clustering method based on social network leveraging the fact that each namesake has a unique social community. Nadimi et al. [16] propose an method for author name disambiguation which combines heuristic hierarchical clustering method and social networks to produce clusters. Lang Jun et al. using the co-occurrence of person names in snippets returned by search engine to find and extend the social networks, then automatically clustered into different social communities by the algorithm combining spectral partition and modularity evaluation [17].

Recently, the graph-based method have brought new ideas to the solution of name disambiguation problems. Shin et al. [18]. propose a graph based approach to name disambiguation where the graph model is constructed using the co-author relations. Jin Jiang et al. presented a Chinese expert disambiguation method based on semi-supervised graph clustering, the disambiguation effect is much better than traditional clustering methods [19].

However, traditional graph model can only describe the association between two expert pages, while ignoring the connection between multiple expert pages which have a very important impact on the clustering results. How to describe the complex relationship between multiple pages has become an important issue in name disambiguation. Hypergraph is the generalization of traditional graphs, and which can describe the relationship between multiple entities. In this paper, we use the disambiguation algorithm based on hypergraph partitioning to describe the association between multiple expert pages. The experiment result proves the effectiveness of hypergraph in name disambiguation.

3 Hypergraph Partitioning

Hypergraph based on set theory and graph theory, first proposed by Berge in 1970. Up to now, hypergraph theory has been developed in the field of computer science and artificial intelligence. In mathematics, hypergraph is a generalization of graph, in a graph, an edge only connects two vertices, which means the two vertices are related. In real life, however, the relationship between each object is much more complex. The main difference between hypergraph and graph is that a hyperedge in a hypergraph can contain two or more vertices. For many problems, graph cannot fully express the relationship between each object, as illustrated in Fig. 1. A set of documents need to be differentiated according to topic, the only known information is the author of each document, If we use graph model to express this kind of situation, a vertex represent a document, two vertices are connected to an edge which means the two documents have the same author. This method obviously lost the information that a person who is the author of three or more documents, document written by the same author will most likely belongs to the same topic, and this information is very important. A hyperedge in a hypergraph can contain multiple vertices, therefore, when we use hypergraph to express this kind of situation, a hyperedge can represent a writer, all of the author's document are included in this hyperedge, so compared with graph, hypergraph is more better in expressing the relationship between documents and authors [20].

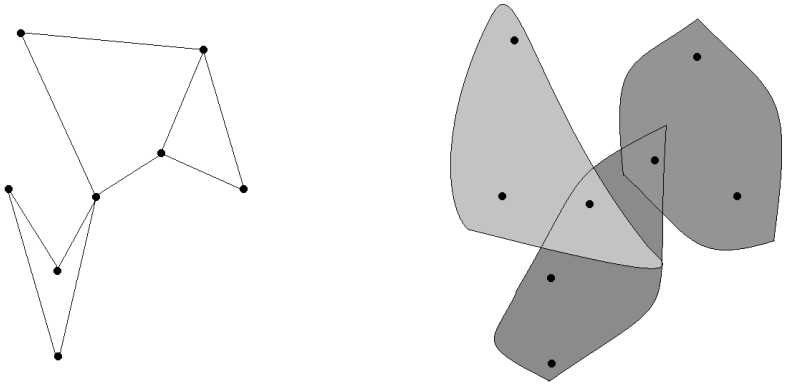

Fig. 1. Hypergraph and graph

Let H(V, E, w) denote a hypergraph where V={v1, v2, v3, ..., vn-1, vn} is the set of vertices, E={e1, e2, e3, ..., em-1, em} is the set of hyperedges, and each hyperedge $e \in E$ is a subset of V. The degree of a hyperedge e is defined by $\delta(e) = |e|$, that is, the cardinality of e. Usually, w is the weight of vertices and hyperedge of hypergraph.

A hypergraph as illustrated in Fig. 2, the hypergraph contains 6 vertices and 4 hyperedges, that is, V={v1, v2, v3, v4, v5, v6}, E={e1, e2, e3, e4}, e1={v1, v2, v3}, e2={v2, v3}, e3={v3, v5, v6}, e4={v4}, the degree of e1 is 3. It can be seen that, if every hyperedge has a degree of 2, the hypergraph reduces to a simple graph.

