International Journal of Database Theory and Application
Vol.9, No.6 (2016), pp.151-160
http://dx.doi.org/10.14257/ijdta.2016.9.6.15

Simultaneous Entities and Relationship Extraction from Unstructured Text

Jingtai Zhang, Jin Liu and Xiaofeng Wang

College of Information Engineering, Shanghai Maritime University, 201306 Shanghai, China
{jtzhang, jinliu, xfwang}@shmtu.edu.cn

Abstract

Entity recognition and entity relationship extraction are two very important tasks in information extraction. Most research work in the literature treats these two work independently when processing the text. This paper proposes a novel method for performing entity recognition and entity relationship extraction simultaneously from unstructured text based on Conditional Random Fields (CRFs). This method makes use of entity features, entity relationship features and features of the triples which is composed of entities and their relationship to conduct the model training. Experiment results show that this method can recognize entity and extract entity relationship effectively.

Keywords*: entity recognition, entity relationship extraction, CRFs*

1. Introduction

There exists a vast amount of unstructured text on the Web, including newswire, blogs, email communications, governmental documents, chat logs, and so on. When faced with that amount of information, how could a person be helped to understand all of the data? A popular idea is to turn unstructured text into structured form by manual annotation. Instead, we would like to have a computer annotate all data with the structure of our interest. The idea is that we first annotate the entities over the unstructured text. Then, we are interested in relations between entities, such as person, organization, and location. Examples of relations are person-affiliation and organization-location. Current state-of-the-art named entities recognizers (NER), such as Stanford NLP [14] and BANNER [15], can automatically label data with high accuracy. [13]

Chinese named entity recognition is to identify a specific entity from the Chinese text. It is the basis of information extraction, machine translation, automatic question answering and other natural language processing technology. In the field of entity recognition, many research institutions have made outstanding achievement in the recognition of English entities. However, due to the restrictions of Chinese characters, Chinese named entity has been very difficult to be recognized. Therefore, it is very important to do the study of Chinese named entities recognition in order to promote the development of other technologies and applications.

Chinese entity relationship extraction is also an important task of information extraction. Entity relationship identification was first proposed in 1998 at the MUC conference, the main task is to determine the semantic relationship between the two entities. However, the whole relation extraction process is not a trivial task. The computer needs to know how to recognize a piece of text having a semantic property of interest in order to make a correct annotation. Thus, extracting semantic relations between entities in natural language text is a crucial step towards natural language understanding applications.

In this paper, we perform both the entity recognition and the entity relationship extraction in the same time, and treat these two tasks as a classification task. Based on the above analysis, this novel method utilizes conditional random field [1-2] to recognize

ISSN: 2005-4270 IJDTA

entity and extract relationships among the recognized entities. By constructing the probabilistic graph model, this method can be used to recognize the entities and relationships in the same time. The remaining of this paper is organized as follows, the second section gives a brief review about the research of the entity and relationship recognition. The third section introduces our novel method, and the fourth section presents the experimental result and analysis.

2. Related Work

2.1. Entity Recognition

There are many different methods proposed in the literature to perform entity recognition and entity relationship extraction. However, they treated these two tasks as independent ones.

Regarding the term "Named Entity", the word "Named" restricts the task to those entities for one or many rigid designators which stands as referent. In data mining, a named entity is a word or a phrase that clearly identifies one item from a set of other items that have similar attributes. Usually, rigid designators include proper names, but it depends on domain of interest that may refer the reference word for object in domain as named entities. For these entities, Hidden Markov Model is widely used to perform the recognition work. HMM is a generative model. The model assigns the joint probability to paired observation and label sequence. Then the parameters are trained to maximize the joint likelihood of training sets. It is advantageous as its basic theory is elegant and easy to understand. Hence it is easier to implement and analyze. It uses only positive data, so they can be easily scaled. In order to define joint probability over observation and label sequence HMM needs to enumerate all possible observation sequence. Wang [9] proposed a method for named entity recognition for short text based on HMM.

Conditional Random Field is a type of discriminative probabilistic model. It has all the advantage of MEMMs without the label bias problem. CRFs are an undirected graphical model (also known as random field) which is used to calculate the conditional probability of values on assigned output nodes given the values assigned to other assigned input nodes.

