Name Disambiguation from Scratch

Zhao Fuhui
1335921828@qq.com
School of Computer Science and Technology, Shandong University
Qingdao, Shandong

ABSTRACT

With the progress of the society, the number of papers in the world is increasing greatly. The massive literature knowledge base provides convenient literature retrieval and learning research services for researchers. However, the phenomenon of duplicate authors' names is very common in these documents, which has a serious impact on the accuracy of the retrieval results and the quality of the content. In order to solve the problem of duplicate authors' names disambiguation, the author name disambiguation algorithm based on entity association graph is used in this paper, which makes full use of the attribute characteristics and intrinsic association of the paper. And the combination of Agglomerative clustering algorithm makes the final clustering result more accurate.

KEYWORDS

Duplicate Name Disambiguation, Entity Association Graph, Hierarchical Clustering, Algorithm Theory

1 INTRODUCTION

In the literature knowledge base, with the sharp increase in the number of scientific and technological documents every year, a large number of authors' duplicate names reduce the accuracy of literature retrieval. When users search documents according to the authors' names, there will often be many irrelevant documents published by authors with the same name in other fields. These are not the information users really need, which interferes with users' judgment of retrieval results and delays the cycle of scientific research work. Therefore, the disambiguation problem of authors with the same name needs to be solved urgently. We can solve the problem of duplicate names of authors in documents according to various characteristic attributes used by authors when publishing papers, such as information of published meetings, keywords, abstracts, collaborators, etc., and determine the identity information of authors from various aspects, thus realizing the disambiguation of duplicate names.

What we need to do now is to disambiguate papers with the same author name. Given a pile of papers with the same name, it is required to return a group of paper clusters, so that papers within a cluster are all one person, and papers between different clusters do not belong to one person. The ultimate goal is to identify those papers by authors with the same name that belong to the same person. For example, there are five papers a,b,c,d and e under the name ”Tom”. Now we disambiguate these six papers and get a cluster of papers {a,b} belong to one person named ”Tom” and {c,d,e} belong to another person named ”Tom”. In this paper, author name disambiguation algorithm based on entity association graph is used, and each paper entity with the same name is represented by a graph node. For paper entities with the same author name, if some features of the entities are associated, it is proved that the two documents must also be associated. An edge is added between the two entity nodes, and finally the whole entity association graph is divided into a certain number of unconnected sub-graphs. The point set formed by these sub-graphs is the disambiguated result. Combined with the Agglomerative clustering algorithm, the final clustering result is more accurate.

2 RELATED WORK

The general idea to solve the problem of disambiguation of the author's name is to use clustering algorithm, which usually takes the papers of the same author as the same group. The similarity between papers in the same group calculated by the similarity function determines whether to use some clustering algorithm to cluster the related papers into a cluster. This method usually needs to go through two steps: first, define the similarity function, and then use a specific clustering algorithm to cluster the paper records of the same author according to the similarity. Among them, the similarity function can be divided into three categories. The first type is a predefined similarity function, that is, a classical similarity function based on some attributes [1]; The second type is the similarity function [2] obtained after training the labeled training sets by supervised machine learning methods; The third type is similarity function based on graphs [3], and the graphs used are often obtained by exploring the relationship between authors and co-authors. The clustering algorithm can directly use classical clustering algorithms such as k-means, hierarchical clustering, etc. [4] uses the idea of atomic clustering. The idea is to cluster with strong rules first. For example, if two papers have more than two co-authors, then the two papers belong to the same category, which can ensure the accuracy within the cluster. Then the previous clusters are merged with weak rules, thus improving the recall rate.

Agglomerative clustering

Agglomerative Clutsering is a bottom-up hierarchical clustering method, which can calculate the distance between classes according to the specified similarity or distance definition. There are three calculation methods between classes: the shortest distance method, the longest distance method and the average distance method.

Minimum distance:$d_{min}(C_i, C_j) = \underset{x \in C_i, y \in C_j}{min} \; dist(x, y)$

Maximum distance:$d_{max}(C_i, C_j) = \underset{x \in C_i, y \in C_j}{max} \; dist(x, y)$

Average distance:$d_{avg}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{x \in C_j} dist(x, y)$

The flow of hierarchical clustering algorithm:

(1) Given the sample set, determine the distance measurement function of clustering clusters and the number K of clustering clusters;
(2) Take each sample as a class and calculate the distance between them;
(3) merging the two classes with the smallest distance into a heart class;
(4) Recalculate the distance between the heart class and all classes;
(5) Repeat (3-4) until the required number of clusters is reached.

