Relationship-based Entity Resolution for Organization Alias Mining

Jianjun Cao[1], Yuling Shang[2], Qibin Zheng[2], Xing Zhou[2], Yi Liu[2], Hongmei Li[2] and Qin Feng[2]
1. Nanjing Telecommunication Technology Institute
2. PLA University of Science and Technology

Abstract: Aimed at the problem that many organization names may refer to the same organization entity, we propose a relationship-based organization alias mining method to solve it. We first construct a bipartite graph based on the relationship between organization names and author names, then we define the directed similarity and the enhanced similarity of two sets respectively, and we also prove that enhanced similarity is better than Tanimoto distance. Next, we propose a novel method called Set-to-Numerical Data Space Transformation (SNDST) by combining enhanced similarity with the data space transformation method, and give the mining method. In the end, we conduct several experiments to compare SNDST with other methods to show its superiority.

Keywords: Entity Resolution, Organization alias, Relationship Data, Tanimoto Distance, Enhanced Similarity, Data Space Transformation

1 INTRODUCTION

In one or more data sets, an object may have many different representations. The aim of Entity Resolution is to find out different representations of the same object from one or more data sets, and correctly recognize all the different entities (Christen, 2012). Entity Resolution is also called Record Linkage, Duplicate Detection, Deduplication, Data Matching, etc. (Gao & Zhang, 2015; Tan et al, 2014).

The research and application of Entity Resolution have achieved remarkable progresses, especially in Data Warehouse, Data Mining and Information Retrieval. Currently many information technologies are developed on assuming that data are correct, however, there are ubiquitously kinds of data quality problems, which have seriously affected their application. Entity Resolution is one of the most typical technologies for data quality management, and it also plays an important role in improving data quality and data application (Christen, 2012; Cao et al, 2010).

The methods of Entity Resolution can be categorized into three kinds for the different use of information, which are feature based methods, context based methods and relationship based methods. Feature based methods recognizes entities based on the similarity of each attribute of the record, and they are the most basic method; context based methods not only takes the record attributes into account, but also the context attributes or the attributes from the context records; all kinds of relationships among entities are fully utilized to recognize entities in the relationship based methods (Wang et al, 2015).

In this paper, the relationship based methods are studied, information which contains organization names, author names and the affiliation relationship between them are used to mine organization names referring to the same organization entity, since these organization names refer to the same entity, they are also called organization alias, such as "Wuxi light industry university" and "Jiangnan University". According to the used period of organization names, organization alias can be divided into coexist organization alias, rename organization alias, and combined rename organization alias etc. And according to the type of organization, it also can divided into the industry and the research organization, but no matter which type of organization, the organization and their authors(staffs) can be got in different ways, and this data can help to mine organization alias.

The reasons leading to organization alias mainly include: the extension, combination, merger, etc. of academy, university, research institution, etc. leading to the change of organization names, for example, "Donghua University"(1999~now) and "Textile University of China"(1985~1999); and some organization entities for security or other reasons, the same entity may have different organization names inside and outside the organization simultaneously, such as "Design Institute of China Electric Power" and "Beijing Electric Power Research Institute". The mined organization alias of academy, university, research institution, etc. can be used to retrace the history of organization entities and check compliance of the usage of organization names.

In the initial period of our research in Organization Alias Mining, we don't consider the type of organization alias, but our feature researches are about the subdivide of organization alias, which are coexist organization alias, rename organization alias, and combined rename organization alias.

Considering the peculiarity of this issue, we construct the bipartite graph to show the organization names, author names and the relationship between them, and solve this issue by finding the sub-bipartite graph; the effective mining of organization alias is achieved by combining enhanced similarity with the data space transformation method.

The rest of this paper is organized as follows: section 2 constructs of Organization-Author bipartite graph. section 3 gives the similarity of set. section 4 introduces data space transformation. section 5 gives organization alias mining method. section 6 is the experiments. section 7 is conclusion.

