Journal of Software Engineering and Applications, **2012, 5, 29-32**
http://dx.doi.org/10.4236/jsea.2012.51005 Published Online January 2012 (http://www.SciRP.org/journal/jsea)

Name Disambiguation Method Based on Attribute Match and Link Analysis*

Yu-Feng Yao

Changshu Institute of Technology, Computer Science and Engineering College, Suzhou, China.
Email: fengsingal81@gmail.com

Received October 30th, 2011; revised November 31st, 2011; accepted December 13th, 2011

ABSTRACT

A name disambiguation method is proposed based on attribute match and link analysis applying in the field of insurance. Aiming at the former name disambiguation methods such as text clustering method needs to be considered in a lot of useless words, a new name disambiguation method is advanced. Firstly, the same attribute matching is applied, merging the identity of a successful match, secondly, the link analysis is used, structural analysis of customers network is analyzed, Finally, the same cooperating information is merged. Experiment results show that the proposed method can realize name disambiguation successfully.

Keywords: Name Disambiguation; Data Mining; Attribute Match; Link Analysis

1. Introduction

Data Mining [1] is searching large amounts of data, reveals the hidden laws in it, and further models it to the advanced and effective method according to the established business objectives. It is the process of extracting the potentially useful information and knowledge in a lot of not complete, noisy, fuzzy, random data. Name disambiguation is to divide the different people using the same name to represented according the context or chapter information. The current systems can manage name disambiguation are such as SnakeT, Vivisimo and Apex, etc., the drawback is only the name of people is managed as an ordinary word. The clustering result labels are the vocabulary related with people name, can not distinguish the disambiguation results. The literature [2] proposed the character name disambiguation method based on social network. Bollegala D *et al*. [3] achieve clustering for the final result by calculating the vector similarity. The above methods [2-5] all considered a lot of useless words, and the digestion process has the strong reliance with the information extraction.

A new name disambiguation method is proposed in this paper, the method concludes three procedures: Firstly, the attribute match can be done, the attribute math algorism is used to calculate the similarity between the customers; then the link analysis is followed, using the relationships between the customers to recognize the customer disambiguation problem. The experiment shows the method is this paper can optimize the match processing of the different people having the same name.

*Supported by the research startup fund for young teachers of the Changshu Institute of Technology (Grant No. QZ0912).

2. The Respective Work

The original data of Field insurance records the contract number, insured name, age of the insured, the insured employment, insured unit, occupation, age and other information. The following are the definitions of several variables:

Identifier: The cooperative network node extracted from the original nodes data, is the carrier of customer information, with the letter "r" to represent;

Entities: The true customer related with the identifier, and it construct the relation of one to many, using the letter "e" to represent;

Attribute: Describe the author attribute such as address, profession etc. using the character "a" to represent, $r_i \cdot a_j$ represent the attribute of identifier a_j; $r_i \equiv r_j$ represent r_i and r_j are the identifier of the same entity.

Relation: The relations between the two entities are called the relation, relation is a wide notion.

3. Name Disambiguation Method

3.1. Name Disambiguation Flow

Name disambiguation procedure can be mainly concluded as three steps: the first step is attribute match, to match the same identifier customer information, and combine the attribute with the successful matched identifier. The

Second step is the link analysis, through the structural analysis for the customer cooperative net to find the customers with the same identifier is cooperated with the same agent. entity, therefore, the customers having the same identifier is the same entity, and combine and analyze the entity with the same cooperative information. The steps are showed as **Figure 1**.

3.2. Attribute Match

For the customers having insured in some insurance company, the attribute has the name, customer name, purchased product name, and the customer information. Firstly, divide the two attributes c of the input entity, and match the first part of it showed as Equation (1):

$$\text{SubStringMatch}\left(r_i \cdot \text{Attr} \cdot \text{FirstPart}, r_j \cdot \text{Attr} \cdot \text{FirstPart}, R_l\right) \tag{1}$$

In Equation (1), R_l is the threshold value between 0 and 1, given the relative short string s in c, the relative long character string is l, and given the maximum long string of l in s is m, so the Equation (2) can be concluded.

$$\text{SubStringMatch}=\begin{cases}\text{true}, s\cdot LENGTH \le m\cdot LENGTH \times R_l \\ \text{false}, s\cdot LENGTH > m\cdot LENGTH \times R_l\end{cases} \tag{2}$$