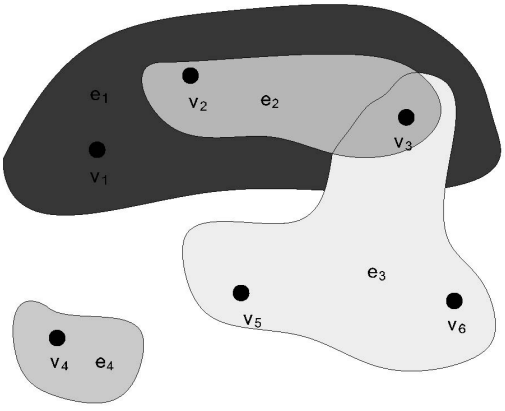

Fig. 2. Hypergraph

Let H(aij) denote the adjacency matrix of hypergraph where i=1, 2, 3, ..., n is number of vertices, j=1, 2, 3, ..., m is number of hyperegde.

$$a_{ij} = \begin{cases} 1, v_i \in e_j \\ 0, v_i \notin e_j \end{cases} \quad \textbf{(1)}$$

The adjacency matrix of hypergraph in Fig. 2 can be expressed as Table 1.

Table 1. The adjacency matrix of hypergraph

	v1	v2	v3	v4	v5	v6
e1	1	1	1	0	0	0
e2	0	1	1	0	0	0
e3	0	0	1	0	1	1
e4	0	0	0	1	0	0

Hypergraph partitioning is to achieve an best cutting method according to the given constraints. Hypergraph partitioning always divided into two categories according to different objective, minimum cut and normalized cut.

$$Min\_cut(A,B) = \sum_{u \in A, v \in B} w(u,v). \quad \textbf{(2)}$$

$$Ncut(A,B) = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(A,B)}{assoc(B,V)}. \quad \textbf{(3)}$$

A and B represent two sub graphs which are non-intersect after cutting, V represents the complete graph, u and v are vertices in the hypergraph. The objective of Min cut is to find a cutting method to minimize the weight loss of hypergraph when hyperedges are cut. Ncut is required to minimize the contacts between subgraphs of the hypergraph after cutting. The objective of cut is often required to maxmize the contacts between vretices in same subgraph and minmize the contacts between vretices in different subgraphs [21].

4 Building the Model of Expert Disambiguation Based on Hypergraph Partitioning

4.1 Selecting the Experts Attribute Association Relationship Features

Chinese experts disambiguation can be transformed into a document clustering of expert pages. In simple terms, put the same expert's pages into a cluster, put different expert's pages into different cluster. So in the process of disambiguation, the expert attribute features in expert pages are very important, we choose expert attribute such as email, phone number, gender and place of birth that can clearly identify themselves [22], then we choose organization, date of birth, position and co-occurrence names in expert pages. These attributes are shown in Table 2.

Table 2. Experts attribute

Serial Number	Feature Name	Example
1	email	12345@hotmail.com
2	phone number	000-123456
3	gender	male
4	place of birth	Harbin
5	organization	Harbin institute of technology
6	date of birth	1939.10.20
7	position	The dean of Computer College
8	co-occurrence names	Jim、Jake

Expert attribute is the key to distinguish experts, so how to extract expert attribute is very important. Bootstrapping algorithm is a kind of machine learning method which can learn out of a large number of

high accurate patterns for attribute feature extraction.

Bootstrapping algorithm start with a set of extraction attribute as seeds, and then applying an incremental iterative procedure to find new features [23], process is as follows:

(1) For a expert feature, searching the Internet through the seed words and getting feature values;

(2) Calculating credibility of all feature values;

(3) Put five feature values with the highest credibility into the feature values dictionary;

(4) Traverse all of the documents to get context of the 5 feature values, then make the context as a candidate patterns;

(5) Calculating credibility of all candidate patterns;

(6) Put three candidate patterns with the highest credibility into the pattern store;

(7) Repeat the process until reaching the threshold or no new pattern generated.

After extracting expert attribute, we define the relationship features of expert attributes as Table 3.

