A novel clustering algorithm for Unsupervised Relation Extraction

Jing Wang[1] ,Yang Jing [1*]
Department of Computer Science and Technology,
East China Normal University
Shanghai, China
{jwang@ica.stc.sh.cn, jyang@cs.ecnu.edu.cn}

Yue Teng[2],Qingling Li[2]
Department of Computer Science and Technology,
East China Normal University
Shanghai, China
{tengyue5i5j@163.com, *qlli@ica.stc.sh.cn}*

***Abstract*—Clustering of entity pairs is the core content of the unsupervised relation extraction method. However, most of the clustering algorithm in the previous unsupervised relation extraction does not take into account the influence of the duality between entity pairs and the relationship characteristics on clustering results. In order to overcome this defect, this paper proposed a novel clustering algorithm for unsupervised relation extraction. It introduces co-clustering theory on the basis of the k-means clustering, not only clustering the entity pairs but also clustering the relationship characteristics to make full use of the duality of the clustering dataset. The final experimental results demonstrate that our clustering algorithm get more higher accuracy rate than k-means clustering algorithm in unsupervised relation extraction.**

Keywords- unsupervised relation extraction; co-clustering theory; duality

I. INTRODUCTION

With the development of information technology, especially the development of the Internet, a lot of information emerges in electronic form. There is an urgent need for automated tools to get useful information from huge amounts of data quickly to cope with the challenges of information explosion. In this case, information extraction produced. At the same time，relation extraction as one of the subtask and key technologies of information extraction has also been more and more attention.

Currently, researchers are mainly concentrated in the supervised and weakly supervised of machine learning methods. Reference [1] used “Relation” and “Event” two ontology databases to carry out supervised relation extraction. Reference [2] used Winnow and SVM two machine learning algorithm to implement weakly supervised relation extraction. However, supervised and weakly supervised relation extraction cannot automatically identify the relationship which is not pre-defined. This is not appropriate for some applications. Therefore, researchers have begun to study on unsupervised relation extraction method.

In 2004, Hasegawa [3], et al. put forward an unsupervised approach for relation extraction from large text corpora. First, they applied named entity recognition technology for the identification of entities. Second, they extracted the contexts of each entity pairs as relationship characteristics. Then according to the contexts, they clustered the entity pairs by a clustering method. Final, they selected the most frequent words in the contexts to represent the relation that holds between the entities.

Chen [4] improved the method of Hasegawa. They proposed a clustering quality assessment formula, which can get the optimal number of clusters and the feature vectors, and then applied the k-means clustering algorithm to cluster. At the same time, they used DCM (Discriminative Category matching) method [4] to choose relations description words. Reference [5] was mainly through the improvement of k-means clustering method to improve the overall performance of unsupervised relation extraction.

In short, Hasegawa first proposed the basic steps of unsupervised relation extraction method, and laid an important foundation for it. Fig. 1 shows the general process of unsupervised entity relation extraction method.

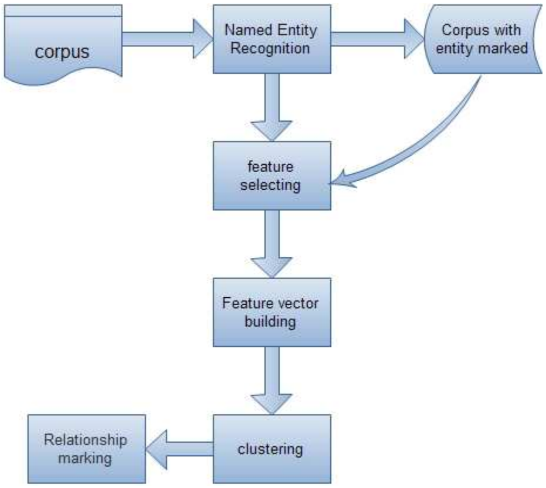

Figure 1. the basic flowchart of unsupervised relation extraction

However, the researches of unsupervised relation extraction method are still in its preliminary stage. There are still some inadequacies, such as not enough reasonable clustering results of entity pairs, low accuracy of relation extraction results, few researches on Chinese relation extraction and so on.