2 BIPARTITE GRAPH MODEL

In this section, organization names, author names and the relationship between them in information are used to mine organization alias. Based on the relationship between organization names and author names, Organization-Author bipartite graph $G=(V, E)$ is constructed, where V is the set of vertexes, E is the set of edges. $A=\{a_k \mid k=1, 2, \ldots, m\}$ is the set of author names, a_k is the k-th author name; $O=\{o_i \mid i=1, 2, \ldots,n\}$ is the set of organization names, and o_i is the i-th organization name; and $A \cup O=V$, $A \cap O=\varnothing$, $m=|A|$, $n=|O|$, $|A|$ and $|O|$ are the size of sets A, O respectively; $E=\{<o_i, a_k>| \ o_i \in O \wedge \ a_k \in A \wedge p(o_i, a_k)\}$, where $p(o_i, a_k)$ represents the existing relationship between o_i and a_k, i.e. author a_k is affiliated with o_i. The Organization-Author bipartite graph is shown in **Figure 1**.

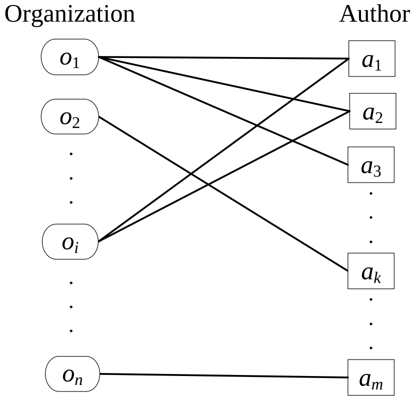


Figure 1: The Organization-Author Bipartite Graph

From the analysis above, we find that in case of different situations, the same author's papers may use different affiliated organization names. According to the different results of recognization, Entity Resolution can be categorized into pairwise comparison method and group comparison method (Tan et al, 2014). Since the mining of organization alias can be classified to be organization alias and others, the pairwise comparison method is suitable for organization alias mining. What's more, mining organization alias from information is equivalent to find subgraph from the Organization-Author bipartite graph where a set of author names affiliated with two different organization names, i.e. these two organization names are organization alias, and they correspond to the same organization entity. In **Figure 1**, if {a_1, a_2, a_3} is the author names set of o_1 and o_i, then the subgraph which satisfies the above conditions is shown in **Figure 2**.

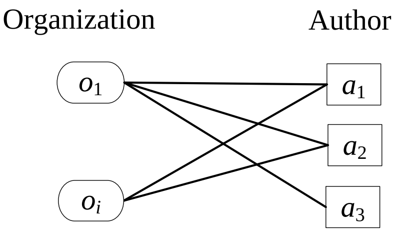


Figure 2: The Subgraph of the Same Author Names Set of Two Different Organization Names

$\forall < o_i, o_j >$, $1 \le i \le n, 1 \le j \le n$, if o_i, o_j are organization alias, they must have some common author names. A_i and A_j are two author names sets affiliated with the same organization names o_i and o_j respectively, then the similarity between o_i and o_j is equivalent to the similarity of sets A_i and A_j. That is

$$s_{ij}(o_i, o_j) = s_{ij}(A_i, A_j) \tag{1}$$

where $0 \le s_{ij} \le 1$, o_i, o_j are organization alias if and only if $s_{ij}(A_i, A_j) > \delta$, δ is the similarity threshold, and $0 < \delta < 1$。

3 Measurement of Set Similarity

In this section, enhanced similarity is proposed to better measure similarity between each two organization names based on their author names sets.