When the Equation (1) returns value is true, then the second part is matched by Equation (3):

$$\text{CountMatch}\left(r_i \cdot \text{Attr}, r_j \cdot \text{Attr}, R_c\right) \tag{3}$$

In Equation (3), R_c is also the threshold between 0 and 1, given s is the relative short string between r_i.Attr and r_j.Attr, and the long string is l, and given the appearance times of *s* in string l is n, so the Equation (4) can be conclude:

$$\text{CountMatch} = \begin{cases}\text{true}, n \ge s.LENGTH \times R_c \\ \text{false}, n < s.LENGTH \times R_c\end{cases} \tag{4}$$

The attribute match function showed as Equation (5) can be obtained by combined Equations (1) and (3):

$$\begin{aligned}&\text{Match}_2(r_i, r_j, R_l, R_c) = \\ &\text{SubStringMatch} \\ &\quad\left(r_i.DEPT.\text{FirstPart}, r_j \cdot DEPT \cdot \text{FirstPart}, R_l\right) \\ &\wedge\text{CountMatch}\left(r_i \cdot DEPT, r_j \cdot DEPT, R_c\right)\end{aligned} \tag{5}$$

Considering the below case: Identifier A is the same with identifier B, and they have the insurance cooperative relation with the identifier C, then identifier A and the identifier B correspond with the same customer entity, showed as the Equation (6):

$$\begin{aligned}&\text{If } \exists r_x, \left(r_i, r_x\right) \in S_c \wedge \left(r_j, r_x\right) \in S_c \\ &\wedge r_i \cdot NAME = r_j \cdot NAME, \text{then } r_i \equiv r_j.\end{aligned} \tag{6}$$

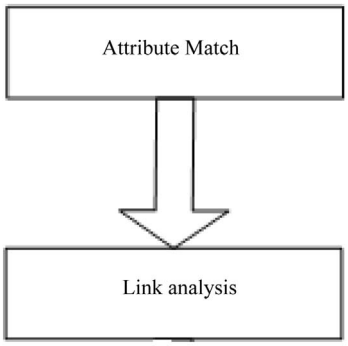

Figure 1. Name disambiguation flow.

4. Simulation Experiment

4.1. Experiment Parameters

The data set used in the experiment is from the core business system of some big insurance company, mainly contains integrated business manage system, universal system and pension system etc. The insured information contains customer number, social insurance number, certificate type, certificate number, occupation categories, subcategories, small categories, health state, smoke years, the smoked number, marriage status, the relationship with the insured; the information of the insured contains customer number, certificate type, certificate number and name etc. Insurance information contains insured number, sign date, effective date, paid premium and policy duration. Insurance category contains insurance code, insurance name, insurance type, duration and the agent-related information. The data set is showed as **Figure 2**.

4.2. Attribute Match Simulation

It is a match experiment showed in **Table 1** after applied Equation (5) in the dataset.

4.3. Link Analysis Simulation

Using the node to represent the insurance relation person such as insurer, insured, and agent, the node size represents the degree of the node, the edge of connecting node represents the cooperative relations between authors. The customer name is labeled at the edge. For the researched sample object, the unit information of the customer can be labeled. For the link analysis, the perfect manage result is: the same name customer not in the same cooperative customer entities are different customer entities, and the same name author in one cooperative entities are attribute to the customer entity. Before the link analysis simulation, the net of three people called Zhao aixiang is showed as **Figure 3**, and the resulted result after link analysis is showed as **Figure 4**.

For the distance between two same name authors called

Table 1. Attribute match.

	Id	Gender	Name	birthday	age	Body state	Order date
Before Match	1315	F	Wang Si	1953.10.23	55	1	2008-5-7
	1328	F	Wang Si	1953.10.23	55	0	2008-5-7
After Match	1328	F	Wang Si	1953.10.23	55	rep	2008-5-7