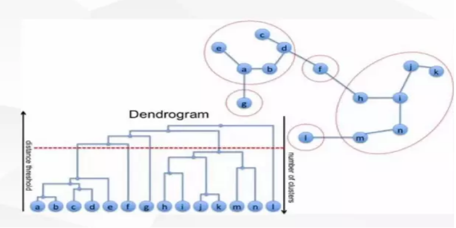


Figure 1: Agglomerative Clutsering

The clustering algorithm needs to indicate the number of clustering results. In this experiment, the algorithm is combined with the algorithm based on association graph, and the number of classes obtained by association graph algorithm is used in the clustering algorithm.

3 APPROACH

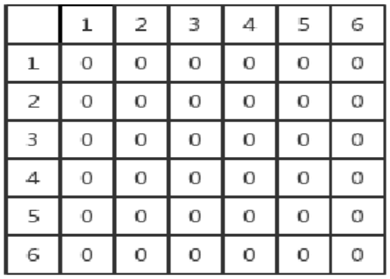
Data preprocessing

The first step is to preprocess the data. The format of many attribute information in the paper is not uniform, which is not convenient for later processing, so we need to preprocess the data and modify all the information into a uniform format. The attributes to be modified include the name of the author of the paper and the name of the organization to which the author belongs. For the author's name, we have changed it to use underscore to connect the parts.For the expression of organization name, some use abbreviations and some use full names,We will change the abbreviated names to full names.For example, "Sch" is changed to "School" and "Dept" is changed to "Department".

Disambiguation method

In this paper, the entity association graph based on graph theory is used to eliminate the phenomenon of duplicate names of authors. The paper entities are represented by graph nodes. For two paper entities with the same author name, if some features of the entities are associated, it is proved that the two paper entities are also associated. An edge is added to the two paper entity nodes, and the strength of the association relationship between different features of the entities is expressed by the weight of the edge. Then the different correlation states are divided into a certain number of disconnected subgraphs, and the point set formed by these subgraphs is the disambiguation result.

There are many papers under each author's name. We construct a two-dimensional array A[n][n] for each author's name to represent an undirected graph,$i, j \in n$ present nodes of graph. If A[i][j]=1, it means that papers i and j under the author's name are related, and there is an edge between the two nodes, belonging to the same class. If A[i][j]=0, it means that the paper i under the author's name is not related to j. At the beginning, we assign all initial values of 0 to the two-dimensional array, indicating that there is no correlation between each paper. Use a case where there are six papers under one author's name as an example:

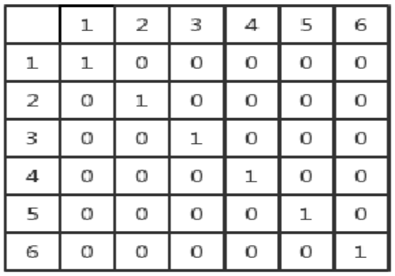


	1	2	3	4	5	6
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0

Figure 2: Array initialization

Since a paper is definitely related to itself, I set the diagonal part of the two-dimensional array to 1.

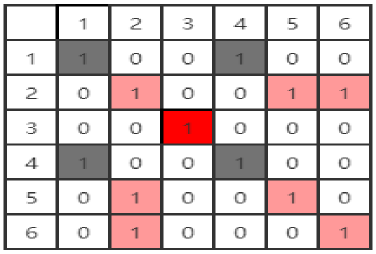
We will consider the number of collaborators and the organizations to which the papers belong to determine whether there is a correlation between the two papers. In reality, a man named "Tom" has two friends named "Tony" and "Mike" who wrote a paper together, while another man named "Tom" happens to have two friends named "Tony" and "Mike"



	1	2	3	4	5	6
1	1	0	0	0	0	0
2	0	1	0	0	0	0
3	0	0	1	0	0	0
4	0	0	0	1	0	0
5	0	0	0	0	1	0
6	0	0	0	0	0	1

Figure 3: Diagonal set to 1

who also co-wrote a paper. The probability of this happening is very small. And if it is common for several people to work together on several papers in the same field, for example, ”Tom”, ”Tony” and ”Mike” are classmates in the same laboratory, then they may work together on many papers in the same field. Therefore, if we find that the two articles have two or more identical collaborators, we think that the two articles are related. In reality, the probability of having two people with the same name under the same organization is also very small, so if the two papers belong to the same organization, then the two papers are considered to be related and should be grouped into one category. According to the above judgment, we get the following two-dimensional array settings:



	1	2	3	4	5	6
1	1	0	0	1	0	0
2	0	1	0	0	1	1
3	0	0	1	0	0	0
4	1	0	0	1	0	0
5	0	1	0	0	1	0
6	0	1	0	0	0	1

Figure 4: Final setting of dimension array

Then the correlation states between the papers are divided into a certain number of disconnected subgraphs, and the point set composed of these subgraphs is the disambiguation result. The corresponding method is to conduct a breadth search operation on the two-dimensional array obtained above, and finally divide several sub-graphs as shown in fig. 5. It can be seen that papers 1 and 4 belong to the same author, papers 2, 5 and 6 belong to the same author, and paper 3 belongs to the same author.