Table 3. Experts attribute association relationship features

Serial Number	Feature Name	Feature type	Eigenvalue (fm)	Feature Weight(αm)
1	Whether the same email	Boolean	0,1	α1
2	Whether the same phone number	Boolean	0,1	α2
3	Whether the same gender	Boolean	0,1	α3
4	Whether the same place of birth	Boolean	0,1	α4
5	Whether the same organization	Boolean	0,1	α5
6	Whether the same date of birth	Boolean	0,1	α6
7	Whether the same position	Boolean	0,1	α7
8	Whether the same co-occurrence names	Boolean	0,1	α8

4.2 Expert Attribute Feature Constrainting

In the experts attribute features which are extracted in expert pages, there are some attributes that can accurately judge whether the expert mentioned on both pages is the same person. For example, experts of different gender is certainly not the same person, experts have the same phone number or email is certainly the same person. In order to effectively use these attributes to distinguish experts, we set two kinds of constraints which are strong connection and negative connection.

(1) strong connection: vertices that satisfy the constraint must be in the same hyperedge.

(2) negative connection: vertices that satisfy the constraint must not be in the same hyperedge.

In the Table 4 as below, it defines the constraint rules for “strong connection” and “negative connection”.

Table 4. strong connection and negative connection

strong connection	negative connection
Email is the same	Differnet gender
Date of birth is the same	Different date of birth
Phone number is the same	Different place of birth

Setting strong connection and negative connection can optimize the effect of clustering, each of these constraints can independently determine whether the two experts are the same person.

4.3 Experts Disambiguation Model-building Process

After extracting features of expert pages and setting constraints, we can start to build expert disambiguation model based on the hypergraph partitioning.

Let H(V, E) denote a hypergraph where V is the node set of the expert pages, E is the set of hyperedges. In the process of hypergraph model building, vertex similarity can directly decide the establishment of the hyperedge, so how to calculate vertex similarity is very important. In this paper, we selected eight attributes in expert pages, but on account of the frequency that the attribute we selected

appears in all of the expert pages and the resolution capability featured by this attribute in different pages, calculate the attribute feature weight [8] through the TF-IDF algorithm with which calculation is made based on word frequency. Then we represent a document in the form of a vector, each vector will be expressed by its feature term and the weight to construct a document vector space.

The cosine of the vector space angle of the two documents is employed for defining the similarity between the two expert evidence-page nodes, based on which we can get the initial similarity matrix A.

Suppose there are two arbitrary age nodes x_i, $x_j \in V$, TF-IDF formula is as follows:

$$W_{t,x} = TF_{t,x} \times IDF_{t,x} \tag{4}$$

$$TF_{t,x} = \frac{N}{M} \tag{5}$$

$$IDF_{t,x} = \log \frac{X}{X_t} \tag{6}$$

$W_{t,x}$ is the weight of the feature item t in the document x, $TF_{t,x}$ represents the occurrence frequency of t in the document x. $IDF_{t,x}$ is known as the document frequency of features to reflect the distribution of feature item t in the whole document set and the distinction ability of this feature item to a certain extent, and where X represents the number of all the documents in the document set, X_t represents the occurrence frequency of t in the document set.

After that, the initial similarity A_{ij} of two page nodes will be defined as below through the included angle cosine of both document vectors:

$$A_{ij} = Sim\left(x_i, x_j\right) = \cos\theta = \frac{\sum_{t=1}^{n} W_{t,x_i} \times W_{t,x_j}}{\sqrt{\left(\sum_{t=1}^{n} W_{t,x_i}^2\right)\left(\sum_{t=1}^{n} W_{t,x_j}^2\right)}} \tag{7}$$

After calculating the initial similarity, we can use strong connection and negative connection to correct the initial page similarity matrix A_{ij} to obtain the final similarity matrix A*, Specific approach is as follows: If two vertices satisfy the "strong connection" constraint, there should be certain strong association between the two vertices, so we triple the similarity of the two vertices. However for the "negative connection", as soon as the rules for "negative connection" constraints are applied, set the similarity as 0 in despite of the fact that there are many same or similar attribute values between two pages. As shown in the formula 8:

$$A^* = \begin{cases} 3 \cdot A_{ij}, (x_i, x_j) \in S \\ 0, (x_i, x_j) \in N \\ A_{ij}, (x_i, x_j) \notin S \,\&\, (x_i, x_j) \notin N \end{cases} \tag{8}$$

Where S is "strong connection", N is "negative connection".

