The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016)

Name Disambiguation Method based on Multi-step Clustering

S. Gu, X. Xu, J. Zhu, L. Ji

* College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
b College of Computer Yancheng Big Data Research Institute, Nanjing University of Posts and Telecommunications, Yancheng 224005, China

Abstract

Author name disambiguation is a very important and complex research topic. During the retrieval and research of literatures, the quality of the investigation results has been reduced because of the high probability of different authors sharing the same name, which lengthens the whole cycle of the scientific research. Therefore, it is necessary to find a reasonable and efficient method to distinguish the different authors who share the same name. In this paper, an author name disambiguation method based on multi-step clustering (NDMC) is proposed to disambiguate author names. First, the framework combines the brief and clear characteristics of literature system information with the comparison of co-authors’ similarity to realize the initial clustering. Then, author's information is extracted from the Baidu Encyclopedia, and the semantic similarity of subordinate units is compared, as the basis of identity discrimination in the second step clustering. Finally, after extraction of two step clustering paper keywords in each class cluster, combined into corpus collection, through the characteristics of the semantic comparison, cancellation of indeterminacy results further adjustment, so as to complete the multi-step clustering. We extract literature information from the China National Knowledge Infrastructure (CNKI) to implement experiments. The experimental results show that the hybrid disambiguation framework is feasible and efficient.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Peer-review under responsibility of the Conference Program Chairs

Keywords: Name disambiguation; Feature extraction; semantic recognition; hierarchical clustering

1. Introduction

Authors identification online needs to be addressed. DBLP [1] (Digital Bibliography & Library Project) is first appeared in the author integrated system as the core of literature, which includes almost all computers in the major international journals published in English literature, and the meeting every quarter for a data update, the academic
literature database can be a very good Computer technology, by the user to retrieve the author's name, you can find all the documents to the name of the author's record, but did not do the nuptial disambiguation.

C-DBLP [2] is developed by imitating DBLP, with the author as the core of literature integration system, and based on the co-author relationship characteristics. It has the name disambiguation function with high accuracy. However, its recall rate is relatively low.

In this paper, we focus on the Chinese literature system of the name disambiguation problem. When we retrieve objects to the author, retrieves a lot of the authors and the paper with the same information, if the name is more common, and the same information would be redundant. Often, large literature database will provide advanced retrieval function, provides the function of the restrained, but usually only from two direction constraints of units or journal, its result is just the constraints under the condition of information, can't really do the name disambiguation. Based on this, we combine the characteristics of the information system, the paper brief refining, to provide the user name repetition experts under the true character of comprehensive paper information as the goal, to automatic disambiguation methods in the literature to provide technical support to make better use in the database.

The main contributions of this work include:

- We combine the Baidu Encyclopedia classification information and the co-author information as the foundation, and then combine keyword corpus collection to obtain the higher disambiguation accuracy and recall rate.
- We utilize the retrieve full rate (RFR) according to the actual users to the author for the demand of the retrieval objects, and design the algorithm with people-oriented thoughts.
- We obtain identity recognition unit classification threshold through the experiments and improve the accuracy and recall rate of the algorithm.

The rest of the paper is organized as follows: in Section 2, a model of author and its features are presented. Section 3 presents the results and experiments and performance analysis. Section 4 concludes this paper.

2 Model of NDMC

First of all, according to the cooperator information of the information of the paper, we get different clustering by conducting the first clustering. These clusters are based on the condensed level of clustering idea, and they will only gather more and more instead of being parted again. After the first clustering, there are still many complete information remaining apart. Then, we base on the item information in the Baidu Encyclopedia and recognize multiple identity information under the name. Through the experiment, we will choose the value of the threshold to distinguish the size of clusters in order to choose clusters that are used as molding snow balls to take over other combinations of clusters. Last, basing on the first two steps of clustering, we extract keywords in the paper from all information-focused clusters, and combine them as feature corpus to compare with the rest clusters that relatively include fewer chapters in the similarity of features. So we can finish the final clustering. Detailed procedures are following:

2. 1 Similarity computation
Take $S$ and $T$ as two strings, $S = \{s_1, s_2, \ldots, s_n\}, T = \{t_1, t_2, \ldots, t_n\}$, and $s_i, t_j$ represent the characters in the two strings. Take the similarity level between String $S$ and String $T$ as $sim(S, T)$, therefore,

$$sim(S, T) = 2 \ast \frac{\text{card}(S \cap T)}{(\text{card}(S) + \text{card}(T))}$$

where $\text{card}(s)$ represents the number of elements in the Set $S$.