Definition 3.1 (Directed Similarity). X_1 and X_2 are two non-empty sets. Then

The directed similarity of $X_1 \to X_2$ is

$$D(X_1, X_2) = \frac{|X_1 \cap X_2|}{|X_2|} \tag{2}$$

The directed similarity of $X_2 \to X_1$ is

$$D(X_2, X_1) = \frac{|X_1 \cap X_2|}{|X_1|} \tag{3}$$

Definition 3.2 (Enhanced Similarity). X_1 and X_2 are two non-empty sets, $D(X_1, X_2)$ and $D(X_2, X_1)$ are the directed similarities of them, then the enhanced similarity of X_1 and X_2 is

$$\begin{aligned} E(X_1, X_2) = E(X_2, X_1) &= \max\{D(X_1, X_2), D(X_2, X_1)\} \\ &= \max\left\{\frac{|X_1 \cap X_2|}{|X_2|}, \frac{|X_1 \cap X_2|}{|X_1|}\right\} \\ &= \frac{|X_1 \cap X_2|}{\min\{|X_1|, |X_2|\}} \end{aligned}$$

(4)

Tanimoto distance is one of the most common methods to measure the similarity of two sets, and the definition of Tanimoto distance is shown in definition 3.

Definition 3.3 (Tanimoto Distance). X_1 and X_2 are two non-empty sets, then the Tanimoto distance between X_1 and X_2 is

$$T(X_1, X_2) = 1 - \frac{|X_1| + |X_2| - 2|X_1 \cap X_2|}{|X_1| + |X_2| - |X_1 \cap X_2|} \tag{5}$$

Theorem 3.1 X_1 and X_2 are two non-empty sets, $E(X_1, X_2)$ and $T(X_1, X_2)$ are the enhanced similarity and Tanimoto distance of X_1 and X_2 respectively, then $E(X_1, X_2) \geq T(X_1, X_2)$.

Proof: $E(X_1, X_2)$ is the enhanced similarity of X_1 and X_2, and $T(X_1, X_2)$ is the Tanimoto distance between X_1 and X_2.

$\because |X_1 \cap X_2| \leq |X_1| \quad \therefore 0 \leq |X_1| - |X_1 \cap X_2| \qquad \therefore |X_2| \leq |X_1| + |X_2| - |X_1 \cap X_2|$

Similarly, $|X_1| \leq |X_1| + |X_2| - |X_1 \cap X_2|$

$\therefore \min\{|X_1|, |X_2|\} \leq |X_1| + |X_2| - |X_1 \cap X_2| \qquad \therefore \frac{1}{\min\{|X_1|, |X_2|\}} \geq \frac{1}{|X_1| + |X_2| - |X_1 \cap X_2|}$

And $\because |X_1 \cap X_2| \geq 0$

$$\begin{aligned}\therefore E(X_1, X_2) &= \frac{|X_1 \cap X_2|}{\min\{|X_1|, |X_2|\}} \geq \frac{|X_1 \cap X_2|}{|X_1| + |X_2| - |X_1 \cap X_2|} \\ &= \frac{|X_1| + |X_2| - |X_1 \cap X_2| - (|X_1| + |X_2| - 2|X_1 \cap X_2|)}{|X_1| + |X_2| - |X_1 \cap X_2|} \\ &= 1 - \frac{|X_1| + |X_2| - 2|X_1 \cap X_2|}{|X_1| + |X_2| - |X_1 \cap X_2|} = T(X_1, X_2)\end{aligned}$$

That is $E(X_1, X_2) \geq T(X_1, X_2)$.

For two non-empty sets X_1 and X_2, the enhanced similarity in equation (4) is more suitable for measuring their similarity than the Tanimoto distance in equation (5). And our experiments in section 6 will show that enhanced similarity has a better effectiveness than Tanimoto distance.

4 DATA SPACE TRANSFORMATION

If the number of an attribute's value can be listed, we called it nominal data. Paper (Qian et al, 2015) proposed a Nominal-to-Numerical Data Space Transformation (NNDST) method, and after the transformation from nominal data to its data space, it achieves a better classification effectiveness than used the nominal data. Since the organization can be represented by its author set, and the set data which is similarity to nominal data but it is unordered. When the set data is used to mine organization alias, has a lower discrimination and a worse mining effectiveness. To improve the similarity of set and the effect of organization alias mining, data space transformation method is introduced. The description of NNDST is shown as follows.