After get the final similarity matrix, we can build the hypergraph model, when the similarity between vertices is higher than the threshold, put the vertices in a hyperedge.

The vertices and hyperedges in hypergraph should have a weight, which is convenient for us to accurately cut the hypergraph.

The weight of the vertices in the hypergraph defined as follows:

$$W_{vi} = \sum_{m=1}^{8} \alpha_m \cdot f_m \tag{9}$$

Where α_m is the page-associated feature weight obtained by training. In Table 3, e-mail, phone number and date of birth are important factor which are determined whether the two page is the same person, therefore, the weight of each feature are shown in Table 5.

Table 5. Weight of experts attribute

Serial Number	Feature Name	Feature Weight(αm)
1	email	α1=0.6
2	phone number	α2=0.6
3	gender	α3=0.5
4	place of birth	α4=0.5
5	organization	α5=0.5
6	date of birth	α6=0.8
7	position	α7=0.5
8	co-occurrence names	α8=0.4

Assuming a hyperedge e_j contains n vertices, the weight of the hyperedge e_j in the hypergraph defined as follows:

$$W_{ej} = \frac{\sum_{i=1}^{n} W_{vi}}{n}, v_i \in e_j \tag{10}$$

After building the hypergraph model, we can start the hypergraph partitioning.

4.4 Hypergraph Partitioning

The partitioning method in this article we used is the hMeT is algorithm which developed by University of MINNESOTA United States [24-26], these algorithm consist of three phases: coarsening phase, initial partitioning phase, and uncoarsening and refinement phase. During the coarsening phase, the method firstly create a small hypergraph, bisection of the small hypergraph is not significantly worse than the bisection directly on the original hypergraph, and in coarsening phase, the size of hyperedges and vertices was reduced; during the initial partitioning phase, a bisection of the coarsest hypergraph is computed; and during the uncoarsening and refinement phase, a partitioning of the coarser hypergraph is used to obtain a partitioning for the finer graph, the method successively projecting the partitioning to the next level finer hypergraph and using a partitioning refinement algorithm to reduce the cut and improve the quality of the partitioning. Process of the method is illustrated in Fig. 3.

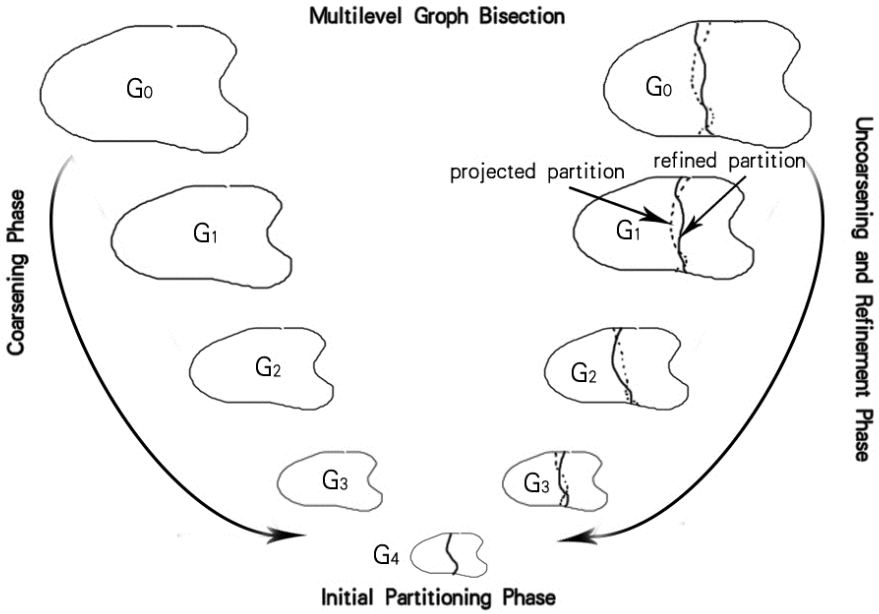

Fig. 3. Process of the method based on hypergraph partitioning

For the hypergraph and segmented results, a cost function defined as follows:

$$cut\cos t(H(A_k)) = \frac{1}{n(k-1)}\sum_{i=1}^{k}\frac{E(P_i)}{w(P_i)} \tag{11}$$