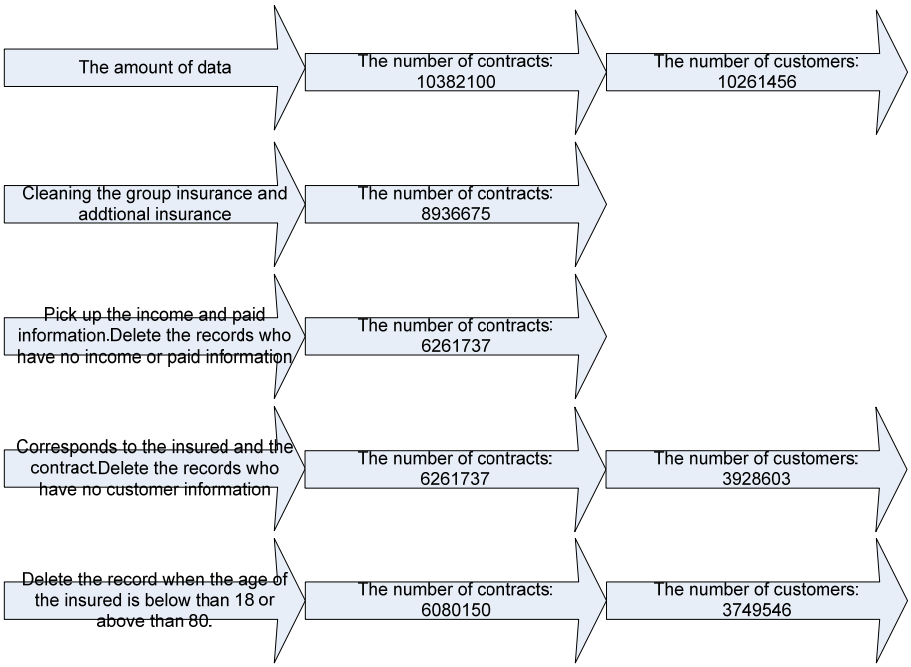

Figure 2. Data set of some insurance company.

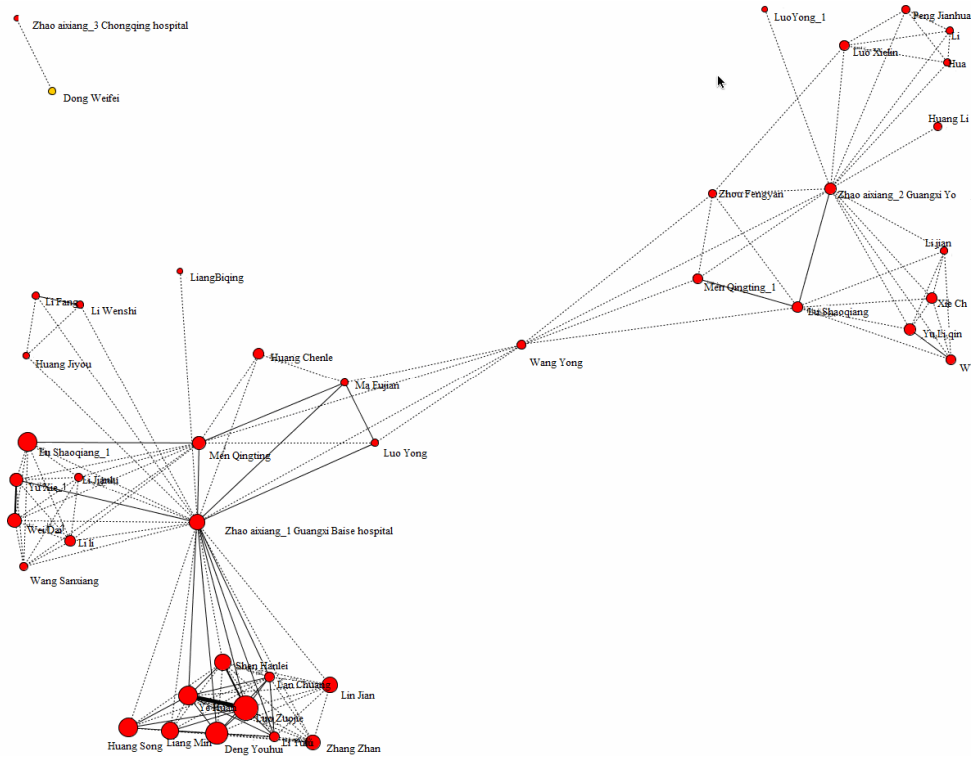

Figure 3. Customer net of "Zhao aixiang" before link analysis.