Table 1 shows a set which has *n* ($n \in Z^+$) records, and each record has *m* ($m \in Z^+$) nominal attributes.

In **Table 1**, *X*={ $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, …, $\boldsymbol{x}_i$, …, $\boldsymbol{x}_n$} is the nominal data set with *n* records, and *B*={ b_1, b_2, …, b_k, …, b_m} is its attributes set, x_{ik} ($1 \leq i \leq n, 1 \leq k \leq m$) is the value of the *k*-th attribute b_k of *i*-th record $\boldsymbol{x}_i$. Then the probability of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ to be the same entity is

$$p_{ij} = \frac{\sum_{k=1}^{m} \theta_k(\boldsymbol{x}_i, \boldsymbol{x}_j)}{m} \tag{6}$$

Nominal Attribute Records						
Record	b_1	b_2	…	b_k	…	b_m
$\boldsymbol{x}_1$	x_{11}	x_{12}	…	x_{1k}	…	x_{1m}
$\boldsymbol{x}_2$	x_{21}	x_{22}	…	x_{2k}	…	x_{2m}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\boldsymbol{x}_i$	x_{i1}	x_{i2}	…	x_{ik}	…	x_{im}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\boldsymbol{x}_n$	x_{n1}	x_{n2}	…	x_{nk}	…	x_{nm}

Table 1: Nominal Attribute Records

where $\theta_k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is the equation to decide whether the *k*-th attribute of records $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ equivalence or not, and it is defined as

$$\theta_k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \begin{cases} 1, & x_{ik} = x_{jk} \\ 0, & x_{ik} \neq x_{jk} \end{cases} \tag{7}$$

As to the records in **Table 1**, after calculation using equations(6) and (7), we can get the numerical matrix $p = \left[p_{ij} \right]_{n \times n}$, then the records in **Table 1** can be transformed to its data space, as shown in **Table 2**.

The Numerical Matrix						
Record	c_1	c_2	…	c_k	…	c_n
$\boldsymbol{x}_1$'	p_{11}	p_{12}	…	p_{1k}	…	p_{1n}
$\boldsymbol{x}_2$'	p_{21}	p_{22}	…	p_{2k}	…	p_{2n}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\boldsymbol{x}_i$'	p_{i1}	p_{i2}	…	p_{ik}	…	p_{1n}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\boldsymbol{x}_n$'	p_{n1}	p_{n2}	…	p_{nk}	…	p_{nn}

Table 2: The Numerical Matrix

In **Table 2**, $c = \{c_1, c_2, \cdots, c_k, \cdots, c_n\}\,(1 \le k \le n)$ is the new attribute set of the record, and $\boldsymbol{x}$'={ $\boldsymbol{x}_1$', $\boldsymbol{x}_2$', …, $\boldsymbol{x}_k$', …, $\boldsymbol{x}_n$'} is the transformed record set; $\boldsymbol{x}$' is used instead of $\boldsymbol{x}$ for classification, and it achieves a better effectiveness.

Based on the proposed method NNDST in paper (Qian et al, 2015), enhanced similarity is proposed to measure the similarity of sets, and a novel method Set-to-Numerical Data Space Transformation method (SNDST) is proposed. SNDST is an effective similarity measurement of sets, and it has a better classification effectiveness in organization alias mining.

5 ORGANIZATION ALIAS MINING METHOD

The steps of organization alias mining are:
a) Based on the relationship between author names and organization names, construct a bipartite graph as shown in **Figure 1**.
b) In **Figure 1**, get the corresponding author names sets of organization names $o_1, o_2, \cdots, o_i, \cdots, o_n$ are $A_1, A_2, \cdots, A_i, \cdots, A_n, 1 \le i \le n, i \in Z^+$.
c) According to equation (4), the enhanced similarity of each two different organization names $o_i, o_j\ (i \ne j)$ is e_{ij}, since their corresponding author names sets A_i, A_j are known, then the enhanced similarity matrix $\boldsymbol{e}$ is