Where H(Ak) is the k-way partitioning of hypergraph, n is the number of vertices, k is the number of sub hypergraph, Pi is the sub hypergraph, w(Pi) is the sum of the weight of all the vertices in the hyperedge which are cut in the sub hypergraph i.

In the article, we use hMeT is to compute a 2-way partitioning (for k = 2), each time we cut the original hypergraph into two sub hypergraph, if the weight of the hyperedge which are cut in this process do not exceed the threshold we set in advance, we continue to cut the sub hypergraph into two, until the threshold is exceeded.

The specific method is as follows:

```
program 2-way partitioning: (original hypergraph H, threshold value
μ, Output: partitioned hypergraph H)
     For each cutting point in H do;
          Calculate cut cost(H(A2));
     End for;
     Get min(cut cost(H(A2)));
     Generate two sub hypergraph;
     If cut cost(H(A2)) < μ then;
          For each sub hypergraph do;
               Cut the sub hypergraph with a 2-way partitioning;
          End for;
     Else;
          Output the current sub hypergraph;
     End if.
```

5 Experiments and Analysis

5.1 Experiment data preparation

In order to verify the effectiveness of Chinese experts disambiguation method based on hypergraph partitioning, we have designed three contrast experiments in this paper. The experimental data is from the CNKI-based (Chinese National Knowledge Infrastructure) resource network to crawl the name of 2,000 Chinese experts devoted to the information processing in the computer field. We have sorted out the name of 200 experts whose information is ambiguous on the Baidupedia, and reserved the top 10 web pages recalled through Baidu search as the page-associated documents for these experts to make the total expert-related pages up to 2,000. In this test, we've made choices based on four test sets. Each time, we chose totally 800 pages with the name of 80 experts having been selected randomly from the raw data set as the test data set. The expert feature constraints ("strong connection" and "negative connection") is between 0 and 300 derived from the experiments which are randomly generated from the training data set. Finally we use clustering evaluation index to evaluate the clustering results of test data set.

5.2 Experimental Results and Analysis

The quality of result based on clustering algorithm should be evaluated by a fair and objective evaluation method, in this paper, we use the F-value [27] which is defined as below:

$$\Pr e = \frac{T_p}{T_p + F_p} \tag{12}$$

$$\mathrm{Re}c = \frac{T_p}{T_p + F_n} \tag{13}$$

$$F = \frac{2\Pr e \times \mathrm{Re}c}{\Pr e + \mathrm{Re}c} \tag{14}$$

Where T_p is the number of evidence-page documents that the two documents together in one cluster are classified correctly, F_p represents the number of evidence-page documents that the two should not be placed in one cluster are divided into one falsely, Fn is the number of evidence- page documents that the two should not be separated are parted wrongly.

Experiments in this paper are as follows:

Experiment 1 tests the influence of different number of feature constraints ("strong connection" and "negative connection") on the result of hypergraph partitioning method. Firstly we determine the number of strong connection and negative connection, twenty experiments were carried out for each given number of constraints. There are four test document sets in the experiment and finally the average output of each test represents the clustering performance. Fig. 4 reveals the changes of the P, R and F-value indexes on the four test sets with different numbers of constraints.

