Research of People Disambiguation by Combining Multiple knowledges

Erlei Ma
School of computer science and technology, harbin institute of technology
elma@insun.hit.edu.cn

Yuanchao Liu
School of computer science and technology, harbin institute of technology
ycliu@hit.edu.cn

Abstract

With the rapid development of Internet and many related technology, Web has become the main source of information. For many search engines, there are many different identities in the returned results of character information query. Thus the Research of People disambiguation is important. In this paper we attempt to solve this problem by combing different knowledge. As people usually have different kind of careers, so we first utilize this knowledge to classify people roughly. Then we use social context of people to identify different person. The experimental results show that these knowledge are helpful for people disambiguation.

1 Introduction

For the real world, many people share one name; this is a very common phenomenon. According to the third national census sample survey conducted by the State Language Committee in 1989, the duplicate names rate for single name was 67.7%, whereas that of double name was 32.4%.

There are two commonly used name disambiguation approach, one is based on the vector space model, and the other is based on social networks.

The first is text-based vector space clustering approach. An entity can be expressed as one vector which is formed according to the content word of the original document. And then the similarity is used to merge documents or classify documents.

The second method is based on social networks. The first step of the method is to build social networks, by analyze the relationship of different people. Generally if two people's name always occurs in same document or very near context ,they will have close relations, one of them will be helpful for disambiguate the other.

In this paper, we first use the domain of character's document to classify roughly, and then context information using social networking is considered again to disambiguate person's name again.

2 the principle of our system

Fig.1. shows the basic principle of our system. The basic steps are:

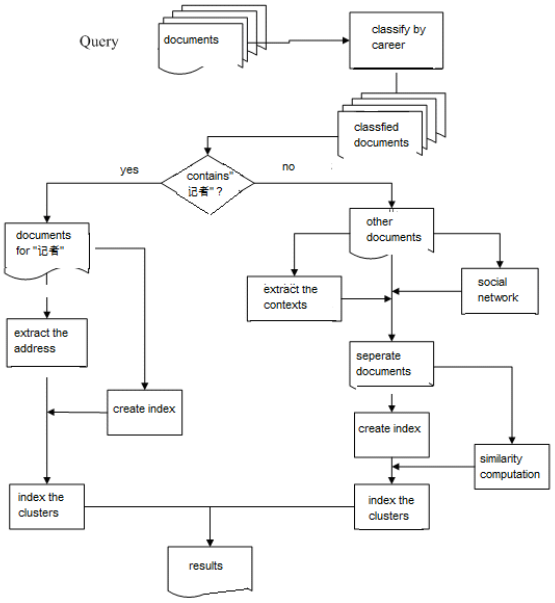


Fig.1. the general framework of our approach

1) documents with same people's name are input;
2) classify these documents into seven careers which include Cultural, administrative, military, science, education, sports, health, economic and etc;

3） Judge if the people are reporter in document, if yes; separate them according to their address.

4） Separate documents by using social networks. This is because different people usually have different social relations. Different social relations usually means different people and different identity. The social network of one people is gained by counting its co-occur frequency with other peoples.

3 experimental results

3.1 evaluation method

Here are the evaluation formula provided by SIGHAN 2010:

$$\text{Precision}_i = \frac{\sum_{S_i \in S} \max_{R_j \in R} |S_i \cap R_j|}{\sum_{S_i \in S} |S_i|} \quad (1)$$

$$\text{Recall}_i = \frac{\sum_{R_i \in R} \max_{S_j \in S} |R_i \cap S_j|}{\sum_{R_i \in R} |R_i|} \quad (2)$$

B-Cubed：

$$\text{Recall}_i = \frac{\sum_{R_i \in R} \sum_{d \in R_i} \max_{S_j \in S; d \in S_j} \frac{|R_i \cap S_j|}{|R_i|}}{\sum_{R_i \in R} |R_i|} \quad (3)$$

$$\text{Precision}_i = \frac{\sum_{S_i \in S} \sum_{d \in S_i} \max_{R_j \in R; d \in R_j} \frac{|S_i \cap R_j|}{|S_i|}}{\sum_{S_i \in S} |S_i|} \quad (4)$$

$$F - measure_i = \frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \quad (5)$$

The overall precision and recall is as follows:

$$\text{Precision} = \frac{1}{n} \sum_{i=1}^{n} \text{Precision}_i \quad (6)$$

$$\text{Recall} = \frac{1}{n} \sum_{i=1}^{n} \text{Recall}_i \quad (7)$$

$$F - measure = \frac{1}{n} \sum_{i=1}^{n} F - measure_i \quad (8)$$

3.2 The performance of our system

By only utilizing the career domain knowledge, the performance is shown in table 1. Obviously the people in this division of the seven categories, the accuracy is low and the recall rates were high. The reasons include the following:

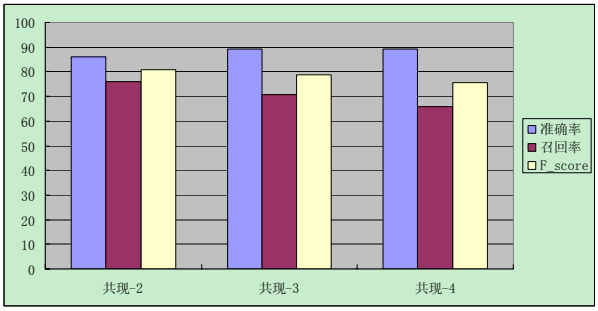
First, in he document pre-classification processing, the named entity recognition has not been carried out in the text dealing with the classification of the document. Some of them are not the people's name.

Second, different people may have same domain, thus the accuracy is adversely affected.

Table 1 . The performance after the first-step classification

	precision	recall	Fmeasure
B-Cube	28.78	99.97	44.69
P_IP	42.82	99.97	59.96

By adding the knowledge of social networks, the performance is shown in Fig.2-Fig.3.



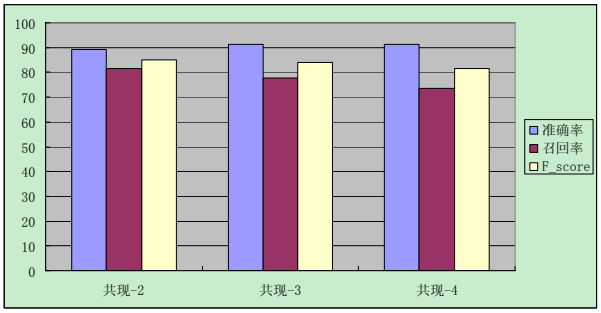


Fig.2 result of B-Cubed

Fig.3. result of P_IP

Clearly the experiment showed that after matching character society attribute information, the recall rate increased significantly, and the F value also have increased. .

4 Summaries

In this paper, we utilize two kind of knowledges: 1) people always have his own career; 2) people have his own social circle. We think these information will be more helpful for disambiguation. Thus we attempt to solve this problem by combing different knowledge. As people usually

have different kind of careers, so we first utilize this knowledge to classify people roughly. Then we use social context of people to identify different person. In the future we wish to address the following aspects: 1) add and improve name recognition accuracy; 2) extract and select the useful context of person's name, which is the problem of information extraction; 3) recognize some kind of public people such as political leaders, famous singers and etc. to improve the effect of social networks.

References

[1]Amit Bagga and Breck Baldwin.Entity Based Cross-Document Coreferencing Using the Vector Space Model In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics(COLING-ACL'98),1998 :79-85.

[2] Gideon S. Mann and David Yarowsky. Unsupervised Personal Name Disambiguation In Proceedings of the seventh conference on Natural language learning at HLT-NAACL, 2003: 33-40.

[3] Bollegala, D., Y. Matsuo, M. Ishizuka. Disambiguating Personal Names on the Web Using Automatically Extracted Key Phrases. In: Gerhard Brewka, Silvia Coradeschi, Anna Perini, Paolo Traverso, eds. Proc. of the 17th European Conference on Artificial Intelligence. Riva del Garda, Italy: IOS Press, 2006:553-557

[4] Bekkerman, Ron, Andrew McCallum. Disambiguating Web Appearances of People in a Social Network. In: Allan Ellis, Tatsuya Hagino , eds. Proc. of the 14th international conference on World Wide Web. Chiba, Japan: ACM Press, 2005:463-470

[5] Javier Artiles, Julio Gonzalo, Satoshi Sekine. The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task. IN: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007),2007: 64–69

[6] Malin, Bradley. Unsupervised Name Disambiguation via Social Network Similarity. In: Hillol Kargupta, Jaideep Srivastava, Chandrika Kamath, Arnold Goodman, eds. Proc. of the Workshop on Link Analysis, Counterterrorism, and Security, in conjunction with the SIAM International Conference on Data Mining. Newport Beach, California, USA: SIAM, 2005:93-102

[7] Nahm, U. Y. and Mooney, R. J.; Text Mining with Information Extraction. In Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases, Stanford, CA, March 2002: 60-67.

[8] Yang, Y., and Jan O. Pedersen. A comparative study on Feature Selection in Text Categorization. Proceedings of the Fourteenth International Conference on Machine Learning Table of Contents, 1997: 412-420.