$$\boldsymbol{e} = [e_{ij}]_{n \times n} = [E(A_i, A_j)]_{n \times n} \tag{8}$$

And $\boldsymbol{e}_i$ also is the numerical matrix of the transformed set data.
d) Cosine similarity is used to calculate the similarity of $o_i, o_j\ (i \ne j)$ on their *i*-th and *j*-th row vector of *e*, that are $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$, and cosine similarity between them is

$$s_{ij}(o_i, o_j) = \cos\left\langle \boldsymbol{e}_i', \boldsymbol{e}_j' \right\rangle = \frac{(e_{i1}, e_{i2}, \cdots, e_{in}) \cdot (e_{j1}, e_{j2}, \cdots, e_{jn})}{\left|(e_{i1}, e_{i2}, \cdots, e_{in})\right|\left|(e_{j1}, e_{j2}, \cdots, e_{jn})\right|} \tag{9}$$

e) o_i and o_j are organization alias if and only if $s_{ij}(o_i, o_j) > \delta$, δ is the similarity threshold, and $0 < \delta < 1$.

6 EXPERIMENTS

Experiments are set to validate the effectiveness of SNDST, and the superior of Enhanced Similarity by comparing them with other methods (shown in section 6.2). Precision, Recall and F-measure are used to measure the effectiveness of methods, and fisher discrimination rate and the rate of infra-class and inter-class distance are used to measure the discrimination of classes, which are given in section 6.3. Section 6.4 is the results and analyses of experiments.

6.1 Data Preparation

The data of 203 organization names is exported from CNKI[1]. And this data has 82484 items in total about organization names, author names and the relationship between them, including 63676 author names, 203 organization names, and in which there are 28 pairs of organization alias as shown in **Table 3**, which contains 37118 author names and 50 organization names.

Organization Alias in Experiment Data Set		
No.	o_i	o_j
1	Mechanical & Electrical College of Anhui Polytechnic University	Anhui Institute of Information Technology
2	Anqing Normal University	Anqing Normal College
3	Donghua University	East China Institute of Textile Science and Technology

[1] Http://epub.cnki.net.

4	Donghua University	Textile University of China
5	East China Institute of Textile Science and Technology	Textile University of China
6	Guangdong Pharmaceutical University	Guangdong College of Pharmacy
7	Guangdong Medical University	Guangdong Medical College
8	Harbin Shipbuilding Engineering Institute	Harbin Engineering University
9	Hebei GEO University	Shijiazhang University of Economics
10	Henan University of Finance and Economics	Henan University of Economics and Low
11	Wanfang Institute of Science and Technology of Henan Polytechnic University	Zhengzhou Technology and Business University
12	Henan University of Traditional Chinese Medicine	Henan University of Chinese Medicine
13	Hubei Normal University	Hubei Normal University (College)
14	Jiangnan University	Wuxi Light Industry University
15	Jinzhou Medical University	Liaoning Medical University
16	Jingdezhen Geramic Institute	Jingdezhen Geramic Institute (College)
17	Nanjing Communications Institute of Technology	Nanjing Communications Institute
18	Shanghai Institute of Technology	Shanghai Institute of Technology (College)
19	Shenyang University of Chemical Technology	Shenyang Institute of Science and Technology
20	Suzhou University of Science and Technology	Suzhou University(College) of Science and Technology
21	City College Wenzhou University	Wenzhou Business College
22	Xuzhou Medical University	Xuzhou Medical College
23	Zhejiang Ocean University	Zhejiang Ocean College
24	China Electric Power Design Institute	Beijing Institute of Power Grid
25	China Jiliang University	China Jiliang College
26	Henan University of Technology	Zhengzhou Institute of Technology
27	Henan University of Technology	Zhengzhou Grain College
28	Zhengzhou Institute of Technology	Zhengzhou Grain College

Table 3: Organization Alias in Experiment Data Set

6.2 Methods of Experiments

To validate the effectiveness of SNDST, we compare it with other methods:

Method 1: The Tanimoto distance $T(A_i, A_j)$ of each two different organization names o_i and o_j $(i \neq j)$ is calculated, and threshold δ_1 $(0 < \delta_1 < 1)$ is set. If $T(A_i, A_j) > \delta_1$, o_i and o_j are organization alias.