This paper mainly studies on the clustering method of unsupervised relation extraction and proposes a novel clustering algorithm. It makes full use of the duality of the clustering dataset and uses the advantages of the k-means clustering in time and space. The experimental results show

978-1-4673-2430-4/12/$31.00 ©2012 IEEE

that our method obtains a higher accuracy rate than k-means clustering algorithm in unsupervised relation extraction.

The remainder of this paper is organized as follows. In section II, the related work is surveyed. The section III presents the related details of our approach. The experiment results are shown in section IV. The conclusion and future work are drawn in section V.

II. RELATED WORKS

Currently, more and more researchers have begun to try to find a good clustering algorithm to improve the overall performance of the unsupervised relation extraction. Hasegawa [3] adopted a fully connected clustering method in hierarchical clustering method to cluster the contexts of entity pairs in Clustering module. However, hierarchical clustering doesn't have good scalability and need to check and estimate a large number of objects or grouping in splitting or mergering module. In addition, the time complexity of hierarchical clustering is $O(n^2)$ .It is inappropriate when dealing with large data sets. For solving the above problems, Reference [5] put forward a new clustering method. First, during the clustering they applied the hierarchical clustering method in the sample subset in order to obtain the initial cluster centers and the number of clusters. Second, they used the cluster center and the number of clusters obtained as a k-means the input parameters of the algorithm to cluster on the entire data set. Chen [4] proposed a clustering quality assessment formula, which can get the optimal number of clusters and the feature vectors, and then applied the k-means clustering algorithm to cluster, thereby improving the overall performance of relation extraction.

From the above study, it can be seen that most clustering method for entity pairs are based on k-means when people research unsupervised relation extraction method. That’s because the k-means clustering algorithm has been proved to be more suitable for large data sets and more suited for finding that the globular cluster in data set. In relation extraction, dimensions and number of clustering objects are relatively large and in [5] pointed out that in the natural language corpus of this field, the contexts of entity pairs quantized was usually based on high-dimensional local "spherical" distribution, which was consistent with the distribution of natural language words. Therefore, most clustering method for entity pairs is based on k-means in unsupervised relation extraction that has a certain basis. So, this paper also bases on k-means clustering algorithm to cluster for entity pairs.

However, we can see that most clustering method for entity pairs is a one-way clustering in unsupervised relation extraction. That is only to calculate the similarity between the entity pairs and cluster the entity pairs. Without taking into account the correlation between the feature vectors which entity pairs corresponds to and the mutual influence between the correlation and clustering for entity pairs. Reference [6] proved that the cluster matrix of data points and corresponding to the relationship feature vectors were interdependencies and the interdependencies was called as characteristics of "duality".

Therefore, in order to take full advantage of the characteristics of the "duality" between entity pairs and the feature vectors, this paper proposes a new clustering method. It introduces co-clustering theory [7] on the basis of the k-means clustering, not only clustering for entity pairs but also clustering for the relationship feature vectors of entity pairs to get a more reasonable clustering result.

III. MY METHOD

In this paper, we propose a novel clustering algorithm in unsupervised relation extraction. This section presents the related details of our approach, including analysis of the duality, co-clustering theory and the steps of the algorithm.

A. Data Duality analysis

Generally, most clustering data are expressed by the form of matrix. The row represents the clustering objects, while the column represents the features of data. The clustering algorithms mainly make similar objects in one group, while objects should not be in the same group with little similarity. In this paper, on the basis of better clustering results, we can get better relationship extraction results of different entities in unsupervised relation extraction.

Reference [6] pointed that there existed relationships among features of columns, which directly affected the similarity calculation of different line objects in clustering process; all of these finally determined the rationality of clustering results. Also the author [6] took an example to explain the above phenomenon of relationships among features of objects. In Fig. 2(a), supposing that X, Y and Z axis represent the features of objects in three-dimensions; also in the space existing three data points as point a, b, c. If we don’t take the relationship of features among objects in columns into consideration, then the distances of points as |ab|, |ac| and |bc| are all the same. However, if there exists a plane in XY containing some datasets, including point a, b, for plane G in Fig. 2(b).From the Fig we can see that the distribution of datasets in X,Y axis possess some potential relationship according all the points lies in G plane. On the other hand, considering the relationship mentioned above, we could conclude that the distance of |ab| must be shorter than |ac| and |bc|, like this we only can get better clustering results.

