Communications and Network, **2012, 4, 30-33**
http://dx.doi.org/10.4236/cn.2012.41004 Published Online February 2012 (http://www.SciRP.org/journal/cn)

An Improved Name Disambiguation Method Based on Atom Cluster

Yu-Feng Yao
Computer Science and Engineering College, Changshu Institute of Technology, Suzhou, China
Email: 16341942@qq.com

Received December 13, 2011; revised January 25, 2012; accepted February 10, 2012

ABSTRACT

An improved name disambiguation method based on atom cluster. Aiming at the method of character-related properties of similarity based on information extraction depends on the character information, a new name disambiguation method is proposed, and improved *k*-means algorism for name disambiguation is proposed in this paper. The cluster analysis cluster is introduced to the name disambiguation process. Experiment results show that the proposed method having the high implementation efficiency and can distinguish the different people with the same name.

Keywords: Relation; Name Disambiguation; Data Mining; Entity

1. Introduction

Recently, with the development of the endowment insurance and medical treatment, a variety of core business processing systems came into being. But in the core business processing systems database, due to the limitations of information recording, it is difficult to find the sole primary key for the real customers. The first problem faced is the customer name collisions problem in insurance business dealing process. Data Mining [1-3] 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.

The literature [4,5] represent the content of name disambiguation to the vector space model to realize the name disambiguation. The literature [6] realized the name disambiguation by further extracting some information, such as the race, gender, education, work, family relationships, and then calculated the figure similarity. The above methods 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 and the atom cluster is used to improve the traditional clustering algorism. The experiment shows, the method is this paper can optimize the match processing of Chinese characters, having the high efficiency, it is suitable to be applied in the insurance field having amount of data.

2. System Frame and Algorism Principle

The same name disambiguation is a regular problem in insurance field. The clustering algorism is used to disambiguate the same name.

Definition 1. Same name disambiguation problem of insurance can be defined as: for the given customer identifier set S_r, the cooperation relation set $S_c = \{(r_i, r_j)\}$, for the same or similar customer identifier of S_r such as r_i and r_j, to calculate the set $S_c = \{e_j\}$, in which S_e is a entity set, and e_i represent the entity corresponding with the customer identifier.

3. Name Disambiguation Based on Atom Cluster

Atom clusters refer to the entity with strong ties in the clustering process will not be dismantled. The same name disambiguation flow of Clusters is showed as **Figure 1**.

The same name disambiguation process based on atom clusters mainly contains two steps: identification of clusters and the same name disambiguation. In the first step, the entity with a strong relationship is identified and it is as the input of the name disambiguation, using the classifier based on AdaBoost to calculate the connection level between entities to justify it satisfy the standard of atom cluster or not; and then the *k*-means cluster is used to name disambiguation using the output of the first step.

The principle of AdaBoos algorism is given a train set such as $(x_i, y_i), \cdots, (x_n, y_n)$, in which x_i belongs to a domain or instance space X, $y_i \in \{-1, +1\}$. In the initial time, the distribution of the given train set of AdaBoost is 1/m, and according to the distribution using the weak learner to train the train set, after train, according the

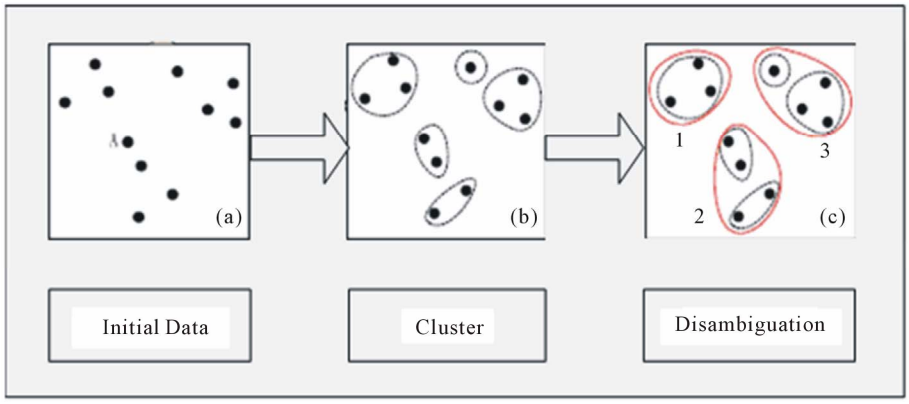

Figure 1. Name disambiguation flow based on atom cluster.