Method 2: The enhanced similarity e_{ij} of each two different organization names o_i and o_j $(i \neq j)$ is calculated, and threshold δ_2 $(0 < \delta_2 < 1)$ is set. If $e_{ij} > \delta_2$, o_i and o_j are organization alias.

Method 3: The method SNDST proposed in section 5.

6.3 Results Evaluation Indicators

Organization alias mining is a binary classification problem, since the data set of organization names, author names and the relationship between them can be classified into two classes: the organization alias and the others. In this section, fisher discrimination rate and the rate of infra-class and inter-class distance are used to measure the discrimination of these two classes (Hu et al., 2006; Yang et al, 2004).

The fisher discrimination rate of binary problem is

$$Fisher = \frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}} \tag{10}$$

In equation (10), μ_1, μ_2 are mean values and σ_1^2, σ_2^2 are variances of the 1-th class and the 2-th class, respectively.

The bigger *Fisher* indicates the better discrimination.

The rate of infra-class and inter-class distance of binary problem is

$$OI = \frac{|d_1 - d_2|}{d_1 + d_2} \tag{11}$$

where $d_i = \frac{\sum_{t=1}^{n-1} \sum_{k=(i+1)}^{n} |s_{it} - s_{ik}|}{n(n-1)}$ is the mean distance of the *i*-th class and *i*={1, 2}, *n* is the number of the *i*-th class, s_{it} and s_{ik} are the *t*-th value and the *k*-th value of the *i*-th class respectively.

The bigger *OI* indicates the better discrimination.

Precision (*P*) is the rate of real organization alias pairs and the checked set of organization alias pairs, Recall (*R*) is the rate of real organization alias pairs and the set of the whole organization alias pairs, and *F*-measure is the harmonic mean of Precision and Recall. These three indicators are used to measure the effectiveness of organization alias mining, which are used to measure the detected approximately duplicate records in (Mong et al, 2000). And the formula of *P*, *R* and *F*-measure are

$$P = \frac{|N_1 \cap N_2|}{|N_2|} \times 100\% \tag{12}$$

$$R = \frac{|N_1 \cap N_2|}{|N_1|} \times 100\% \tag{13}$$

$$F\text{-measure} = \frac{2 \times P \times R}{P + R} \times 100\% \tag{14}$$

where N_1 is the set of the whole organization alias pairs, N_2 is the checked set of organization alias pairs, and $N_1 \cap N_2$ is the real organization alias pairs set in N_2.

Usually, the bigger *F*-measure is, the better effectiveness of organization alias mining has.