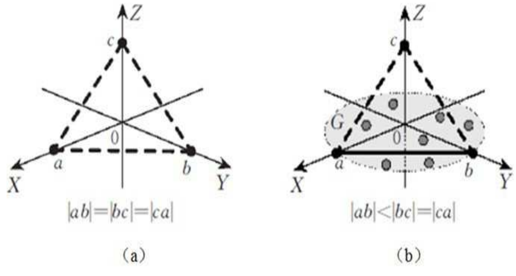

Figure 2. the instance of Data Duality

Reference [6] concludes the phenomenon as the data duality between dataset and features of matrix column. That is to say, first there is a certain correlation between the different features in given features space. Second the distribution of features should affect the clustering of data objects.

On the basis of analysis above, we can conclude that if there is a data duality characteristic in a clustering matrix dataset. In order to get reasonable clustering results, we should not only consider the clustering of data objects, but also the influence of the correlation of features on clustering. So in this paper, we proposed a new clustering algorithm according to date duality in dataset matrix, the next section we will introduce our clustering algorithm.

B. Co-clustering Theory Analysis

Considering the data duality, in order to get reasonable clustering result, here we introduce co-clustering idea to k-means clustering algorithm to solve the data duality problem mentioned above, clustering column and line at the same time. Co-clustering explores the relationship of same dimension, not like traditional clustering only considered column or line. While co-clustering can be well applied in the below situation: the quality of matrix line clustering determined by columns, the reverse is also satisfied. Compared with traditional clustering, when introducing co-clustering idea into k-means clustering; we can get more advantages [8] such as:

- Mining the potential partial pattern in clustering matrix. Owing to co-clustering both considers column and line in clustering matrix; fully consider the data duality in clustering matrix.
- Potentially decreasing two dimensions of clustering matrix. Traditional clustering only decreases column or line of clustering matrix in single dimension, while co-clustering in column and line one time, even clustering in one dimension can also get better clustering result. Also co-clustering can solve matrix sparse problem in some extent.
- Time complexity. As we know that with the growth of the clustering data matrix, a stronger clustering algorithm shows more important for some systems and applications. Since co-clustering decreases scale of dataset matrix in two dimensions, which can reduce time complexity clearly in later computation.

Combined with advantages of k-means clustering algorithm in time and space complexity, also the situation to data duality of co-clustering, this paper proposed a new clustering algorithm based on k-means clustering and co-clustering theory to improve the performance of unsupervised relation extraction.

C. Clustering Algorithm Steps

To unsupervised relationship extraction, entities are divided into some general genres, clustering often are based on the genres. The genres can be generated from the type of entities-self. Take ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) for example, using this tool we can get the genres of entities, the type of “李彦宏” (“Yanhong Li”) is “nr”, “百度” (“Baidu”) will be taken as “nt” and so on. Here “nr” represents “person”, “nt” represents “organization”. All description of entities is the subset of genres. So in this paper we also adopt this tool to tag all entities, then on the tagged entities to make clustering. With no special situation, all the clustering is on basis of genres of entities-self. Then we can get the clustering dataset as table I:

TABLE I. ENTITY-FEATURES MATRIX

	Word-1	**Word-2**	**…**	**Word-j**	**Word-n**
Entity-1	R_{11}				R_{14}
Entity-2		R_{22}	R_{23}		R_{24}
…					
Entity-i	R_{31}	R_{32}		R_{ij}	
Entity-m					

On the other hand, we suppose all the entity pairs set as E and relation features set expressed as W, the relationship between the two are showed in table I, R_{ij} represents the number of features which entities contains. Now we will introduce the new clustering algorithm as follows:

Input：Entity pairs set expressed as E，features set as W，Entity pairs-Features matrix as A，cluster number will be selected as k.