Zhang [10] proposed a method extract opinion target and polarity on iterative two-stage CRF model, and the two CRF model reached an F-score of 0.505 on the COAE2014 evaluation data.

Li [11] proposed a method for medical named entity recognition using combining CRF and rule. The algorithm made initial entity recognition by CRF and then applied a rule based recognition method to improve the accuracy, whose rules included the rules from decision tree and domain knowledge. The results show that the algorithm has high accuracy and recall performance at records entity recognition that is up to 91.03% and 87.26%.

2.2. Entity Relationship Extraction

Many researchers present the task of entity relationship extraction as a classification task. Given a sentence $S = w_1, w_2, ..., e_1..., w_j, ...e_2..., w_n$, where e_1 and e_2 are the entities, a mapping function f can be given as

$$f_R(T(S)) = \begin{cases} +1 & \text{if } e_1 \text{ and } e_2 \text{ are related according to relation R} \\ -1 & \text{Otherwise} \end{cases} \quad (1)$$

Where T(S) are features that are extracted from S. Essentially the mapping function f decides whether the entities in the sentence are in a relationship or not. Put in another way, the task of entity-relationship extraction becomes that of entity-relationship detection. If a labeled set of positive and negative relationship examples are available for training, the function f can be constructed as a discriminative classifier like Perceptron, Voted Perceptron or Support Vector Machines (SVMs). These classifiers can be trained using a set of features selected after performing textual analysis. Depending on the nature of input to the classifier training, supervised approaches for relationship extraction are further divided into feature based methods. Here is a sample of "Bag of features kernel" method. The sentence "In 1975, Gates and Paul Allen co-founded the Microsoft". Given that Gates and Paul Allen are the named entities, the words co-founded indicate a person-organization relationship between the two entities. Thus we can conclude that the context around the entities under question can be used to decide whether they are related or not.

Qin [4] proposed a method of unsupervised Chinese open entity relationship extraction by using the relationship between demonstratives description method to solve the pre-defined relationship type system. In the PER-PER relationship between the words of the experiment, the average F-measure reached 64.25%.

Recently, Deep Belief Networks (DBN) were reported to be used in the task s of nature language processing and obtained satisfactory results. Liu [5] proposed a named entity relationship extraction based on the positive and negative training SVM. Chen [6] proposed a method of extracting the entity relationship based on DBN [8].

DBN represents a network of Restricted Boltzmann Machines (RBM). In contrast CRF is a class of statistical modeling method. In this method, the characters of the entity, the entity type, the relative position between the entities are easy to be extracted, and the accuracy is not affected by the lexical analysis. The difference between CRF and DBN is that DBN is more suitable for image recognition while CRF is applied to a variety natural language processing tasks.

3. Simultaneously Entity and Relationship Extraction (SERE)

3.1. The Overall Process

First, large amount of unstructured texts are obtained from the Internet to be the training and validation data set. Then, we tag all the texts to build the training file. For the sake of simplicity and clarity, we restrict our objective to binary relations between two entities. Therefore, we use a training tool which is called CRF++ to train the model. Last, we can use the trained model to identify entities and relationships.

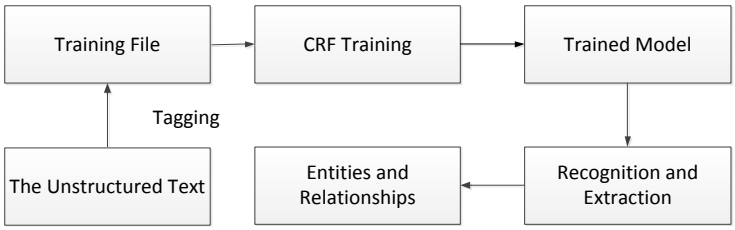

Figure 1. The Overall Process

3.2. Recognition Principle

Different from the aforementioned methods for entity recognition and relationships extraction, our method performs these two tasks simultaneously in the extraction process. To better illustrate our method, we only focus the relationship between the two entities in one sentence without considering the relationship between the entities cross the sentences. Then we can simplify the task of dealing with them. Given a sentence, the task is to recognize the entities and the relationship in it. The following are the detailed steps in this process:

Step 1: Tagging all the entities in the sentences.

