Research Article **Open Access**

Jianhou Gan, Peng Huang, Juxiang Zhou, and Bin Wen*

Chinese open information extraction based on DBMCSS in the field of national information resources

https://doi.org/10.1515/phys-2018-0074
Received April 19, 2018; accepted June 12, 2018

Abstract: Binary entity relationship tuples can be applied in many fields such as knowledge base construction, data mining, pattern extraction, and so on. The purpose of entity relationship mining is discovering and identifying the semantic relationship. As the relationship between entities are different from the general domain, using supervise learning methods to extract entity relationships in the field of ethnicity is difficult. After research,we find that some words can be used in the context of a sentence to describe the semantic relationship. In order to salve the existing difficulties of building tagged corpus and the predefined entities-relationships model, this paper proposes a method of density-based multi-clustering clustering of semantic similarity (DBMCSS) to mine the binary entity relationship tuples from the Chinese national information corpus, which can extract entity relationships without a training corpus.

Keywords: Open information extraction, DBSCAN, Entity relationship

PACS: 07.05.Kf, 07.05.Mh

Jianhou Gan: Key Laboratory of Educational Informatization for Nationalities Ministry of Education, Yunnan Normal University, Kunming, China, Kunming, E-mail: ganjh@ynnu.edu.cn
Peng Huang: Key Laboratory of Educational Informatization for Nationalities Ministry of Education, Yunnan Normal University, Kunming, China, Kunming, E-mail: ganjh@ynnu.edu.cn
Juxiang Zhou: Key Laboratory of Educational Informatization for Nationalities Ministry of Education, Yunnan Normal University, Kunming, China, Kunming, E-mail: 406430159@qq.com
***Corresponding Author: Bin Wen:** College of Computer Science and Technology, Yunnan Normal University, Kunming, Yunnan Province, China, E-mail: wenbin@ynnu.edu.cn

1 Introduction

The research on information extraction (IE) is being developed into open information extraction, extracting open categories of entity relations and events from open domain text resources [1, 2]. Wu [3] uses Wikipedia infobox attribute values to construct training data, using the CRF method to choose relational words, then using the matching template to handle an unbounded set of semantic. This help the IE system get better results while also speeding more time. Shan [4] uses the shallow parsing theory to improve the IE system's quantity of information, chooses POS (part of speech) tagging and nominal phrase (NP) chunks instead of full parse trees to be the features, and chooses a logistic regression classifier. This will bring some new errors, such as "He lent me a book", the result is (He, lent, me), but as a matter of fact, it's a ternary relation task. Cafarella [5] uses an open information extraction system to extract entities from the web [6]. Yates [7] proposes a system named TextRunner for extraction tasks. Del and Gemulla [8] think of extract open information as a clustering problem, Xavier [9] extracts open information based on semantics.

Relation extraction is an important task in text mining [10], it can reflect the relationship between the named entities and help to find implicit knowledge in the substantial data and text [11]. We find that it can use some words in the context of a sentence to describe the semantic relationship. To solve the known difficulties and problems in the setup of a tagged corpus and predefine the entities-relationships model, this paper proposes a method of density-based multi-clustering of semantic similarity (DBMCSS) to extract the binary entity relationship tuples.

In recent years, supervised learning of relation-specific examples has huge adoption in information extraction systems, but the training data is very hard to obtain, especially in the field of ethnicity [12], and it hasn't built a training corpus and predefined the entities-relationship model. To solve this problem, we give an ex-

tract entity relation triples without a tagged corpus and pre-defined relation types. In the experiments, we have to perform better in precision of some relation type tasks than others.

2 The method

2.1 Named entity pair extraction

To extract Binary entity relationship tuples from texts. Firstly, we need to extract named entity pairs. Since the named entity pairs are between in a sentence and a group of sentences, we need to use the method of HMM to recognize the named entity [13]. Our task is to extract the named entity pair from a sentence. How to find the named entity pair from a sentence which includes many entities influences the precision. So we defined the entity set as follows:

Entity = (entity1, type1), (entity2, type2), (entity3, type3),...

We find that most of the entity pairs are neighboring by analyzing the distribution of entity pairs in a sentence. According to this principles, the entity pairs are selected. We defined the candidate entity pair set as follows:

Candidate Entity=<entity1,entity2>,<entity2, entity3><entity3,entity4>...