Output: C_E(clustering result of E), C_W(clustering result of W).

Step1： Randomly select k objects from E and W as the initial clustering centers of C_E and C_W respectively.

Step2： Select one object excluding k clustering centers from W to calculate the similarity with k clustering center, finally the object will be adhered to the most similar center, also delete the object column.

Step3： Select one object excluding k clustering centers from E to calculate the similarity with k clustering center, finally the object will be adhered to the most similar center, also delete the object line.

Step4: Go to Step2 and Step3 unless the matrix is empty.

Step5: If the line of A is empty，go to Step2，if column of A is blank，go to Step3.

Step6： Update all the clustering results, that is re-calculate the mean value of the clustering center in each cluster of C_E and C_W.

Step7： For every clustering center, take the object as new clustering center object, which is most similar to original clustering center.

Step8： If the clustering criterion function does not converge, Re-constitute the matrix A, then repeat steps 1 to 7.

Step9： Until getting the convergence of the criteria function.

IV. EXPERIMENTS

A. *Experimental DataSet*

In this paper, the experimental dataset is taken from 2000 documents in Web. In order to evaluate the final results of relation extraction, we analyze the data set manually and identify the relation types for two different domains like ACE corpuses. One is the PERSON- PERSON (PER-PER) domain. We get 119 distinct entity pairs and classify them into 7 relations (classes) manually. The other is the PERSON–ORGNIZATION (PER-ORG) domain. We obtain 70 distinct entity pairs and classify them into 5 classes manually. Table II shows the types of classes and the number in each class.

TABLE II. EXPERIMENTAL DATASET

Domain	Class	The number of Entity pairs
PER-PER	夫妇(couple)	32
	扮演者(actor)	24
	学者 (scholar)	19
	经纪人(broker)	14
	爱将(pupil)	12
	父亲(father)	10
	经济学(economist)	8
PER-ORG	校长(principal)	21
	董事长(chairman)	18
	任教(teach)	15
	局长(director)	9
	就读(studying)	6

B. *Evaluation Metric*

In this paper, we prove the effectiveness of our method in unsupervised relation extraction by evaluating the final results of relation extraction using precise from information retrieval [9].

Precise can be defined as follows:

$$P = \frac{N_{\text{correct}}}{N_{\text{correct}} + N_{\text{error}}} \quad (1)$$

While $N_{correct}$ is the number of correct results of relation extraction, N_{error} is the number of error results of relation extraction. During conducting the assessment, we determines the results are correct or not by artificial judgment. For example, we get a result that the relationship of entity pairs “A” and “B” is “husband”, then manually compare it with the sorted entities in table II. In our data sets, the relationship of entity pairs “A” and “B” is “couple”. Since “husband” can reflect the mean of “couple”, this paper thinks the result is correct.

C. *Parameters Settings*

The number of clustering k is very important, for its significant effect on the final cluster results. Reference [10] proposed that in order to determine the optimal value of k, it normally needed to do the following works: determining the range of the number of clusters depending on the size of the dataset, gaining the clustering results in different number of clusters by the clustering algorithm, selecting the validity index to evaluate the results obtained by clustering, by the results of the assessment to get the optimal number of clusters.

This paper uses Silhouette index [11] as the evaluation metric to obtain the optimal value of k. The range of the value of Silhouette index is [-1, 1], the value reflects the compactness within the class and the separation between classes. In a dataset, the higher the average value of Silhouette index, the better the clustering results.

Fig. 3 and Fig. 4 show the Silhouette index value corresponding to the different values of k (from 2 to $\sqrt{n}$, n is the total number of data objects) respectively using k-means clustering method and co-kmeans clustering method in the domains of “PER-PER” and “PRE-ORG”.