Step 2: Tagging all the relationships in the sentences.

Step 3: If there are both entities and relationships which are tagged before, we can tag the triple symbol on them.

The specific tagging features are shown in figure:

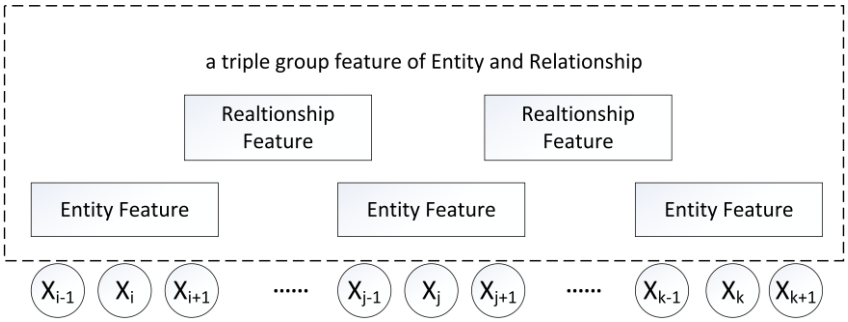

Figure 2. Figure of Tagging Features

In Figure 2, X (the unstructured text) is a random variable over data sequences to be tagged, and Y (entity feature, relationship feature, feature of triple group) is a random variable over corresponding label sequences.

Then we convert these tag features into variables of conditional random field in the model. Conditional Random Fields (CRFs) is presented by Lafferty, McCallum and Pereira [1]. It can be used as a framework for building probabilistic models to segment and label sequence data. $X = (X_1, X_2, \cdots, X_n)$ is a random variable over data sequences (the unstructured text) to be labeled, and $Y = (Y_1, Y_2, \cdots, Y_n)$ (entity feature, relationship feature, feature of triple group) is a random variable over corresponding label sequences. A CRF is an undirected graphical model whose nodes can be divided into exactly two disjoint sets X and Y, the observed and the output variables, respectively, the conditional distribution $P(Y \mid X)$ is then modeled. For entity and relation extraction tasks, the train model can be simplified to a first-order linear-chain CRFs model. Based on the above theory, we can define $P(Y \mid X)$ as an ordinary linear-chain CRF:

$$P(\mathrm{y} \mid \mathrm{x}) = \frac{1}{Z(\mathrm{x})}\left(\sum_{i,k} \lambda_k t_k(\mathrm{y}_{i-1}, \mathrm{y}_i, \mathrm{x}, \mathrm{i}) + \sum_{i,l} \mu_i s_l(\mathrm{y}_i, \mathrm{x}, \mathrm{i})\right) \tag{2}$$

$$Z(\mathrm{x}) = \sum_{y} \exp\left(\sum_{i,k} \lambda_k t_k(\mathrm{y}_{i-1}, \mathrm{y}_i, \mathrm{x}, \mathrm{i}) + \sum_{i,l} \mu_i s_l(\mathrm{y}_i, \mathrm{x}, \mathrm{i})\right) \tag{3}$$

By the formula (1), it can be seen that the linear chain of conditional random field is a log-linear model. Where, t_k and s_l is the characteristic function, λ_k and μ_l is the corresponding weight function of each feature value, $Z(x)$ is a normalized factor for

probability constraints to make it satisfy the probability axioms. The sum operation is performed on all possible output sequences. So, linear chain conditional random field is completely determined by the characteristic function t_k , s_l and the corresponding weights λ_k and μ_l . Therefore, SERE consists of the following three steps:

1) Feature extraction. Extract the features from the text to determine the characteristic function. Here, the main feature is the set of three kinds of features described above.
2) Parameter estimation. Use the selected features to train the model to calculate the weights λ_k and μ_l .
3) Results labeled. Input the test data, using the trained CRF model to do entity recognition and relation extraction tasks.