There is too much noise in the candidate entity pair set, so we used some rules to filter out the noise. The rules are used by us as follows:

Rule1: We observed that most of entity pair are not too far away, so we give a threshold to filter some candidate entity pair of noise;

Rule2: The entity pair are not the same if the entity pair exist relationship, so we can filter some noise;

Rule3: To statistic frequency of the relationship demonstrative between the entity pair type. We can filter some noise if the frequency is below a threshold value.

2.2 The contextual feature extraction of named entity pairs

We find some conditions by counting the entity pair distribution of the Chinese corpus. With the Chinese syntactic changeable, we need to simplify the condition, so we can mainly divide the condition into three types:

(1) The relationship demonstrative is between the entity pair, such as : "the ancient of Hmong firstly lived in the middle and lower reaches of Yellow River", the relationship demonstrative "lived in" is between the entity of "the ancient of Hmong" and "the middle and lower reaches of Yellow River".

(2) The relationship demonstrative is on the right of the entity pair, such as: "the swinging dance is the most influential large dance of Tujia ethnic minority", the relationship demonstrative "the large dance" is on the right of the entity pair of "the swinging dance" and "Tujia ethnic minority".

(3) The relationship demonstrative is on the left of the entity pair, such as: "As a leader, Mrs. Washi led the Zhuang ethnic minority people to resist the Japanese", the relationship demonstrative "leader" is on the left of the entity pair of "Mrs. Washi" and "the Zhuang ethnic minority people".

We chose the entity pair, the verb or noun between the entity pair, the verb or noun on the right of entity pair, the verb or noun on the left of entity pair, and the POS of all of the words as the feature for binary entity relationship tuples extraction.

2.3 Semantic similarity computation of feature vector

During the cluster analysis course, calculating the similarity of feature vector is important. We suppose that if entity pairs have the same relationship the feature vectors are more similar. For the N-dimensional feature vector to describe the binary entity relationship tuples we use the distance to measure the feature vector's similarity. We choose the Manhattan distance to describe the two feature's similarity. For the vector $V_1(v_{11}, v_{12}, \ldots, v_{1n})$, $V_2(v_{21}, v_{22}, \ldots, v_{2n})$, the Manhattan distance formula as follows:

$$D(V_t, V_j) = \sum_{k=1}^{m} |v_{tk} - v_{jk}| \quad (1)$$

We choose the TongYiCi CiLin Extended Edition dictionary written by the HIT to calculate the similarity of two words, the TongYiCi CiLin Extended Edition record 70000 entry [14]. The TongYiCi CiLin Extended Edition uses a five-layer classification system to describe the hierarchical relation of entries. Fig.1 is its hierarchical structure figure.

Such as: the word "Hmong" code is Di04B10#, "D" is the Level 1, "i" is level 2, "04" is level 3, "B" is level 4, "10" is level 5, and the "#" has other uses. Table 1 gives the word encoding rules.

The eighth encode has 3 labels, including the label "#", "=" and "@". The "#"represent the unequal, the "=" represents equation, and the "@" represents independent.

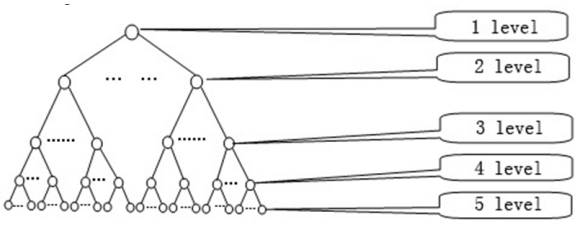

Figure 1: Monitoring Properties of Sensor NetworkThe hierarchical structure of TongYiCi CiLin Extended Edition.

Table 1: Description of sub-process monitoring indicators

Encode address	1	2	3	4	5	6	7	8
Example	D	i	0	4	B	1	0	m/# /@
Type	large	medium	little		word-group	little-word-group		
Level	1	2	3		4	5		

For the two words w_1 and w_2, the similarity evaluation formula is as follows:

$$Sim(w_1, w_2) = l_n \times cos(n \times \frac{\pi}{180})(\frac{n-k+1}{n}) + C \quad (2)$$

Because the density-based spatial clustering of application with noise (DBSCAN) method (see algorithm 1) has better performance to cluster the information from the noise, so we used this method to cluster the binary tuples about the entity relationship. We can use this clustering method to filter some noise and improve the accuracy of the entity relationship extraction. This method first defines two parameters: radius and density. If accord with these conditions, those binary entity relationship tuples will be clustered. We can find those sample points which are in keeping with these two conditions by constantly extending the search, clustering the information from the noise.