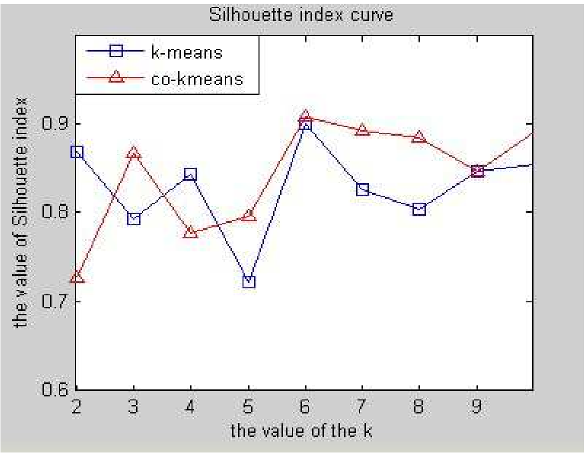

Figure 3. Silhouette index curve in “PRE-PRE”

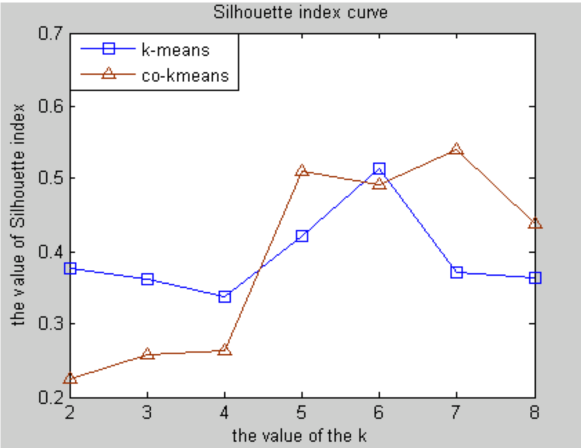

Figure 4. Silhouette index curve in "PRE-ORG"

By the two figures above, we can get the optimal value of k in k-means method and co-kmeans method. Table III shows the specific content.

TABLE III. THE OPTIMAL VALUES OF K AND SIL

Optimal values of k and sil Domain	k-means method	co-kmeans method
PER-PER	6(0.9008)	6(0.9075)
PER-ORG	6(0.5144)	7(0.5393)

D. *DCM Relationship Marked Results*

According to the optimal values of k in Table 3, we can get the clustering results of the two algorithms (k-means and co-kmeans). In this paper, we verify the validity of our cluster algorithm by its influence on the results of the unsupervised relation extraction, so after getting the clustering results, we uses DCM weights calculation method to allocate weight values for the words of each cluster and chooses the two largest DCM weight value in each cluster corresponding to the characteristic words as class descriptors.

Through the two feature words of each cluster, we can roughly understand the relationship of entity pairs in a cluster. However, we hope that only chose a feature words to describe the relationship of an entity pair. After getting the two feature words of each cluster, we match them with the relationship characteristics words of each entity pairs in the cluster and according to the following matching rules to mark the relationship of entity pairs:

（1） If successfully matching with one word, choose the word as the relation description word of the entity pair.

（2） If successfully or fail matching with two word, choose the word with the largest DCM value in two words as the relation description word of the entity pair.

Table IV shows the partial marked results of entity pairs accordance to the above matching rules.

TABLE IV. PARTIAL MARKED RESULTS OF ENTITY PAIRS

Entity Pairs	marked relationship	true relationship
李松山，韩蓉 (Songsan Li, Rong Han)	丈夫 (husband)	夫妇 (couple)
沙僧，闫怀礼 (Seng Sha, Huaili Yan)	丈夫 (husband)	扮演者 (actor)
狄云，吴樾 (Yun Di,Yue Wu)	电视剧 (TV series)	扮演者 (actor)
琼瑶，蒋勤勤 (Yao Qiong, Qinqin Jiang)	剧中 (TV series)	爱将 (pupil)
马云，淘宝 (Yun Ma,TaoBao)	公司 (company)	董事长 (chairman)
朱维铮，复旦大学 (Weizheng Zhu, Fudan University)	教授 (professor)	任教 (teach)
…	…	…

E. *Experiment and Result And Analysis*

According to the evaluation metric in this paper, we can get the final performance of the two methods (k-means and co-kmeans) in our dataset as table V shows:

TABLE V. THE PERFORMANCE OF DIFFERENT METHODS

The Accuracy(P) Domain	k-means method	co-kmeans method
PER-PER	35.31%	37.05%
PER-ORG	54.41%	60.29%

It can be seen from the experiment results that:

（1）Compared with the traditional k-means clustering algorithm, the proposed clustering algorithm in this paper obtains a higher accuracy rate in unsupervised relation extraction .It is because that our method not only takes into account the duality characteristic between entity pairs and relationship characteristics but also uses of the advantages of the k-means clustering in time and space.

（2）Observing the accuracy of the experimental results in both domains, we can find that the accuracy in “PRE-PER” is lower than in “PER-ORG”. In order to get the reason of this phenomenon, we analyze the clustering dataset of the both domains and find that the relationship characteristics in “PER-ORG” are more effective than in “PER-PER”. In other words, the effectiveness of clustering dataset will affect the clustering results in unsupervised relation extraction.

V. Conclusion and Future Work

The research of entity relation extraction technology is the need of the development and application of multi-disciplinary, and has far-reaching theoretical significance and broad application prospects. Unsupervised relation extraction starts late and still exist many problems to be solved. This paper proposes a new clustering algorithm for unsupervised relation extraction. It brings into the thinking of co-clustering in order to make full use of the duality of the clustering dataset and uses of the advantages of the k-means clustering in time and space. The experimental results show that our method obtains a higher accuracy rate than k-means clustering algorithm in unsupervised relation extraction.

For future work, on one hand, we will continue committing to study the clustering approach in order to improve the performance of unsupervised relation extraction. On the other hand, we are trying to get more valid clustering dataset and parameters for improving the accuracy of our clustering method.

Acknowledgment

The authors would like to thank the support from National High Technology Research and Development Program of China (863 Program) (No.2011AA01A107), the Shanghai Science and Technology commission Foundation (No.10dz150010B and 11511502203) and the Opening Project of Shanghai Key Laboratory of Integrate Administration Technologies for Information Security. Also we would like to thank other members in ICA lab for their support.

References

[1] C．Aone，M．Ramos Santacruz．REES：A Large Scale Relation and Event Extraction System,In：Proceedings Of the 6th Applied Nature Language Processing Conference[C]，2000：76—83.

[2] Wanxiang Che, Ting Liu, Sheng Li. Automatic Entity Relation Extraction. Journal of Chinese information processing. 2004, 19 (2):1-6.

[3] Takaaki Hasegawa, Satoshi Sekine and Ralph Grishman. Discovering Relations among Named Entities from Large Corpora. Proceeding of Conference ACL,Barcelona, Spain, 2004:415-422.

[4] Jinxiu Chen, Donghong Ji, Chen Lim Tan, Zhengyu Niu. Unsupervised Feature Selection for Relation Extraction. International Joint Conference on Natural Language Processing. 2005:262-267.

[5] Zhitian Zhang.The Reaeach of Relation Extraction with Unsupervised Method[D]. Harbin Institute of Technology, 2007.

[6] Peng Wang,Shiqiang Yang,Zhiqiang Liu.Information-Theoretic Co-Clustering for Video Shot Categorization[J].Chinese Journal of Computers,2005(10).

[7] Q.Gu and J.Zhou.Co-clustering on manifolds.In Proc.of KDD’09,pages 359–367,2009.

[8] Chongtie Ye. Researeh and APP1ication On Co-C1uster1ng A1gor1thms for High D1mens1onal and Very Large Data[D]. Zhejiang Gongshang university.2010.

[9] Gerard Salton. Automatic Text Processing: The transformation, analysis,and retrieval of information by computer. Addison-Wesley, 1989.

[10] Shibing Zhou,Zhenyuan Xu,Xu Tang. New method for determining optimal number of clusters in K-means clustering algorithm.Computer Engineering and Applications[J]. 2010,46(16) : 27-31.

[11] Dudoit S,Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset.Genome Biology,2002,3(7):1-21.