PROCEEDINGS OF SCIENCE

Entity Relation Extraction Method based on Improved K-means Clustering

Bin Yu[1]
School of Computer Science and Technology, Xidian University
Xi'an, 710071, China
E-mail: yubin@mail.xidian.edu.cn

Ke Pan[2]
School of Computer Science and Technology, Xidian University
Xi'an, 710071, China
E-mail: try_panky@163.com

Chen Zhang[3]
School of Computer Science and Technology, Xidian University
Xi'an, 710071, China
E-mail: zhangc@xidian.edu.cn

Yu Xie
School of Computer Science and Technology, Xidian University
Xi'an, 710071, China
E-mail: sxlljcxy@gmail.com

Jiangyan Sun
School of Engineering, Xi'an International University
Xi'an, 710077, China
E-mail: 275125609@qq.com

This paper presents an unsupervised method of extracting entity relation from large-scale corpus which is based on the hypothesis that a named entity with the same relation has a similar context, analyzes the co-reference relation between the co-reference substance to be tested and the object to be resolved, completes the construction of the entity according to the adjacent principle of the type entity and the core word principle, and uses the relative position restriction rule to combine the context window method to extract the feature and construct a feature sequence. In the end, the completion of the entity relation extraction task is based on the improved K-means clustering algorithm. The experimental results show that the new method can effectively improve the effect of entity relation extraction with a certain practical value.

ISCC2017
16-17 December 2017
Guangzhou, China

[1]Speaker: Bin Yu

[2]This work is supported by the National Natural Science Foundation of China (No.6150 2365), the Fundamental Research Funds for the Central Universities (No.JB170305)

[3]Correspongding Author: Chen Zhang

PoS(ISCC 2017)001

1.Introduction

With the arrival of the era of big data, information resources show explosive growth. Its scope, scale, the spread of the speed and breadth have reached an unprecedented level. Data has become an important basis and strategic resources to drive the growth of economic and the progress of social. How to use the massive data serves people effectively is the hot topic in today's society, the information extraction research comes into being under this kind of background. Information extraction is the transformation of unstructured text into structured information for user to query, further analysis and utilization, and information extraction system is the extraction of the specific entity from the text. However, in practical applications, we should not only identify the entities in the text, but also determine the relationships among these entities, which we call entity relation extraction.

Entity relation extraction task refers to automatically identifying the association between the entities from a given corpus. The so-called entity refers to the real facts that exist in the real world, such as time(TIME), person(PER) and so on[1]. The association can refer to a pre-defined relation, or an undefined but occurring frequently descriptor that reflects the relation between the knowledge elements[2].

The extraction of entity relation is a powerful tool for people to obtain information. The research on entity relation extraction can be divided into two stages[3]. The first stage is the extraction of entity relation for specific fields and specific relations, which mainly uses the template matching method to work in the specific corpus environment, and achieves good results. The advantage of this approach is its strong relevance and high efficiency for a specific corpus environment, but the versatility and portability are poor. The second phase of the relevant research begins to focus on the introduction of various types of machine learning algorithm, achieves the relevant practice and parameter correction through the use of supervised, semi-supervised or unsupervised learning methods on the algorithm model, and continuously improves the accuracy and recall rate of entity relation extraction in massive data.

The paper introduces the extraction of the entity relation in general, and mainly classifies and analyzes the existing entity relation extraction algorithms, then puts forward the improvement of the K-means clustering algorithm and realizes the extraction of the entity relation. At last, this paper proposes the verification of the improved algorithm and summarizes the current research status, and prospects the future research direction.

2.Related Work

The current method of entity relation extraction is based on knowledge engineering method and machine learning method. The knowledge engineering method is mainly achieved through expert writing rules and pattern matching. The method of machine learning[4] is based on the statistical model, and the relation extraction is transformed into the classification problem. The method of entity relation extraction based on machine learning can be divided into supervised method, semi-supervised method and unsupervised method.

The supervised relation extraction trains the relation extraction model on the corpus of entity relation which has been annotated manually, and then use this model to implement the relation extraction of the corpus which is to be tested. Such as Gumwon Hong of the university of Michigan have achieved SVM[5] relation extraction system based on supervised eigenvector

method, Chen Peng[6] proposes a method of Chinese domain entity relation extraction based on convex combination kernel function, which solves the problem of the difference of different kernel functions in relation extraction. Although with high accuracy, these methods are not suitable for large-scale applications because they need a lot of manual intervention.