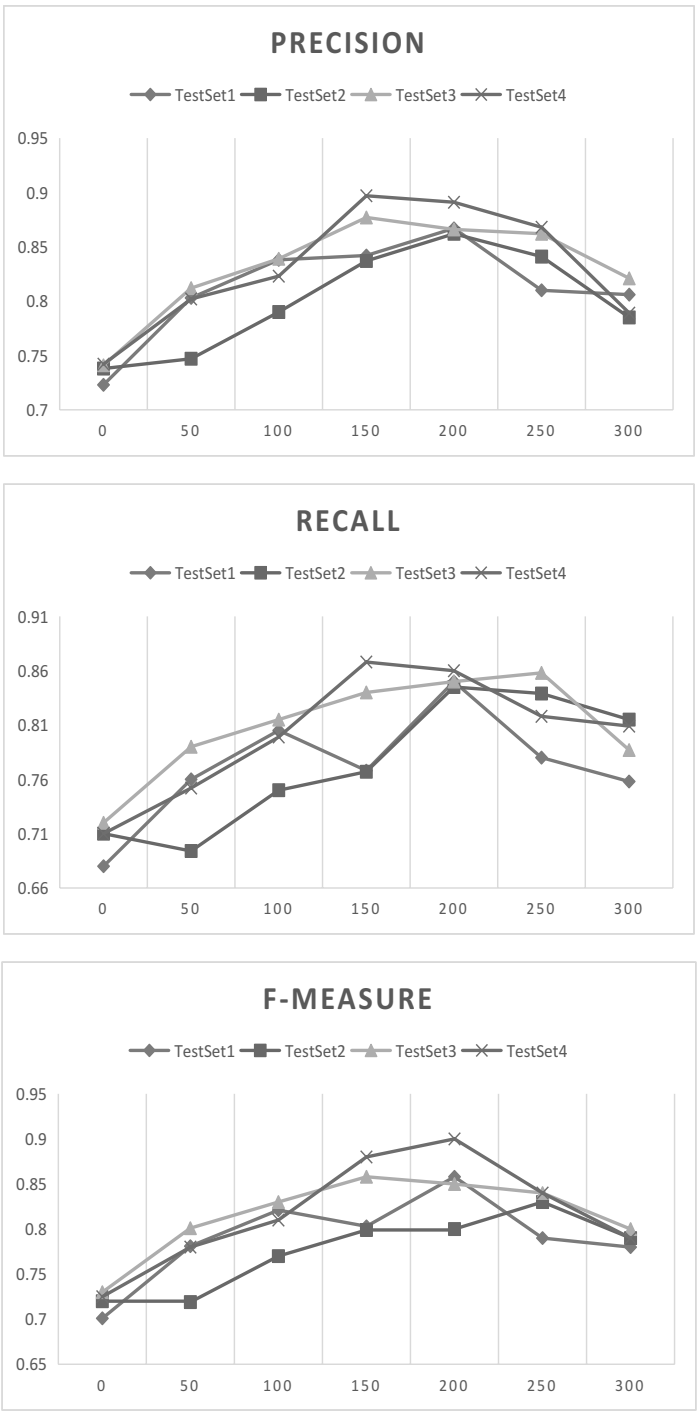

Fig. 4. Influence of different number of constraints on P, R and F-value in the test sets

It can be seen from Fig. 4 that the P, R and F-value indexes obtained by the method based on hypergraph partitioning show upward trend as a whole with the increase in the number of constraints. When the number of constraints is between 150 and 200, the P, R and F-values reached the maximum and the number of clusters obtained at the moment is optimal. It's also obvious that when the number of constraints is greater than 200, the clustering performance in the test begin to degrade slowly.

Experiment 2 is to test the influence of hypergraph partitioning by comparing before and after adding constraint rules, the number of constraint in the test are 200, the result is in Table 6.

Table 6. Influence of hypergraph partitioning before and after adding constraint

Sets|	Clustering method without constraint			Clustering method with constraint		
Methods	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
TestSet 1	76.62	72.58	74.55	85.72	85.06	85.38
TestSet 2	76.87	73.53	75.16	86.20	84.55	85.37
TestSet 3	75.45	70.13	72.69	86.67	85.11	85.88
TestSet 4	77.21	73.39	75.25	89.14	86.07	87.58

It can be seen from Table 3 that the P, R and F-value of our method with the expert feature constraint are significantly improved than that without the constraint. Therefore the feature constraints has brought a significant effect on the result of the expert name disambiguation.