2. 2 Model

2. 2. 1 Feature disambiguation

There are titles, cooperators, work units and such information. When we deal with the disambiguation, these features have common practicability. We derive paper information using the structure—Title, Names of Cooperators (NC), Work Units of Authors (WUA), Publishers, Keywords, Publishing Time, which we use as the first step of multistep disambiguation. Our aim is to gather the paper which have links on the surface together as much as possible. Therefore, we should base on the cooperator information. Algorithm 1 shows us the clustering algorithm of step 1.

2. 2. 2 Disambiguation of identity based on the Baidu Encyclopedia

The Baidu Encyclopedia is a Chinese-language collaborative web-based encyclopedia provided by the Baidu search engine. As of January 2016, Baidu Encyclopedia has more than 13 million articles. As for editors of personal information, Baidu Encyclopedia classifies duplicate identities using items. If we can reasonably use these items when we conduct Duplication Disambiguation to the document database, it will be more beneficial to identity recognition.

Scholars publish literatures in different work units, it will bring difficulty to automatic disambiguation, because this literature can belong to the same author, or a real another one. We crawl information under the name from the Baidu Encyclopedia. In general, a well-known figure can be marked by Baidu Encyclopedia. For example, after the step 1, we assume that different type of clusters show the scholar from the work unit A and B. If the information from Baidu Encyclopedia includes these work units, it will judge these clusters represent different people; if the information from Baidu Encyclopedia only includes work unit A, we will carry out further work, but not directly merge these clusters, because the same school also has the same name of the scholars.

When the information from Baidu Encyclopedia only includes part of work units, we should judge which cases that different clusters can be merged. We carry on threshold calculation:

$$\text{Sec}_T = \text{MAX}[\text{SMAX}\{A_1, A_2, \ldots, A_n\}_A, \ldots, \text{SMAX}\{N_1, N_2, \ldots, N_n\}_B]$$

where $\text{Sec}_T$ represents divided cluster threshold that has been chosen, $\text{SMAX}\{A_1, A_2, \ldots, A_n\}_A$ represents that a cluster is dominated by Author A, and the chapters included are only next to those included in the largest panel point.

However, based on the accuracy of the identity of the Baidu Encyclopedia criterion, this step relies on the accuracy of classification of the Baidu Encyclopedia, without edited the author information, it maybe affect the result of disambiguation. Algorithm 2 shows the clustering algorithm of step 2.

Algorithm 1

Input:
A list of papers that share the same name
Output
A list of clusters that have the same NC
Initialization:
Create an empty set FinalCluster, read NC,Cluster C,Paper P
Method
Put each Paper P into Cluster C
Computer the similarity of NC between Pi and Pj
If ( similarity of NC= 1)
   Add Pj into Ci and Del Cj
   If ( num of paper in Cj >= 2 )
      If ( the similarity of NC =1&& Cj < Ci )
         Add all papers into Ci and Del Cj
      Else
         Preserve Cj
   Else
      Preserve Cj
Put all the information of clusters into FinalCluster
Return FinalCluster

Algorithm 2

Input:
The result of first step FinalClusters
Output
A list of clusters after the second step
Initialization:
Create two set temp_large and temp_small,Cluster C,Paper P
Method
For each cluster Ci, and Pi ∈ Ci do
   For each cluster Cj, and Pj ∈ Cj do
      If size(Ci) < σ
         Put all papers ∈ Ci into temp_small
      Else
         Put all papers ∈ Ci into temp_large
      Compare workspace with the record from internet
      If sim(WUA)< δ
         New Cluster Ci = Merge(C with the same WUA)
      Else
         Preserve Ci
Put all the information of clusters into tempCluster
Return tempCluster

Algorithm 3

Input:
The result of second step tempCluster
Output
A list of clusters with different author
Initialization:
Create two set temp_large and temp_small,Cluster C,Paper P
Method
For each cluster Ci, and Pi ∈ Ci do
   For each cluster Cj, and Pj ∈ Cj do
      If size(Ci) < V
         Put all papers ∈ Ci into temp_small
      Else
         Put all papers ∈ Ci into temp_large
      Compare workspace with the record from internet
      If sim(WUA)< δ
         New Cluster Ci = Merge(C with the same WUA)
      Else
         Preserve Ci
Put all the information of clusters into tempCluster
Return tempCluster