One important step is to estimate the parameters. In CRF model, the characteristic function is divided into the transfer characteristic function $t_k(\mathrm{y}_{i-1}, \mathrm{y}_i, \mathrm{x}, \mathrm{i})$ and the state function. $s_i(\mathrm{y}_i, \mathrm{x}, \mathrm{i})$. In this paper, we define the t_k :

$$(\mathrm{y}_{i-1} = \mathrm{state}1, \mathrm{y}_i = \mathrm{state}2, \mathrm{x}_w = \mathrm{text}) \tag{4}$$

Where, y_{i-1} indicates the current position of the previous tag inspection location, y_i indicates the current position of the current tag inspection location. state1, state2 values should be obtained from the tag symbols. Given m transfer characteristic and n state characteristic, $K = m + n$ so the characteristic function is as follows:

$$f_k(\mathrm{y}_{i-1}, \mathrm{y}_i \mathrm{x}, \mathrm{i}) = \begin{cases} t_k(\mathrm{y}_{i-1}, \mathrm{y}_i \mathrm{x}, \mathrm{i}) & k = 1, 2, \ldots m \\ s_l(\mathrm{y}_i \mathrm{x}, \mathrm{i}) & k = m + l; l = 1, 2, \ldots \mathrm{m} \end{cases} \tag{5}$$

Then, we use log-likelihood objective function to estimate the parameters:

$$L(\lambda) = \sum_{x,y} \tilde{p}(x, y) \sum_{i=1}^{n} \left(\sum_j \lambda_j f_j(\mathrm{y}_{i-1}, y_i, x, i) \right) - \sum_x \tilde{p}(x) \log Z(x) \tag{6}$$

And by formula (7), Derivation of λ_i can be obtained:

$$\frac{\partial L(\lambda)}{\lambda_j} = \sum_{x,y} \tilde{p}(x, y) \sum_{i=1}^{n} f_j(y_{i-1}, y_{i,x)} - \sum_{x,y} \tilde{p}(x, y) \mathrm{p}(\mathrm{y} \mid \mathrm{x}, \lambda) \sum_{i=1}^{n} f_j(y_{i-1}, y_{i,x}) \tag{7}$$

We can calculate the λ_i by L-BFGS [18] iterative algorithm effectively. Finally, we can use the model with parameter λ_i to do entity recognition and relationship extraction. Based on the principle, we propose an algorithm SERE.

Algorithm1 SERE

1: **Input:** a set of unlabeled data X and the corresponding symbols of tagging Y

2: **Repeat:**

Train classifier C on Y

Calculate the probability P(y|x)

Calculate the parameter λ_i according to P(y|x)

Use algorithm L-BFGS to converge the model

Until convergence criteria is reached

L-BFGS shares many features with other quasi-Newton algorithms, but is very different in how the matrix-vector multiplication for finding the search direction is carried

out $d_k = -H_k g_k$. There are multiple published approaches using a history of updates to form this direction vector. Here, we give a common approach, the so-called "two loop recursion."[21]

We'll take as given x_k , the position at the k iteration, and $g_k = \nabla f(x_k)$ where f is the function being minimized, and all vectors are column vectors. We also assume that we have stored the last m updates of the form of

$$s_k = x_{k+1} - x_k \tag{8}$$

And

$$y_k = g_{k+1} - g_k \tag{9}$$

We'll define $\rho_k = \frac{1}{y_k^T S_k}$, and H_k^o will be the 'initial' approximate of the inverse Hessian that the estimate at iteration k begins with. Then we can compute the direction as follows:

Algorithm2 L-BFGS[18]

$q = g_k$

For $i = k-1, k-2, ..., k-m$

$\alpha_i = \rho_i s_i^T q$

$q = q - \alpha_i y_i$

$H_k^o = y_k^T s_{k-1} / y_{k-1}^T y_{k-1}$

$z = H_k^o q$

For $i = k-m, k-m+1, ..., k-1$

$\beta_i = \rho_i y_i^T z$

$z = z + s_i(\alpha_i - \beta_i)$

Until $H_k g_k = z$

3.3. Samples of Tagging

The unstructured text must consist of multiple tokens. We define some symbols to tag the text, such as Date (Date entity), Per (person entity), Org (Organization entity), Founder (ORG-Affiliation relationship) and all the tags represented in IOB2 format. And the third symbol “Tri” means that entities and relationship can form a (entity, relationship, entity) triples. Then, we use these symbols (according to ACE05 [3] annotation guidelines) to tag the text sequence.