DBSCAN is a kind of clustering is the method-based on density. Density based methods considers clusters as dense region of objects that are different from lower dense regions in the data space. Density based regions are more appropriate and applicable in arbitrary shaped clusters but selection of attributes and selection of clusters with algorithms are more complex. It has the feature to merge two clusters that are sufficiently close to each other.

Density biased sampling, DBSCAN (Density Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify Clustering Structure), and DENCLUE (Density Clustering), and so forth are examples of this method[15–18]. DBSCAN requires two important parameters as follows [15, 16].

- Eps is the radius that represents spatial attribute (latitude and longitude) that delimitates the neighborhood area of a point.
- Minpts is the minimum number of points that must exist in the Eps-neighborhood.

Some concepts and definitions of DBSCAN (Density-Based Spatial Clustering of Application with Noise) which are directly and indirectly related to DBSCAN (Density-Based Spatial Clustering of Application with Noise) are explained here [15]:

(1) Cluster: in a database with given data objects as $D = \{O_1, O_2, \cdots, O_n\}$, the procedure of partitioning database D into smaller parts which are similar in certain standards as $C = \{C_1, C_2, \cdots, C_i\}$ is called clustering. $C_j^{'}s$ are clusters, where $C_j \leqslant D(j = 1, 2, 3, \cdots, i)$.

(2) Neighborhood: a distance function (e.g., Manhattan distance and Euclidean distance) for any two points p and q denotes $dist(p, q)$.

(3) Eps-neighborhood:the Eps-neighborhood (threshold distance) of a point p is defined by $\{q \in D | dist(p, q) \leqslant Eps\}$.

(4) Core object: a point p is a core point if at least Minpts points are within distance of it, and those points are said to be directly reachable from p . In other words, a core object is a point that its neighborhood of a given radius (Eps) has to contain at least a minimum number (Minpts) of other points.

(5) Directly density reachable: an object p is directly density reachable from the object q if is with in Eps-neighborhood of q and q is a core object in given data objects as $D = \{O_1, O_2, \cdots, O_n\}$.

(6) Density reachable: a point q is reachable from q if there is a chain $p_1 \cdots p_n$ with $p_1 = p$ and $p_n = q$, where each $p_i + 1$ is directly reachable from p_i with respect to Eps and Minpts, for $1 \leqslant i \leqslant n, p_i \in D$.

(7) Density connected: an object p is density connected to object q with respect to Eps and Minpts if there is an object $o \in D$ such that both p and q are density reachable from o with respect to Eps and Minpts.

(8) Density based clusters: a cluster c is nonempty subset of D satisfying the following "maximality" and "connectivity" requirements:

- p, q: if $q \in C$ and p is directly reachable from q with respect to Eps and Minpts, then $p \in C$.
- $p, q \in C$: p is density connected to q with respect to Eps and Minpts.

(9) Border objects: an object p is a border object if it is not a core object but density reachable from another core object.

(10) Noise: all points are not reachable from any other point, that is, neither a core point nor density reachable. $Noise = \{p \in D | \forall i : p\ notC_i\}$.

Some of the reasons why we have selected DBSCAN are its positive points as discussed in the following[16]:

- It is capable of discovering clusters with arbitrary shapes.
- There is no need to predict the number of clusters in advance and hence it is more realistic.
- There are greedy methods to replace R^*−tree data type greedy queries.
- Selection and application of attributes is always open to improve time and space complexity.
- It is robust to outliers and merging is possible with other clusters if they are similar.

It can be found that density reachable is the transitive closure of directly density reachable, and this relationship is asymmetric, and density connected is a symmetric relationship. The purpose of DBSCAN is to find the largest set of density connected objects.

Eg: hypothesis radius Eps=3, MinPts=3, there is point $\{m, p, p_1, p_2, o\}$ in the Eps-neighborhood of point p, there is point $\{m, p, q, m_1, m_2\}$ in the Eps-neighborhood of point m, there is point $\{q, m\}$ in the Eps-neighborhood of point q , there is point o, p, s in the Eps-neighborhood of point o, there is point $\{o, s, s_1\}$ in the Eps-neighborhood of point s,then there are the core object p, m, o, s(But q is not the core object because the number of points in the Eps-neighborhood of point q is equal to 2, less than MinPts=3). Point m is directed density reachable from point p because m is the Eps-neighborhood of point p, and p is the core object. Point q is reachable from point p density, because point q can reach direct density from point m, and point m can reach direct density from point p. Point q is density reachable from point p because point q can directed density reachable from point m and point m can directed density reachable from point p. Point q is density connected to point s because point q is density reachable from point p and point s is density reachable from point p.