The semi-supervised relationship[7] mainly takes the form of seed expansion, and acquires the new relationship model automatic through statistical learning method, or generates the instances with high credibility to expand the training data. For example, Chen Liwei[8] uses the collaborative training method based on Bootstrapping to strengthen the semi-supervised relation extraction model, and analyzes the collaborative strategy in detail. Wang Mingyin[9] proposes a semi-supervised machine learning method SCOERE to extract the relation tuple from the network text and achieves good results.

The so-called unsupervised relation extraction includes the clustering of relation entity and the selection of relation markers, and do not rely on annotation corpus. Such as Ma Chao[10] uses the unsupervised algorithm to extract the ontology from the web information, and then introduces the conceptual relation weight into the K-means algorithm, which is greatly reduce the corpus annotation cost of traditional algorithm. Wu Sheng[11] studies the unsupervised extraction method of numerical relation, and proves the effectiveness of the three industries in steel, ship and real estate.

3.Entity Relation Extraction Based On Improved K-means Clustering

In this paper, we use the improved K-means clustering algorithm to achieve the extraction of entity relation, which is based on the distribution hypothesis theory that the entities have the same relation have similar contextual content. Then the representative words can be selected from the content to describe the basic relation between entities. Entity relation extraction mainly includes entity co-reference resolution, entity construction and entity feature selection, relation triples clustering and so on.

3.1Entity Co-reference Resolution

Co-reference resolution based on machine learning is actually regarded as a binary classification problem, through the use of different algorithms of the classifier for training, analysis and determine whether the co-reference substance to be tested and the object to be resolved exist co-reference relation. It is divided into three steps: Firstly, the training samples are extracted based on syntactic and semantic analysis, extracting the entity object and the feature vectors based on sentence analysis. Then the classifier based on a machine learning algorithm is used to classify the samples, and the parameters of the classifier are corrected to make them better. Finally, the trained classifier is used to classify the test cases, so that the entities with the same reference attributes are associated with the same co-index chain, thus completing the co-reference resolution task.

3.2Entity Construction and Entity Relation Feature Selection

In this paper, only a specific type of named entity is selected as an entity, represented by (e_1, e_2). For example, in a large number of text data about profile, if you want to extract a person's date of birth information, it is necessary to constitute an entity by the name and time in each text, and then combining with the characteristics of the relation sequence which indicated

by $relationWord$, constructing a relation triple $(e_1, relationWord, e_2)$, and clustering these specific types of relation triples. In the process of clustering, we can classify the entities that represent the same kind of relations according to the characteristic relation. Finally, select the type of relation from the classified entities.

In this paper, two basic principles are used in the process of entity construction, which are the adjacent principle and core word principle. First given the following definition:

- Collection $E_{all}=((e_1, p_1),(e_2, p_2),\cdots(e_n, p_n))$ represents all entities contained in a sentence, where e_i represents a separate entity and p_i represents the attribute of entity e_i, such as name, title, institution and so on.
- Entity $E_i=(e_{i1}, e_{i2})$ represents two entities may have a relation in a sentence.
- For the collection of the characteristic relation sequence $R_n=(R_1, R_2, \cdots, R_n)$, any entity of E_i may find some representations of feature words between e_1 and e_2 in context, which is a collection of these features is its characteristic sequence.

Based on the above definition and the actual statistical analysis, this paper find that a sentence or context-related statement may contain a number of named entities, if all entities are combined to form the entity will reach an order of C_n^2. It is difficult to deal with it when the amount of data is large, so in order to simplify the difficulty of the processing, this paper proposes the adjacent principle of the entity of a specified type. In accordance with this principle, the construction of the entity only in the specified type when two entities are close to each other.

In order to improve the efficiency of the entity construction, the core word principle should be used in the construction process. The so-called core word is the statistics of an article in the high number of occurrences and the preferred field for the type of entity in a database. For example, in the construction of talent information database, the general name as the core word.