train result to renew the distribution of train set, and according to the new sample distribution to train, after the iterative rounds, finally a sequence of estimate can be concluded as $h_1,\cdots,h_r$, every estimate has some weight, the final estimate H is obtained by weight voting, and the probability of every sample appeared in the new train set is obtained by AdaBoost algorism, the train error of the final prediction function *H* is satisfied with the Equation (1):

$$H=\prod\left[2\sqrt{\varepsilon_t\left(1-\varepsilon_t\right)}\right]=\prod\sqrt{1-4\gamma_t^2}\le\exp\left(-2\sum_t\gamma_t^2\right) \quad (1)$$

In Equation (1), ε_t is the train error of prediction function h_t, $\gamma_t=1/2-\varepsilon_t$, and from Equation (1) we can conclude the train error is deduced with *t*.

After the classification of the entities with the different connection intensity using AdaBoos algorism, the *k*-means cluster is used to recognize the name disambiguation as follows.

The core idea of K-Means cluster is to classify to k cluster of *n* data, and make the sum of square for every data of cluster to the cluster, and the algorism is managed as follows.

The Definition 1. *k*-means cluster for name disambiguation algorism.

Input: Cluster counts *k*, the data set contains *n* data object.

Output: *k* cluster.

Step 1. Choose any *k* objects from *n* data objects as the initial cluster center.

Step 2. Calculate the distance of every object to the every cluster center, and the object is assigned to the nearest cluster center.

Step 3. After the assign of all the objects is finished, the k cluster center is re-calculated.

Step 4. Compare with the previous calculated *k* cluster center, if the center of cluster is changed, then goes the step 2 else goes the step 5.

Step 5. Output cluster result.

The details of the algorism can be described as follows.

Firstly, choose *k* object as the initial cluster center from *n* data object, for the other objects, according to their similarity with the cluster center, assign to the similar cluster to them.

Then calculate the new cluster center of every new cluster, repeat the process till all the standard measure function is in convergence, using the standard deviation as the standard measure function, it is defined as follows:

$$E=\sum_{i=1}^{k}\sum_{peC_i}\left|p-m_i\right|^2 \quad (2)$$

In equation, *E* is the sum of square for standard deviation of all the objects of the database, *p* represents a spot in object space, m_i is the average value of cluster C_i.

4. Simulation Experiment

4.1. Problem Description

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.

4.2. Atom Cluster Simulation

In the life insurance field, usually, a policyholder for the insurance and the insured designated beneficiary, so that

the insured, the insured, the beneficiaries are existed the relationship. Any two of the three constitutes two connected network. In the social network analysis, the small network is called atom cluster. Through the cluster analysis for the atom cluster, the target of the same name is realized. **Figures 2** and **3** showed the insurance of the networks before and after the name disambiguation.

The clusters found by comparing the simulation agents can effectively distinguish the same name.

5. Conclusions

A novel improved method based on cluster analysis is introduced. In order to improve the traditional cluster algorism by using the atom cluster, and 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 improv-