Algorithm 1: DBSCAN

Input: E, MinPts, S
Output: Cluster

```
For o from 0 to length of S // iterate over the collection of S
If o==core objects then // if o is core object
R=find1(o)// find o's semantic density directly point
For r from 0 to length of R // iterate over the core object
  semantic //density directly point
Find2(r) //find point r maximum semantic density data set
```

In summary, the method of density-based multi-clustering of semantic similarity (DBMCSS) that we are proposed includes the following steps: (1) Named entities in the text are recognized; (2) The entity pairs in a sentence are selected according to the adjacent principles.This is to find out the entity pairs that have the relationship; (3) Using some rules to filter out the noise entity pairs; (4) Choosing the entity pair, the verb or noun between the entity pair, the verb or noun on the right of entity pair, the verb or noun on the left of entity pair to establish the entity relationship tuples; (5) The feature vectors of the arbitrary entity pairs in the sentences are established.We chose the POS of the words in step (4) as the feature for binary entity relationship tuples extraction; (6) Calculating the similarity of the feature vector by Manhattan distance; (7) Using the DBSCAN clustering algorithm to cluster the feature vector, then the entity relationship tuples are extracted.

3 Experiment

The corpus of experiments collected from the online source of Miao, WA, Tujia, Zhuang, Mongolian, Tibetan and other ethnic data comes from the Baidu encyclopedia of national introduction. The corpus size is 7.9 million words. This paper chooses the entity pair, the N verbs and nouns on the left of the entity pair, the N verbs and nouns between the entity pair, the N verbs and nouns on the right of the entity pair to be the feature, and the part-of-speech of the words as the eigenvectors of the relation extraction. The features include not only the word, but it also contains the part-of-speech. The method to calculate the similarity of a part of speech is as follows: if two words have the same part-of-speech then the value of similarity is 1, and if they are different then the value of similarity is 0. For the various attributes of the feature vector, through experimental analysis, for each value is weighted, the weight is between two entities similar weight value is 1, part of speech similar to the weight value of 0.3, entity of in the context of gerund similar weight of 1.5.

Part of the results of the computation of the similarity of the corpus are showed in Table 2. Some entity relationship extract experiment results are showed in Table 3.

The experiments are a combination of two methods by using and not using the semantic clustering method to extract from the corpus, through during the experimental observations this paper's set up field density is 5, the radius area is 2, then the experiments were conducted, and the results of the entity relationship extraction are showed in table 4 as follows:

Table 2: Entity pair similarity

Entity pair 1	Entity pair 2	Semantic similarity
Peach/ns and Grass/nr	Zhijin County/ns and Hmong/nz	1.840
Peach /ns and Grass/nr	Qiandongnan/ns and Goulang/nr	3.088
Zhaotong/ns and antiphonal singings/nr	Goulang/nr and Guangxi/ns	0.500
Anshun/ns and Hmong/nz	Zhijin County/ns and Hmong/nz	3.100
Yilang/nr and Hmong/nz	Zhang Xiumei/nr and Qiannan/ns	2.271
Hmong/nz and Miao surname/nr	Zhaotong/ns and antiphonal singings/nr	1.200

Table 3: The result of entity relation extraction

Sentence	Entity1	Entity2	Entity2
Part of Hmong migrated to Southeast Asian countries	Hmong	Southeast Asian	migrated to
The Hmong language belong to Sino-Tibetan ethnic languages Hmong	Hmong language	Sino-Tibetan	belong to
From Genghis Khan unified Mongolia into before and after the yuan dynasty established	Genghis Khan	Mongolia	unified
Chiyou first defeat Emperor Yan	Chiyou	Emperor Yan	defeat
Chiyou led by Hmong people fought bravely	Chiyou	Hmong	led
Ethnic Zhuang residential area is located in the South of the Five Ridges	Ethnic Zhuang	South of the Five Ridges	residential area

Table 4: Accuracy comparison of relation extraction with two methods

Entity relationship type	Accuracy of no clustering	Accuracy of using clustering
organization—toponymy	75.0%	78.0%
toponymy—organization	21.5%	37.0%
organization—organization	57.1%	75.6%
organization—character	67.0%	72.2%
Ycharacter—organization	60.0%	76.0%
average correct rate	56.1%	67.7%

The results of the Experiment, using the DBSCAN method to clustering the candidate entity relationship, clearly show a filtering of the noise from the candidate entity relationship. Finally, by comparing the two methods, we find it improves the accuracy of entity relationship extraction if we use the DBSCAN cluster method to filter out the noise.