6.4 Results Analysis

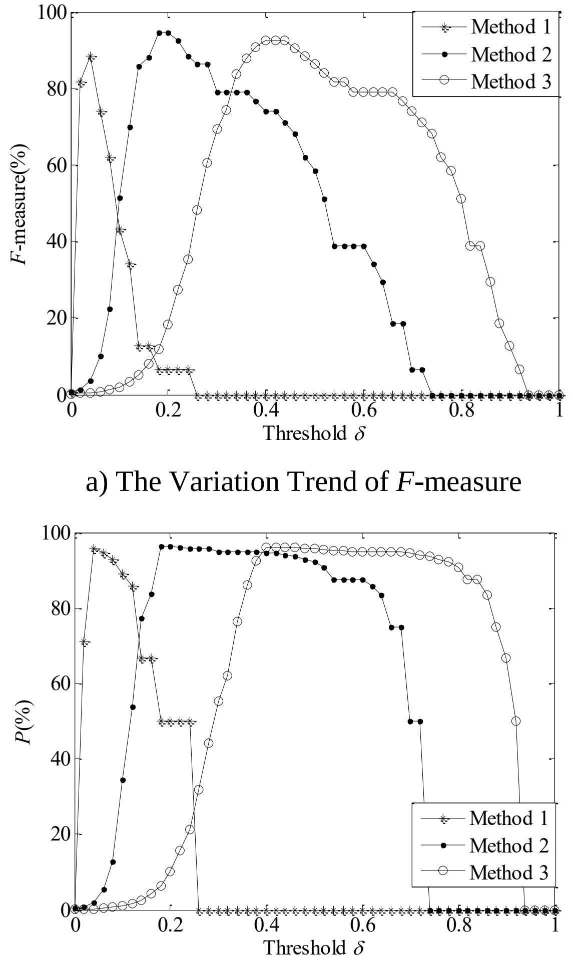
Organization alias mining is a binary classification problem, and the organization pairs can be classified by their similarity into organization alias and others. Then according to the similarity in these two classes, fisher discrimination rate *Fisher* and the rate of infra-class and inter-class distance *OI* of the methods in section 6.2 can calculate by the equation (8) and (9), and **Table 4** is the results.

Discrimination Indexes Value of the Methods Above			
Method	Method 1	Method 2	Method 3
Fisher	1.6243	2.5768	3.4821
OI	2.8381	4.0606	5.1839

Table 4: Discrimination Indexes Value of the Methods Above

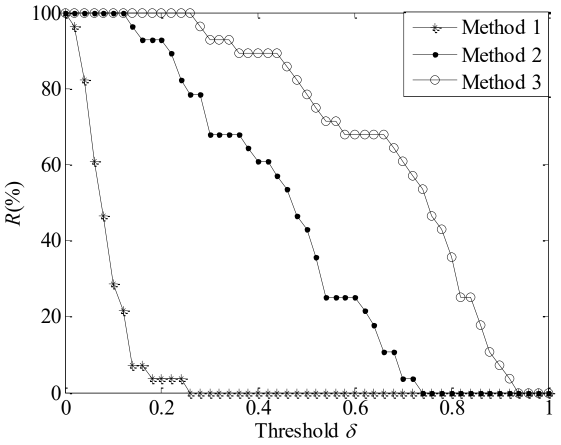
From **Table 4**, we know that a) both *Fisher* and *OI* of Method 1 are smaller than those of Method 2, which means enhanced similarity can enlarge the discrimination of classes, and it is more suitable for organization alias mining; b) both *Fisher* and *OI* of Method 2 are smaller than Method 3, which means the data space transformation also can enlarge the discrimination; c) Method 3 has the biggest *Fisher* (*Fisher* = 3.4821) and *OI* (*OI*=5.1839), which means it has the best discrimination.

For organization alias pairs can be classified by their similarity into organization alias and others. Aimed at the mining of organization alias, data in section 6.1 is used to test the effectiveness of methods in section 6.2. Firstly, we set different thresholds and compare the similarity with them. Secondly Precision *P*, Recall *R* and *F-measure* are calculated according to the compared results. Then the changing curves of Precision *P*, Recall *R* and *F*-measure for different thresholds are shown in **Figure 3**.



a) The Variation Trend of *F*-measure

b) The Variation Trend of Precision



c) The Variation Trend of Recall

Figure 3: The Variation Trend about threshold of the methods

Precision (*P*) is the reflected of correct checked out of organization alias, the bigger *P* indicates the better precision; Recall (*R*) is the checked out of the real organization alias pairs' reflected, and the bigger *R* indicates the better checked out of real organization alias pairs; and *F*-measure is the balance reflect of *P* and *R*. From **Figure 3**, we can see that:

a) Method 1 has the worst effectiveness among all these three methods, and its variation trend is very sharp, meaning that it is much more sensitive to threshold than the other two methods;

b) Method 2 is much better than Method 1, and its variation trend is gentler than that of Method 1, i.e. enhanced similarity is suitable for organization alias mining;

c) Method 3 has the best effective among all these three methods, i.e. data space transformation is useful for organization alias mining. Method 3 not only has the best curve of all the methods, but also the most gently variation trend of all, i.e. Method 3 is not sensitivity to threshold;

d) Consider similarity, the bigger the similarity of two different organization names is, the more reasonable they are regarded as organization alias; since the thresholds of both Method 1 and Method 2 are very small, the reasonability of them needs further discussing;