In this paper, we find that the relation between the relation markers in the sentence and the two entities e_1 and e_2 can be divided into five cases, The five cases are: the relative between the two entities, the relative is located on the left or right side of the entity, there is no relation indicator and other circumstances. In addition, the related research shows that when the number of words between two entities is less than 6, the possibility of its existence account for 74.57%. When the number of entities is less than 5, the number of entities whose relation exists account for 98.55%.

Therefore, based on the above research, in the selection of a relation feature, this paper do the following restrictions for context of the search window and rule extraction.

- Extract the verbs and nouns between the two entities as a relation feature.
- Select 2-3 nouns and verbs on the left of first entity e_1, and the distance from e_1 should be less than 4.
- Select 2-3 nouns and verbs on the left of second entity e_2, and the distance from e_2 should be less than 4.

According to the above-mentioned relation markers position restriction rule and context window method to extract the correlation feature, form the feature sequence and constitute the relation triples with the corresponding entity.

3.1K-means Clustering of Relation Triples

In order to facilitate the basic description of the algorithm, this paper first defines the following:

- R_n is a data set containing n data objects to be clustered.
- k is the number of target clusters, generally take $1 \le k \le n$.
- Initial clustering center set $C = (c_i \in R_n, 1 \le i \le k)$.
- Calculate the distance between objects or similarity function $Disp(x, y)$.
- The cluster C_{luci} represents a cluster centered on c_i, where $c_i \in C$.

In this paper, the following improvements are mainly made in the research process:

In the process of dealing with k value problem, the method of combining experience and search is presented to deal with the choice of the value of k, so that it can improve the selection of k value in clustering to a certain extent. As Algorithm 1:

Algorithm 1 The k value algorithm

Input: We evaluate the search range of k value, and select the intermediate value k_i as the initial value, take $f=0$; The clustering algorithm is used to calculate the convergence and classification effect by using the k_i as the value, denoted by Φ_i.

Process:

1: Repeat

2: Increase $d_k = \lambda \Delta k_i$ as a step value on k_i, denoted by $k_{i+1} = k_i + d_k$;

3: Test the clustering with k_{i+1}, count convergence and classification effect, denoted by Φ_{i+1};

4: if $\Phi_i < \Phi_{i+1}$ then

5: $f = f + 1$;

6: if $f > 1$ then

7: if $\Phi_i < \Phi_{i+1}$ then

8: $k = k_i$;

9: else

10: $k = k_{i+1}$;

11: else

12: $d_k = -d_k$;

13: Until select the k value successfully.

Output: The value of k

For the definition of the initial set C of clustering centers, this paper presents the density method to select the representative points. The density method which presented in this paper can improve the effect of clustering while avoiding the intensive problem of the central point in random selection. As Algorithm 2:

Algorithm 2 Representative point selection algorithm

Input: Data set $R_n = (x_1, x_2, \cdots, x_n)$.

Process:

1: Repeat

2: for $i = 1,2,\cdots,k$ do

3: With each sample $x_i (x_i \in R_n)$ as the center of the sphere, make a ball O_i with a radius of ξ;

4: Count the number of samples N_i fall within the range of O_i;

5: Compare the number of samples N_i of O_i and the number of samples N_j of $O_j (1 \le i \le n, 1 \le j \le n, i \ne j)$;

6: Select the maximum density of an O_m's sphere x_m as the first central point;

7: Select the next central point and radius $\xi > 0$, and ensure that the distance between x_i and x_m is greater than ξ;

8: end for

9: Until select out the center value of k.

Output: Center value

The last problem solved by K-means algorithm is the problem of isolated points. The introduction of isolated points usually causes the calculation deviation of cluster center, the accumulation of this deviation will seriously affect the final clustering effect.