Through the method of association graph, we get the number of classes that the paper should divide under a name, then we can use the number of classes to divide the paper again using the traditional clustering algorithm. I used Agglomerative Clutsering algorithm to cluster the paper again. The basis for clustering is the cooperators, affiliated organizations and abstracts of the paper.

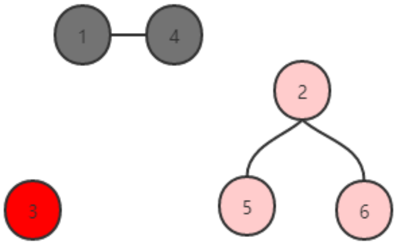


Figure 5: Final disambiguation result

Table 1: The data format of each paper

Domain	Data type	Meaning	For example
id	string	ID of the paper	l7VHz8Vd
title	string	title	Data mining
authors.name	string	The author’s name	Jiawei Han
author.org	string	The author’s organization	Shandong university
venue	string	Meetings/Periodicals	Inteligencia Artificial
year	int	Year of publication	2000
keywords	list of strings	keywords	[”data mining”, ”structured data”]
abstract	string	abstract	Our ability to generate...

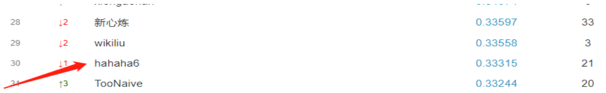
4 EXPERIMENT

Data Set

There are two data files in this task. The data in file 1 is organized into a dictionary (dic1) and stored as JSON objects. The key of dic1 is the author’s name, and the value of dic1 is the set of all paper IDs representing the author with the same name. The data in the other file 2 is represented as a dictionary, its key is the paper ID, and its value is the corresponding paper information, including title, author.name, author.org, venue, year, keywords, abstract. What we have to do is to cluster the papers in the value corresponding to each key in file 1 according to the information of each paper in file 2.The specific format is shown in Table 1.

Method implementation

Write a program to implement the above method. The first step is to preprocess the data, uniformly process the names of the authors of the paper, and restore all the names of the organizations to which the paper belongs to full names. The following is the real disambiguation process, which uses circular operation to construct a two-dimensional array for all papers under each author’s name, uses the two-dimensional array to represent an undirected connection graph, then judges whether the two papers are related or not according to the cooperators and affiliated organizations of the papers, and sets the corresponding two-dimensional array as a set-one process.Then we did a breadth search operation on the last two-dimensional array and obtained several disconnected subgraphs. The number of subgraphs is the number of clustered classes. Finally, the paper's collaborators, affiliated organizations and abstracts are used to cluster again using hierarchical clustering algorithm.In the daily submission, the score of online test is 0.33315, my ranking is 30,and the total number of submissions is 21.



28	↓2	新心炼	0.33597	33
29	↓2	wikiliu	0.33558	3
30	↓1	hahaha6	0.33315	21
31	↑3	TooNaive	0.33244	20

Figure 6: Online ranking

5 CONCLUSION

In this paper, the author's name is disambiguated by the way of association graph, and each paper represents a node. By considering the cooperators of the papers and their affiliated organizations, we can determine whether there is association between the two papers. If there is association, we will establish an edge between the two nodes, and finally we will get several connected subgraphs. The number of subgraphs is the number of classes after clustering. Then we will cluster again with Agglomerative algorithm. The results obtained by this method are good.

In this assignment, I gained a lot and learned the method of disambiguating the author's name and the related knowledge of entity connected graph. I have a deeper understanding of classical clustering algorithms, especially Agglomerative algorithm and DBSCAN algorithm. Before this lesson, I almost had no contact with data mining. Through this semester's study and this competition, I felt the fun and importance of data mining.

REFERENCES

[1] C Lee Giles, Hongyuan Zha, and Hui Han. Name disambiguation in author citations using a k-way spectral clustering method. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'05)*, pages 334–343. IEEE, 2005.

[2] Jian Huang, Seyda Ertekin, and C Lee Giles. Efficient name disambiguation for large-scale databases. In *European conference on principles of data mining and knowledge discovery*, pages 536–544. Springer, 2006.

[3] Byung-Won On, Ergin Elmacioglu, Dongwon Lee, Jaewoo Kang, and Jian Pei. Improving grouped-entity resolution using quasi-cliques. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 1008–1015. IEEE, 2006.

[4] Feng Wang, Juanzi Li, Jie Tang, Jing Zhang, and Kehong Wang. Name disambiguation using atomic clusters. In *2008 The Ninth International Conference on Web-Age Information Management*, pages 357–364. IEEE, 2008.