To sum up, the discrimination of the organization alias of SNDST is clear, its results are not sensitive to threshold, and the precision and recall are quite high, so its threshold can be set in a fixed range. The certain value of threshold can be set according to their trend of precision and recall. The method proposed in this paper is suitable for all data sets with the relationship between organization names and author names, and it is more general for its threshold do not need experts to set or users to debug.

7 CONCLUSION

The data set of organization names, author names and the relationship between them in information were used to construct Organization-Author bipartite graph, it realized organization alias mining by defining enhanced similarity and importing data space transformation. Our works include:

a) Bipartite graph was an effective description of the relationship between organization names and author names. And it was the realization of organization alias mining;

b) The enhanced similarity proposed was a better measurement of two sets, and it was more beneficial for the organization alias mining;

c) We extended the use of data space transformation method from nominal data to set data, achieving the transformation from set to numerical data; SNDST also enlarged the application range of data space transformation, and realized effective mining of organization alias.

However, our method may not work well when a small organization with only several different authors published paper, and unfortunately some of them have same name with an bigger organization.

The method proposed is very important to get information from information, and it also played an important role in analyzing the change of organization names.

ACKNOWLEDGEMENTS

This work was supported by the Natural Science Foundation of China under Grant 61371196, the China Postdoctoral Science Foundation under No. 2015M582832, the Jiangsu Post-doctor Research Fund of China under No. 1402138C, the China Postdoctoral Science Foundation under No. 20090461425, the China Postdoctoral Science Special Foundation under No. 201003797.

REFERENCES

Gao G., Zhang Z. (2015). “Survey on Entity Resolution over Relational Databases”. *New Technology of Library and Information Service*(7:8). pp. 37-47.

Tan M., Diao X., Cao J. (2014). “Survey on Entity Resolution”. Computer Science (41:4). pp. 9-12, 20.

Wang H., Fan W. 2011. “Object Identification on Complex Data”. *Chinese Journal of Computers* (34:10). pp. 1843-1852.

Christen P. (2012), “Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection”. New York, USA, *Springer Science & Business Media*.

Cao J., Diao X., Wang T., et al. (2010). “Research on Domain-independent Data Ceaning: A Survey”. *Computer Science*(37:5). pp. 26-29.

Wang H., Khoshgoftaar T.M., Seliya N. (2015). “On the Stability of Feature Selection Methods in Software Quality Prediction: An Empirical Investigation”. *International Journal of Software Engineering and Knowledge*(25:9). pp. 1467-1490.

Qian Y., Li F., Liang J., et al. (2015). “Space Structure and Clustering of Categorical Data”. *IEEE Transactions on neural networks & learning systems* (27:10). pp. 1-13.

Mong L.L., Tok W.L., Wai L.L. (2000). “IntelliClean: A Knowledge-Based Intelligent Data Cleaner”. *Knowledge Discovery & Data Mining*. pp. 290-294.

Hu Q., He Z., Zhang Z., et al. (2006). “Intelligent Diagnosis for Incioient Fault Based on Lifting Wavelet Package Transform and Support Vector Machines Ensemble”. *Chinese Journal of mechanical engineering* (42:8). pp. 16-22.

Yang B.S., Han T., An J.L. (2004). “ART-KOHONEN Neural Network for Fault Diagnosis of Rotating Machinery”. *Machanical Systems and Signal Processing* (18:3). pp. 645-657.