In order to solve this problem, this paper introduces the isolated point elimination model based on the other related research. By calculating the distance between each point in the cluster $C_i (1 \le i \le k)$ and the mean square deviation of the cluster $C_i (1 \le i \le k)$, the model avoids the point where the absolute value is too large, the specific calculation model is as shown in equation(3.1), where the $Disp$ function is used to calculate the distance, and $\mu_i (1 \le i \le k)$ is the mean square deviation of cluster $C_i (1 \le i \le k)$.

$$D_i = Disp(x_i, c_i) - \mu_i \quad (1 \le i \le k) \tag{3.1}$$

4.Experiment

4.1Evaluation Metrics

The evaluation standard is MUC(Message Understanding Conference), it includes accuracy rate P(Precision), recall R(Recall) and the value of F(F-Measure). The calculation formula is as follows:

$$P=\frac{N_1}{N_2} \tag{4.1}$$

$$R=\frac{N_1}{N_3} \tag{4.2}$$

$$F=\frac{\left(\beta^2+1\right)PR}{R+\beta^2 P} \tag{4.3}$$

The accuracy rate P indicates the ratio of the correct result returned by the system to the total result. The recall rate R represents the ratio of the correct result returned by the system to the correct result that should be obtained, and the value of F is the criterion for evaluating the overall performance of a system, where N_1 is the number of instances the system judges correctly, N_2 is the total number of instances in the system that participate in the judgment, N_3 is the total number of instances in the data set, β is the weight of the recall rate relative to the accuracy rate which can be used to adjust the importance of both N_1 and N_2 in the value of F, we usually take $\beta=1$.

4.2Experimental Data

The experimental corpus comes from web crawler and artificial collection approximately 15000 talent information texts, of which 1/3 as the test data, 2/3 as training data. The system receives the results from the named entity recognition as shown in TABLE 1, selects the specific named entity to construct the entity and extract the characteristic words in the context that may represent the entity relation, and constructs the relation triples with entity. At last, deal with the final clustering result through the clustering algorithm.

Entity Type	**Form and Quantity**	
	The form in English	**Number**
person name	PER	16520
place name	LOC	37451
team name	TEAM	42638
time expression	TIME	60749
any other types of entities	GEN/CON/GRAD/DUTY…	954305
independent entity	NONE	13528

Table 1: The Named Entity Extraction Results In Corpus

The person's name as the center, this paper builds its entities with other types of entities, and the valid entity obtained are shown in TABLE 2. It can be seen from TABLE 2 that some of the entity which have been constructed are far beyond the actual needs of the number, such as the entity consist of PER-TIME. This article only requires data on the date of birth, and if the data is complete, it will be about 15,000. In fact, there is often a lot of time information in the text about learning work experience, which is not needed but will be extracted. And then

gradually screening and filtering in relation clusters. And some data may not be extracted because the text information is incomplete such as e-mail, telephone and so on. The clustering algorithm of relation extraction is based on the 11 named entity shown in TABLE 2, and the relevant feature vectors are extracted for each entity to form the relation triples.

Entity Type	**Form and Quantity**	
	The form in English	**Number**
person name - gender	PER-GEN	14428
person name - nation	PER-NATION	13669
person name - nationality	PER-CON	9638
person name - time	PER-TIME	58447
person name - place name	PER-LOC	33725
person name - team name	PER-TEAM	38465
person name - degree	PER-GRAD	14837
person name - post	PER-DUTY	20436
person name - discipline	PER-SUBJ	32418
person name - email	PER-EMAIL	8792
person name - telephone	PER-NUM	7094

Table 2: Effective Entity Extraction Results

4.3Experimental Design

Traditional K-means algorithm selects the value of k always using experience or random method, and initial center often appear isolated point problem. In order to solve the common problem, the k value of this paper is selected by the combination of experience and optimization algorithm. The selection of initial clustering center by density statistic method will solve the isolated point problem to a certain extent. In order to verify the feasibility of the improved algorithm, this paper also implements the general K-means algorithm as a comparison, the design as shown in TABLE 3.

Sequence Number	**Method**	
	Test group	**Control group**
test 1	The value of k is selected by experiment combines with the optimization method.	The k value is selected by random method. The k value is selected by empirical method.
test 2	The initial center is selected by density method.	The initial center is selected by random method.

Table 3: Relation Extraction Test Scheme

As shown in TABLE 3, test 1 selects the initial center randomly, and respectively uses the random method, the empirical method and the optimization method to select the value of k for the relation extraction under the situation of other conditions being equal. Basing on test 1, test

2 uses the empirical method and optimization method to select the value of k, and the initial clustering center is selected by random method and density method under the same condition.

4.4Experimental Results and Analysis

This paper takes person as the center to extract the entity relation, and the corresponding P-values and R-values are calculated after the relation is extracted by the various clustering methods. In this paper, we only select parts of the entity relation as a representative to calculate the extraction results of the algorithm.