4 Discussion and conclusion

In this paper, we introduce an new extract entity relation triples method that is density-based of multi-clustering semantic similarity (DBMCSS). Due to the effect of word segmentation and named entity recognition there is a direct impact on the effect of relation extraction, the extraction of entity relationships will influence the improvement of the accuracy of word segmentation and named entity recognition. This paper used single layer clustering to filter the noise, so if we were to use a multi-cluster method the results may better than this. In the experiment, we found the computation speed was too slow, so in our future work, we considered using distributed computing to solve the computation speed problem. In the future, we will carry out the following research works: firstly, the method of entity relation extracting is improved to improve the accuracy and recall rate of mining. In addition, there is the application of the relation extracting results in knowledge services and intelligent retrieval.

Acknowledgement: The research is supported by a National Nature Science Fund Project with Nos.61562093, Nos.61661051, and Key Project of Applied Basic Research Program of Yunnan Province Nos. 2016FA024, and Youth

Project of Applied Basic Research of Science and Technology Plan in Yunnan Province Nos. 2016FD022.

References

[1] Duc-Thuan V., Ebrahim B., Open Information Extraction, Encyclopedia with Semantic Computing & Robotic Intelligence, 2016, 1, 1, 1-4.

[2] Gotti F., Philippe L., From French Wikipedia to Erudit: A test case For Cross-domain open information extraction, Comp. Intel., Special Issue of Computational Intelligence based on the 29th Canadian Conference on Artificial Intelligence, 2017, 1-20.

[3] Wu F., Weld D.S., Open information extraction using Wikipedia, Proceedings of the Meeting of the Association for Computational Linguistics, 2010, 118-127.

[4] Shan S.R., Xie J.H., Identifying Relations for Open Information Extraction, Conference on Empirical Methods in Natural Language Processing, Assoc. Comp. Linguistics, 2011, 1535-1545.

[5] Banko M., Cafarella M.J., Soderland S. et al., Open information extraction from the web, Int. Joint Conference on Artificial Intelligence, 2007, 2670-2676.

[6] Zhila A., Gelbukh A., Open Information Extraction from real Internet texts in Spanish using constraints over part-of speech sequences: Problems of the method, their causes, and ways for improvement, Revista Signos, 2016, 49,119-142.

[7] Yates A., Cafarella M., Banko M. et al., Text Runner: open information extraction on the web, Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, 2007, 25-26.

[8] Del C. L., Gemulla R., Claus IE: Clause-based open information extraction, 2013 International Conference on World Wide Web, 2013, 355-366.

[9] Xavier C.C., Lima V.D., Souza M., Open information extraction based on lexical semantics, J. Brazilian Comp. Soc., 2015, 21, 1, 1-14.

[10] Gamallo P., An overview of open information extraction, 3rd Symposium on Languages, Appl. Technol. (SLATE'14), 2014, 13-16.

[11] Gotti F., Langlais P., Harnessing Open Information Extraction for Entity Classification in a French Corpus, Adv. Artificial Intel., 2016, 150-160.

[12] Jain S., Tumkur K. R., Kuo T.T. et al., Erratum to: Weakly supervised learning of biomedical information extraction from curated data, BMC Bioinformatics, 2016, 17, 1, 84.

[13] Morwal S., Jahan N., Chopra D., Named Entity Recognition using Hidden Markov Model (HMM), Int. J. Nat. Lang. Comp., 2012, 1(1), 15-23.

[14] Liang CH.X., Shao Y.Q., Zhao J., Construction of a Chinese Semantic Dictionary by Integrating Two Heterogeneous Dictionaries: TongYiCi Cilin and HowNet, 2013, Int. Joint Confer- ences on Web Intelligence, 2013, 203-207.

[15] Muetzelfeldt R., Duckham M., Dynamic Spatial Modeling in the Similie Visual Modeling Environment, chapter 17, 2005, John Wiley Sons, New York.

[16] Sharma A., Gupta R.K., Tiwari A., Improved Density Based Spatial Clustering of Applications of Noise Clustering Algorithm for Knowledge Discovery in Spatial Data, Mathematical Problems in Engineering, 2016, 1564516, 1-9.

[17] Locke J.B., Peter A.M., Multiwavelet density estimation, Appl. Math. Comp., 2013, 219(11), 6002-6015.

[18] Culbertson J., Guralnik D.P., Stiller P.F., Functorial hierarchical clustering with overlaps, Discrete Applied Math., 2018, 236, 108-123.