2.2.3 Disambiguation of key words based on the result of two step clustering

The task of this phase is to improve the accuracy of disambiguation and compensate for the shortcomings of the first two steps. At this time, the panel spots where information is relatively centralized are mainly different authors and we take them as basic spots. As for the scattered information, we begin with its content and use semantic judgment and mustering similarity. What we are facing are problems of the size of the set and the choice of critical features. Such as, after the first two steps, Author A’s information is gathered to four panel spots: C1,C2,C3,C4. Supposed that the chapters in the C4 are largest, then it will be supposed as the fixed point. We extract all critical information from it to make up the feature corpus and Set U. As for C1,C2,C3, we also extract all key words to make up Set V1,V2,V3. Then we should calculate the semantic similarity between the small set and the large set. Next according to the size of the set, we will design the corresponding amalgamative threshold. All the paper’s information in the Set V will be amalgamated into Set U as long as the similarity between Set V and Set U goes beyond the value of threshold. Algorithm 3 shows us the clustering algorithm of step 3.
3. Experiments

3.1 The source of the test data

The test data set built should meet the following conditions:

a) The name of authors that are chosen should be representative, mainly in common name;

b) The number of literature should be representative, because of the difference of authors;

c) Authors may have more than one work unit;

We give enough consideration to the above conditions, and select 1179 literatures randomly from the CNKI2, consisting of six different names.

3.2 Evaluation criteria

This paper adopts a kind of commonly used evaluation methods [4]: Precision (P), Recall rate (R) and F value measure the advantages and disadvantages of the algorithm. All indexes are defined as follows [5]:

\[
P = \frac{CN}{N} \times 100\% \\
R = \frac{CN}{RN} \times 100\% \\
F = \frac{2 \times P \times R}{P + R} \times 100\%
\]

where \(CN\) is the correct identification entity number appears in the actual results, \(N\) is the number of entities of recognition, \(RN\) is the number of entities in the actual results.

According to the user to retrieve core literature database requirements, RFR is calculated with:

\[
RFR_A = \sum_{j=1}^{k} \max(\mu_j / N_{A_{real}})
\]

For example, there are three authors respectively for the output of the results in the following table:

<table>
<thead>
<tr>
<th>Node</th>
<th>Author_name</th>
<th>(\mu_j)</th>
<th>(N_{real})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node_1</td>
<td>Author_A</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Node_2</td>
<td>Author_B</td>
<td>30</td>
<td>45</td>
</tr>
<tr>
<td>Node_3</td>
<td>Author_B</td>
<td>8</td>
<td>45</td>
</tr>
<tr>
<td>Node_4</td>
<td>Author_C</td>
<td>60</td>
<td>70</td>
</tr>
</tbody>
</table>

Therefore, the results: \(RFR_A = 100.00\%\), \(RFR_B = 66.67\%\), \(RFR_C = 85.71\%\)

3.3 Results of experiments

According to the 2.2.2, through sample data training, we can observe from Fig. 1 that in this time' experiment, \(\text{Sec}_T\) is 14, so it will be more efficient if the threshold takes the value of about 15.

We classify 1179 literatures manually, and compare with the experimental results, as shown in Table 2.

\[2\] http://www.cnki.net/
Table 2. The experimental results.

<table>
<thead>
<tr>
<th>Author Name</th>
<th>Literature Number</th>
<th>RN</th>
<th>N</th>
<th>CN</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>XuXiaolong</td>
<td>236</td>
<td>34</td>
<td>37</td>
<td>30</td>
<td>0.81</td>
<td>0.88</td>
<td>0.84</td>
</tr>
<tr>
<td>Chenwei</td>
<td>223</td>
<td>35</td>
<td>37</td>
<td>32</td>
<td>0.86</td>
<td>0.91</td>
<td>0.89</td>
</tr>
<tr>
<td>Lilei</td>
<td>148</td>
<td>64</td>
<td>66</td>
<td>61</td>
<td>0.92</td>
<td>0.95</td>
<td>0.94</td>
</tr>
<tr>
<td>Liulinfen</td>
<td>87</td>
<td>32</td>
<td>29</td>
<td>26</td>
<td>0.89</td>
<td>0.81</td>
<td>0.85</td>
</tr>
<tr>
<td>Liyun</td>
<td>282</td>
<td>54</td>
<td>56</td>
<td>52</td>
<td>0.92</td>
<td>0.96</td>
<td>0.94</td>
</tr>
<tr>
<td>Chenzhi</td>
<td>203</td>
<td>23</td>
<td>25</td>
<td>20</td>
<td>0.80</td>
<td>0.87</td>
<td>0.83</td>
</tr>
<tr>
<td>Average</td>
<td>196</td>
<td>40</td>
<td>42</td>
<td>37</td>
<td>0.86</td>
<td>0.89</td>
<td>0.88</td>
</tr>
</tbody>
</table>