Experiment 3 is to verify the advantage of hypergraph partitioning method, in this test, we use three Chinese disambiguation methods to compare the clustering results. These three methods are as follows:

T1, a Chinese name disambiguation based on analysis of sentential semantic structure [17]. This method firstly built a graph of social relationships according to the characters of human social relationships, then clustering the relationships by combing with the attributes of the name entity.

T2, a Chinese name disambiguation based on clustering occupation characteristics [1]. This method considers the occupation is a vert important feature, it build a basic occupation dictionary through the Internet, occupation is extracted as a feature, supplemented by person names and works to make up for the problems of occupation missing, then clustering by agglomerative hierarchical algorithm.

T3, a Chinese name disambiguation method based on combined features. The method extract different features related to the name and then creating combined features by vector space model, then clustering by hierarchical algorithm.

T4, The methods mentioned in this article.

The result is in Table 7.

Table 7. Experts name disambiguation results of different algorithms

Sets|	T1			T2			T3			T4		
Method	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
TestSet 1	81.25	79.98	80.61	78.74	77.26	77.99	73.64	70.89	72.24	84.02	80.73	82.34
TestSet 2	81.82	80.43	81.12	78.06	77.35	77.70	72.28	71.33	71.80	86.56	83.42	84.96
TestSet 3	80.70	79.17	79.93	79.40	77.82	78.60	75.52	73.21	74.35	84.43	82.39	83.40
TestSet 4	82.54	80.76	81.64	77.18	74.91	76.03	75.34	72.96	74.13	86.96	86.45	86.70

It can be seen from Table 7, T1 method is based on analysis of sentential semantic structure which can avoid the problem of sparse features of expert pages, T2 method focuses on the occupational characteristics, T3 method combines features together. These three method are only consider the features of expert in a single page, but ignore the relationship of features in multiple pages. So the method based on hypergraph partitioning obtained the best experimental results.

6 Experiments and Analysis

In this paper, we propose a Chinese expert name disambiguation approach based on hypergraph partitioning, expert pages in this method are regarded as the vertexes of hypergraph, the similarity between pages is calculated by features which are extracted from expert pages, then we set "strong connection" and "negative connection" as constraints to optimize and generate hypergraph model.

hypergraph model can well express the relationships between multiple expert pages, the experiment results show the validity and the effectiveness of the method. But this method still have some defect, method rely heavily on features, which are not owned by every expert page, when an expert page only own few low-weight features, the accuracy of similarity calculation will be reduced, and so does the accuracy of page clustering. Our next work will focus on how to improve the accuracy of similarity calculation between expert pages, we will consider adding more features to expert pages, such as research direction and web link, etc., we will also optimize weight-calculation of hyperedge, we hope that through these to improve the accuracy of clustering.

Acknowledgement

This paper is supported by National Nature Science Foundation (No. 61472168,61175068), and The Key Project of Yunnan Nature Science Foundation (No.2013FA130), and Science and technology innovation talents fund projects of Ministry of Science and Technology(No.2014HE001).

References

[1] Y.-L. Yang, J. Zhou, B.-C. Li, Y.-Y. Xi, Name disambiguation based on clustering by step, Journal of Data Acquisition and Processing 1(2016) 213-222.

[2] J. Tang, A.C.M. Fong, B. Wang, J. Zhang, A unified probabilistic framework for name disambiguation in digital library, IEEE Transactions on Knowledge and data engineering 24(6)(2012) 975-987.

[3] B. Zhang, T.K. Saha, M.A. Hasan, Name disambiguation from link data in a collaboration graph. In: Proc. Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on, 2014.

[4] P.-Y. Chen, B. Zhang, M.A. Hasan, A.O. Hero, Incremental method for spectral clustering of increasing orders, in: Proc. KDD Workshop on Mining and Learning with Graphs, 2016.

[5] B. Zhang, S. Choudhury, M.A. Hasan, X. Ning, K. Agarwal, S. Purohit, P.G.P. Cabrera, Trust from the past: Bayesian personalized ranking based link prediction in knowledge graphs, in: Proc. SDM Workshop on Mining Networks and Graphs, 2016.