Table 1. Tagging Samples

The unstructured text(X)	Tagging symbol(Y)
In	N
1975	B-Date
,	O
Gates	B-Per-Tri

and	O
Paul	B-Per-Tri
Allen	I-Per-Tri
co-founded	B-Founder-Tri
Microsoft	B-Org-Tri
.	O

The B- prefix before a tag indicates that the tag is the beginning of a chunk, and an I- prefix before a tag indicates that the tag is inside a chunk. The B- tag is used only when a tag is followed by a tag of the same type without O tokens between them. An O tag indicates that a token belongs to no chunk. A Tri-suffix after a tag indicates that the tag can form a triple composed of entities and relationship with the near tags which have a same Tri-suffix. Through the above analysis, it can be found that SERE treats the two tasks that entity recognition and extraction of their relationships as one problem. So we can recognize entities and their relationship by SERE easily.

4. Experiment

4.1. Data Preparation and Experiment Environment

We gathered more than 10000 news paragraphs in Blue Net Shipping News [7] as a training set. And all the text are tagged using the method according to Section3.3. In order to investigate the classification results in different size of training text, the text set (Corpus) is divided into different proportions, the proportion of training text is 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, respectively, and the rest is the proportion of testing text.

4.2. Performance Evaluation Metrics

In the supervised methods setting, entity recognition and relationship extraction are expressed as classification tasks and hence, metrics like Precision, Recall and F-Measure are used for performance evaluation. These metrics are as follows:

$$\text{Precision } P = \frac{\text{Number of correctly extracted entities or entity relations}}{\text{Total number of extracted entities or entity relations}} \quad (10)$$

$$\text{Recall } R = \frac{\text{Number of correctly extracted entities or entity relations}}{\text{Actual number of extracted entities or entity relations}} \quad (11)$$

$$\text{F-Measure } F1 = \frac{2PR}{P+R} \quad (12)$$

4.3. Experiment Result

In the experiments, we find that taking 80% data as training and taking 20% corpus as test corpus can get the best result. The recognition rate of the names is 90%, the recognition rate is 87%, and the recognition rate is 85%. Located in the relationship recognition rate is 96%, the employment relationship identification rate is 93%, close to the relationship recognition rate is 80%.

Table 2. Results

Type	P/%	R/%	F/%
Person	89.92	88.68	89.30
Location	86.97	86.33	86.65
Organization	88.10	87.49	88.21

Relationship of Located	94.02	89.04	91.65
Relationship of Employ	95.52	93.10	94.30