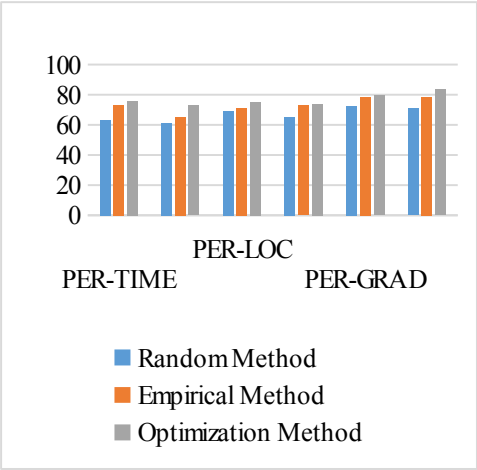

Figure 1: The result of P in test 1

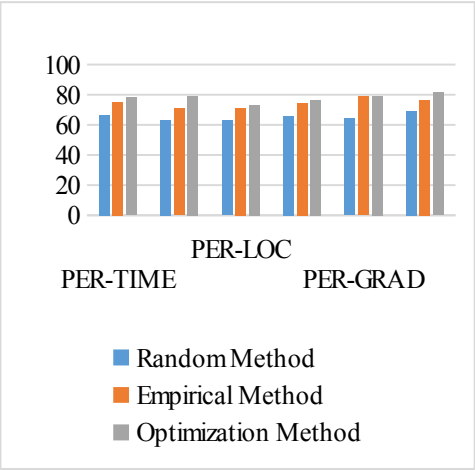

Figure 2: The result of R in test 1

In Figure 1 and Figure 2, the vertical axis represents the precision P and the recall R respectively, the unit is %, and the horizontal axis represents the entity relation type. It can be seen from the Figure 1 that the precision of choosing k value by random method is the worst, the precision of choosing k value by empirical method is in general, and the precision of choosing k value by optimization method is the best. It can be seen from Figure 2 that the recall of choosing k value by random method is the worst, the recall of choosing k value by empirical method is in general, and the recall of choosing k value by optimization method is the best. By comparing the values of P and R of each method in test 1, the results show that the selection of k value by random method obtains the worst effect and the practice statistics show that the effect is unstable. Although the selection of k value by empirical method is better for some types of entity relations, it may not apply to certain types of entity relations and without general and universal. The optimization method of k value can be achieved relatively good extraction effect, and the practice shows that it has good stability.

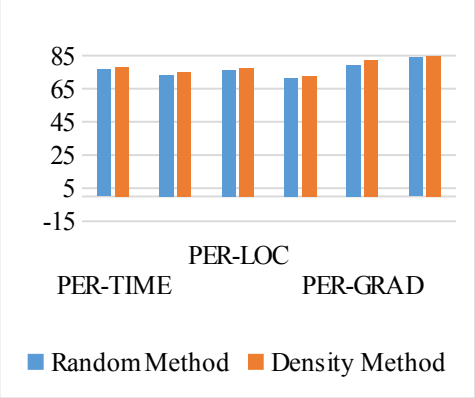

Figure 3: The result of P in test 2

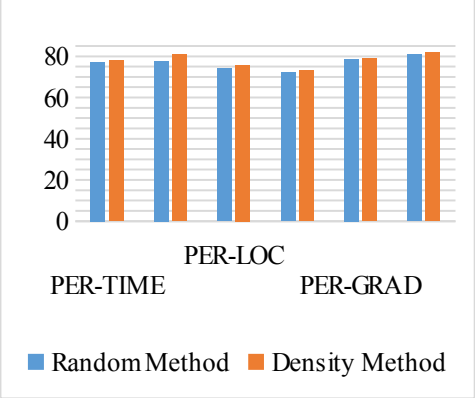

Figure 4: The result of R in test 2

In Figure 3 and Figure 4, the vertical axis represents the precision P and the recall R respectively, the unit is %, and the horizontal axis represents the entity relation type. It can be seen from the Figure 3 that the precision of selecting the initial center by random method is the worst, the precision of selecting the initial center by density method is the best. It can be seen from the Figure 4 that the recall of selecting the initial center by random method is the worst, the recall of selecting the initial center by density method is the best. By comparing the values of P and R of each method in test 2, the results show that the use of density method to select the initial center can effectively improve the precision and recall of the relation extraction.