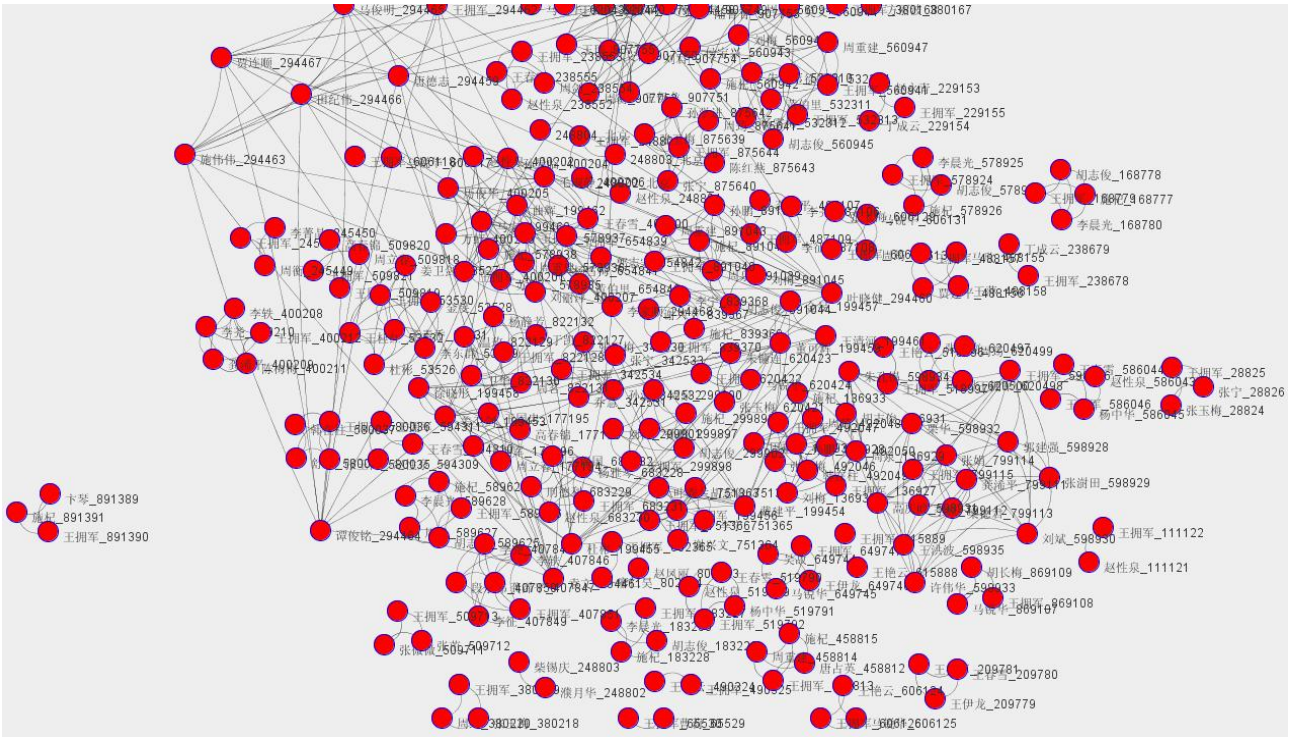

Figure 2. Insurance net of agent "Yongjun Wang" before name disambiguation.

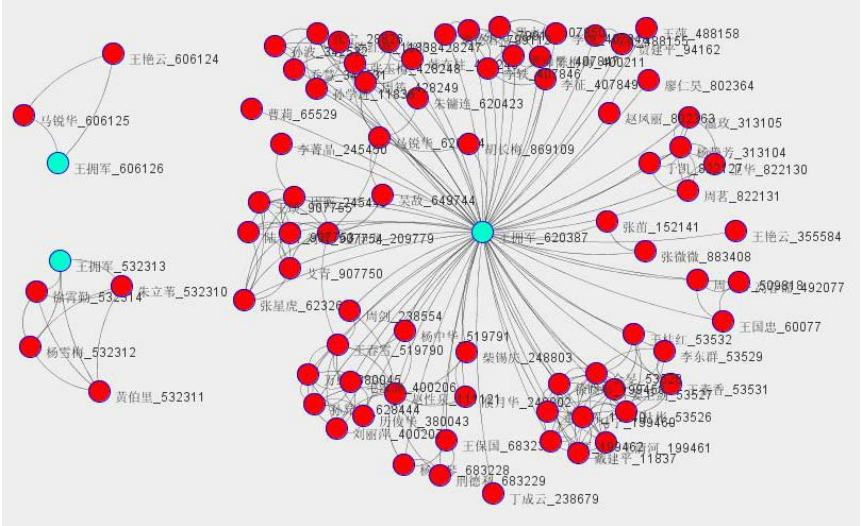

Figure 3. Insurance net of agent "Yongjun Wang" after name disambiguation.

ing 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.

The next work is to evaluate performance difference among the method proposed in this paper with the other name disambiguation methods.

REFERENCES

[1] H. S. Xia, *et al*., "Data Warehouse and Data Mining Technology," Science Press, Beijing, 2004.

[2] J. W. Han and M. Kamber, "Data Mining Concept and Technology," Mechanical Industry Press, Beijing, 2001.

[3] E. G. Mallach, "Decision Support and Data Warehouse System," Electronic Industry Press, Beijing, 2001.

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

[5] Y. M. Quan, "Research on Knowledge Mining in Person Tracking," Ph.D. Dissertation, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 2006.

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