We make statistical analysis of the results after each step. The results are shown in Table 3.

Table 3. Experimental Comparison Results.

<table>
<thead>
<tr>
<th>Step</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step1</td>
<td>0.40</td>
<td>0.75</td>
<td>0.51</td>
</tr>
<tr>
<td>Step2</td>
<td>0.75</td>
<td>0.83</td>
<td>0.79</td>
</tr>
<tr>
<td>Step3</td>
<td>0.86</td>
<td>0.89</td>
<td>0.88</td>
</tr>
</tbody>
</table>

From the perspective on indicators in Table 3, with the increment of feature selection and recognition method, accuracy and recall rate has been significantly improved.

At the same time, we test the RFR to verify that whether nodes contain most of the correct literatures. The results are shown in Table 4.

Table 4. RFR Test.

<table>
<thead>
<tr>
<th>Author name</th>
<th>Real num</th>
<th>&lt;60%</th>
<th>60%~70%</th>
<th>70%~80%</th>
<th>80%~99%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>XuXiaolong</td>
<td>34</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>Chenwei</td>
<td>35</td>
<td></td>
<td>1</td>
<td>1</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>Lilei</td>
<td>64</td>
<td>1</td>
<td>2</td>
<td></td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>Liulinfen</td>
<td>32</td>
<td></td>
<td></td>
<td>2</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>Liyun</td>
<td>54</td>
<td>1</td>
<td></td>
<td>1</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>Chenzhi</td>
<td>23</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>21</td>
</tr>
</tbody>
</table>

From the experimental data set, for example, there are 64 scholars named Li Lei, the complete information of 61 scholars shows in 61 nodes, 70 to 80 per cent of two scholars’ complete information shows in two nodes, less than 60% of one scholar’s complete information shows in the node. Then when users search for Li Lei, they can get correct and complete information of the scholars named Li Lei basically. According to the author in the node number of dominant and the corresponding RFR, as shown in Fig.2.

Next, we compare our proposed NDMC with another disambiguation method based on the fusion of multiple features (DFMF). DFMF is a disambiguation method that calculates the similarity of clusters by the weight of the information elements in the paper, and it is a classical method. Using the same test data, Table 5 shows the results in terms of accuracy. NDMC statistically significantly outperforms DFMF.
From the line chart above (Fig. 2), we can find some regulars. When the literatures are few, there is a high rate of literature searching. The first local minimum point appears in the position of the twelfth literature. The reason may be that the author publish papers independently which causes the less similarity of collaborator among literature. What’s more, with the increase of literature, the RFR will raise gradually. Subsequently, we analyze the data and have some new discoveries. In our test sample, similarity of feature information among these papers is higher in this section of the line chart, especially the similarity of cooperator and work units. As the quantity of papers continues to increase, as for the author, there will appear various types of research area, cooperator and work units inevitably. This phenomenon will result in the reduction of local similarity. Then, when the samples become more and more, the feature corpus will be larger. Under this circumstance, the algorithm of calculating similarity according to semanteme can also play an important role.

4. Conclusion

On the basis of experiment, we put forward new standards and research methods on the application of name disambiguation in literature library. From the experimental results, we can see the method is feasible. To be specific, we measure the standard of algorithm more practical and put forward the idea of RFR to check out the practicality of our algorithm. Moreover, we don’t simply pursue the high precision of the index. Our goal is to achieve that when users are searching for an author, they can retrieve the papers’ information, which are well clustered of all authors under the same name.

Acknowledgements

We would like to thank the reviewers for their detailed comments to help us improve the quality of this paper. This work was jointly sponsored by the National Natural Science Foundation of China under Grant 61472192 and 71401079, and the Natural Science Fund of Higher Education of Jiangsu Province under Grant 14KJB520014.

References