The average of the precision and recall in each method are taken, and $\beta=1$. The results of test 1 and test 2 are shown in TABLE 4 and TABLE 5:

Test 1	**Result**		
	Precision	**Recall**	**F-Measure**
random method	66.85%	65.95%	66.39
empirical method	71.85%	74.85%	73.32
optimization method	78.15%	77.25%	77.69

Table 4: The Comparison Results Of Test 1

Test 2	**Result**		
	Precision	**Recall**	**F-Measure**
random method	77.45%	76.45%	76.95
density method	78.8%	77.35%	78.07

Table 5: The Comparison Results Of Test 2

The results of the two groups show that the improved K-means algorithm has a certain improvement effect compared with the traditional K-means algorithm. The value of P and F rise to a certain extent, and it is proved that the improvement of the selection of k value and center makes extraction effect much better in overall performance.

5.Conclusion

In this paper, we propose a method of entity relation extraction based on improved K-means clustering. Firstly, we complete a series of preprocessing work, such as coreference

resolution, the construction of the entity, relation feature extraction and so on, and then use the improved K-means clustering method to complete the named entity extraction. The value of k is selected by experience combines with the optimization method, and the initial center is calculated by the density method. The isolated point elimination model is used to solve the problem of clustering center deviation caused by the isolated points in the traditional K-means clustering algorithm. Finally, the improved algorithm is proved to ameliorate the overall performance of the entity relation extraction compared with the traditional K-means clustering algorithm. The next step in this paper is to further study and improve the various errors that exist in the current extraction results, and to find a more effective method to improve the performance of entity relation extraction.

References

[1] J. J. Mu, H. Bao. *Research on Chinese entity relation extraction[J]*. Computer Engineering and Design. 30(15), 3587-3590(2009) (In Chinese)

[2] Z. T. Zhang. *The Research of Relation Extraction with Unsupervised Method[D]*. Harbin: Harbin Institute of Technology(2007) (In Chinese)

[3] X. Y. Guo, T. T. He. *Survey about Research on Information Extraction[J]*. Computer Science. 42(2), 14-17(2015) (In Chinese)

[4] F. C. Liu, Z. N. Zhong, L. Lei, Y. Wu. *Entity Relation Extraction Method Based on Machine Learning[J]*. Ordnance Industry Automation. 32(9), 57-62(2013) (In Chinese)

[5] Y. Liu, J. W. Bi, Z. P. Fan. *A method for multi-class sentiment classification based on an improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm[M]*. Elsevier Science Inc., New York. pp, 38-52(2017)

[6] P. Chen, J. Y. Guo, Z. T. Yu, Y. T. Xian, X. Yan, S. C. Wei. *Chinese Field Entity Relation Extraction based on Convex Combination Kernel Function[J]*. Journal of Chinese Information Processing. 27(5), 144-149(2013) (In Chinese)

[7] S. Y. Liu, J. Zhou, B. C. Li, Y. Y. Xi, H. H. Tang. *Entity Relation Extraction Method Based on Multi-SVM-KNN Classifier[J]*. Journal of Data Acquisition and Processing. 30(1), 202-210(2015) (In Chinese)

[8] L. W. Chen, Y. S. Feng, D. Y. Zhao. *Extracting Relations from the Web via Weakly Supervised Learning[J]*. Journal of Computer Research and Development. 50(9), 1825-1835(2013) (In Chinese)

[9] M. Y. Wang. *Research on Chinese Open Entity Relation Extraction[D]*. Beijing: Beijing University of Posts and Telecommunications(2015) (In Chinese)

[10] C. Ma. *Using Improved Unsupervised Relation Extraction Method to Construct Traffic Ontology Based on Web[J]*. Computer System and Applications. 24(12), 273-276(2015) (In Chinese)

[11] S. Wu, M. F. Liu, H. J. Hu, Z. Q. Zhang, J. G. Gu. *Unsupervised Extraction of Attribute-Value Entity Relation from Chinese Texts[J]*. Journal of Wuhan University(Natural Science Edition). 62(6), 552-560(2016) (In Chinese)