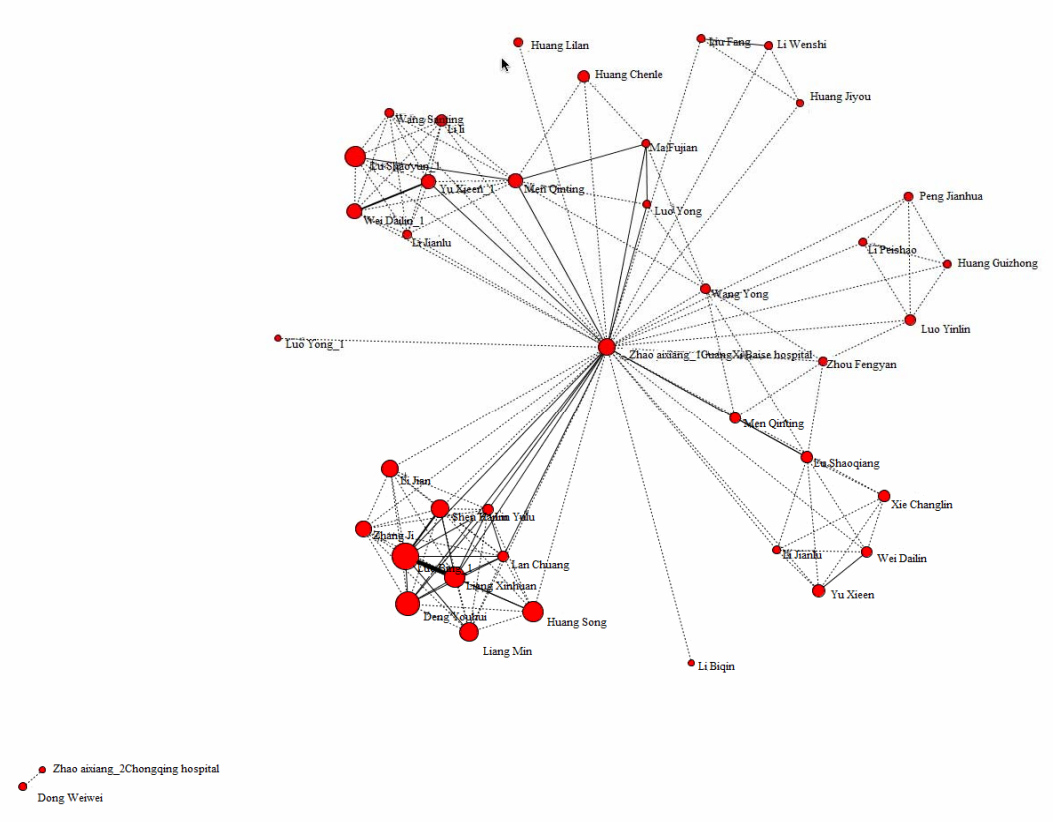

Figure 4. Customer net of "Zhao aixiang" after link analysis.

"Zhao aixiang" is 2, so they are regarded as the same customer entity, remove "Zhao aixiang" in the southweset of Guangxi youjiang medical College hospital, and transfer her the cooperative relation to "Zhao aixiang" 1in Baise people's hospital. From **Figure 3** we can find the original cooperative people of "Zhao aixiang" 2 such as "Lu shaoqiang" is showed have the cooperative relation with "Zhao aixiang" 1.

5. Conclusion

The disambiguation analysis method is proposed in this paper, it can be divided to three steps: Firstly, the attribute match is finished, then using the attribute match algorism to calculate the similarity between same name customers; then by the relations of the customers to recognize the name disambiguation by link analysis; finally, improve the traditional cluster algorism by using the atom cluster. From the compared experiment, the method showed in this paper can solve the name disambiguation, and the executing efficiency can satisfy the practical demands. The method proposed in this paper has successfully applied in some insurance company, the next work is to consider improving the algorism efficiency, and considering extracting the other named entity and combining with the idea of text cluster to realize the name disambiguation of customers.

REFERENCES

[1] M. Q. Yu, "Research on Knowledge Mining in Person Tracking," Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 2006.

[2] J. Lang, B. Qin, W. Song, *et al*., "Person Name Disambiguation of Searching Results Using Social Network," *Chinese Journal of Computers*, Vol. 32, No. 7, 2009, pp. 1365-1374. doi:10.3724/SP.J.1016.2009.01365

[3] D. Bollegala, Y. Matsuo and M. Ishizuka, "Disambiguating Personal Names on the Web Using Automatically Extracted Key Phrases," *Proceedings of the* 17*th European Conference on Artificial Intelligence*, IOS Press, Riva del Garda, 2011, pp. 553-557.

[4] H. F. Wang, "Cross-Document Transliterated Personal Name Coreference Resolution," Springer, New York, Vol. 21, 2010, pp. 11-20.

[5] H. F. Wang and Z. Mei, "Chinese Multi-Document Personal Name Disambiguation," *High Technology Letters*, Vol. 11, No. 3, 2010, pp. 280-283.