[6] S. Choudhury, K. Agarwal, S. Purohit, B. Zhang, M. Pirrung, W. Smith, M. Thomas. NOUS: Construction and querying of dynamic knowledge graphs. in: Proc. Data Engineering (ICDE), 2017 IEEE 33rd International Conference, 2017.

[7] H. Zhang, The Research on Personal Name Disambiguation and Character Relationship Extraction Merging Sentential Semantic Feature, Beijing Institute of Technology, Beijing, 2015.

[8] M.-O. Peng, H.-U. Po, X.-J. Hunag, T.-T. He, A hypergraph based approach to collaborative text summarization and keyword Extraction, Journal of Chinese Information Processing 6(2015) 135-140.

[9] C. Long, L. Shi, Web person name disambiguation by relevance weighting of extended feature sets, in: Prco. Notebook Papers/LABs/Workshops, 2010.

[10] J. Guerreiro, D. Goncalves, D.M. de Matos, Towards a fair comparison between name disambiguation approaches, in: Proc. the 10th Conference on Open Research Areas in Information Retrieval, 2013.

[11] Z.Y. He, S.-J. Liu, M. Li, M. Zhou, L.-K. Zhang, H.-F. Wang, Learning entity representation for entity disambiguation, in: Proc. the 51st Annual Meeting of the Association for Computational Linguistics, 2013.

[12] A.A. Ferreira, M.A. Goncalves, A.H.F. Laender, Disambiguating author names using minimum bibliographic information, World Digital Libraries-An International Journal 7(1)(2014) 71-84.

[13] Q.M. Vu, A. Takasu, J. Adachi, Improving the performance of personal name disambiguation using web directories, Information Processing and Management 44(4)(2008) 1546-1561.

[14] X. Zhou, C. Li, M.-H. Hu, H.Z. Wang, Chinese name disambiguation based on exclusive character attributes, in: Proc. CCTR 2010, 2010.

[15] X.-X. Yang, P.-F. Li, Q.-M. Zhu, Name disambiguation based on query expansion, Journal of Computer Applications, 9 (2012) 2488-2490+2507.

[16] M.H. Nadimi, M. Mosakhani, A more accurate clustering method by using co-author social networks for author name disambiguation, Journal of Computing and Security 1(4)(2015) 102-111.

[17] J. Lang, B. Qin. Person name disambiguation of searching results using social network, Chinese Journal of Computers 7(2009) 1365-1375.

[18] D. Shin, T. Kim, J. Choi, J. Kim, Author name disambiguation using a graph model with node splitting and merging based on bibliographic information, Scientometrics 100(1)(2014) 15-50.

[19] J. Jiang, X. Yan, Z. Yu, J. Guo, W. Tian, A Chinese expert disambiguation method based on semi-supervised graph clustering, International Journal of Machine Learning and Cybernetics 6(2)(2014) 197-204.

[20] J. Xu, A hypergraph-based semantic information fusion method for image scene classification, [dissertation] Beijing: Beijing Jiaotong University, 2014.

[21] W.-P. Chen, Researchon hypergraph partition for coreference resolution, [dissertation] Harbin: Harbin Institute of Technology, 2012.

[22] W. Tian, T. Shen, Z. Yu, J. Guo, Y. Xian, A Chinese expert name disambiguation approach based on spectral clustering with the expert page-associated relationships, in: Proc. 2013 Chinese Intelligent Automation Conference, 2013.

[23] L. Li, Chinese personal name disambiguation based on attribute information, [dissertation] Beijing: Beijing University of Posts and Telecommunications, 2012.

[24] G. Karypis, V. Kumar, Multilevel k-way hypergraph partitioning, VLSI Design 11(3)(2000) 285-300.

[25] G. Karypis, R. Aggarwal, V. Kumar, S. Shekhar, Multilevel hypergraph partitioning: application in vlsi domain, in: Proc. the Design and Automation Conference, 1997.

[26] D. Zhou, J. Huang, B. Scholkopf, Learning with hypergraphs: clustering, classification, and embedding, in: Proc. the NIPS, 2006.

[27] L-Y. Deng, L.-M. Yan, Y.-T. Long, Multiple ant colonies clustering combination algorithm based on undirected hypergraph, Industrial control computer 27(7)(2014) 129-131.