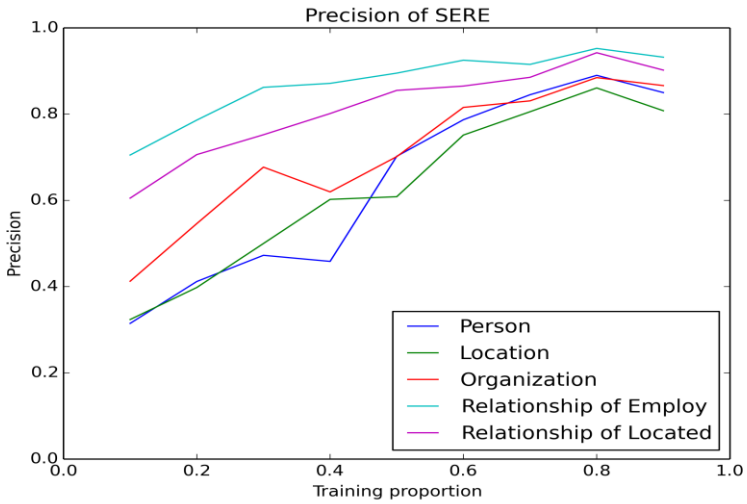

Figure 3. Precision of SERE

4.4. Conclusion

In this paper, we had presented a novel method to extract named entities and their relationships from unstructured text. To the best of our knowledge, the presented work is the first one to combine the two tasks with synthesizing feature set. The experiment results show that SERE can effectively extract entities and their relationships from unstructured text corpus. Users can make full use of SERE to perform information extraction in different topic domains. And this method can also be used in other information extraction system based on CRFs. Conditional Random fields can offer a unique combination of properties and discriminatively trained models for sequence segmentation and labeling by the combination of arbitrary, overlapping and agglomerative observation features from both the past and future. However, this method still can be improved in many ways, such as the capacity to discrete the relationships in more complicated context.

In the future, we will try to obtain more experiment results with different corpus and improve this method to make it applicable to more complex context scenarios. And we would consider some different methods to improve the results.

Acknowledgements

This work was supported by Shanghai Municipal Sci. &Tech. Commission “Science and Technology Innovation Action Plan” project (14511107400), and by Shanghai Maritime University research fund project (No. 20130469), and by State Oceanic Administration China research fund project (201305026).

References

[1] J. Lafferty, A. McCallum and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", **(2001)**.
[2] C. Sutton and A. McCallum, "An introduction to conditional random fields for relational learning", Introduction to statistical relational learning, **(2006)**, pp. 93-128.
[3] ACE, "Chinese annotation guidelines for entities." ACE2005, **(2005)**.
[4] Q. Bing, L. An'an and L. Ting, "Unsupervised Chinese Open Entity Relation Extraction", Journal of Computer Research and Development, vol. 52, no. 5, **(2015)**, pp. 1029-1035.
[5] L. Lu, L. B. Cheng and Z. X. Fei, "Named entity relation extraction based on SVM training by positive and negative cases", Computer Applications, vol. 28, no. 6, **(2008)**, pp. 1444-1446.
[6] C. Yu, Z. D. Quan and Z. T. Jun, "Chinese Relation Extraction Based on Deep Belief Nets", Journal of Software, vol. 23, no. 10, **(2012)**, pp. 2572-2585.
[7] Blue Net Shipping News, http://www.bluehn.com/.
[8] G. E. Hinton, "Deep belief networks", Scholarpedia, vol. 4, no. 5, **(2009)**, pp. 804-786.
[9] W. Dan and F. X. Hua, "Named entity recognition for short text", Journal of Computer Applications, vol. 29, no. 1, **(2009)**, pp. 0143-03.
[10] Z. Sheng and L. Fang, "Opinion Target and Polarity Extraction Based on Iterative Two-Stage CRF Model", Journal of Chinese Information Processing, vol. 29, no. 1, **(2015)**, pp. 163-169.
[11] L. Wei, Z. D. Zhe, L. Bao, P. X. Ming and L. J. Ren, "Combining CRF and the rule based medical named entity recognition", Application Research of Computers, vol. 32, no. 4, **(2015)**, pp. 1082-1086.
[12] N. David and S. Sekine, "A survey of named entity recognition and classification", Lingvisticae Investigations, vol. 30, no. 1, **(2007)**, pp. 3-26.
[13] B. Nguyen and S. Badaskar, "A survey on relation extraction", Language Technologies Institute, Carnegie Mellon University, **(2007)**.
[14] S. Rahul, "Named Entity Recognition: A Literature Survey", **(2014)**.
[15] T. K. Sang, F. Erik and F. D. Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition", Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003. Association for Computational Linguistics, vol. 4, **(2003)**.
[16] Stanfor NLP http://nlp.stanford.edu/software/CRF-NER.shtml.
[17] BANNER http://banner.sourceforge.net/ .
[18] C. Zhu, R. H. Byrd, P. Lu and J. Nocedal, "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization", ACM Transactions on Mathematical Software (TOMS), vol. 23, no. 4, **(1997)**, pp. 550-560.
[19] Z. G. Dong and J. Su, "Named entity recognition using an HMM-based chunk tagger", proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, **(2002)**.
[20] Z. Shaojun, "Named entity recognition in biomedical texts using an HMM model", Proceedings of the international joint workshop on natural language processing in biomedicine and its applications. Association for Computational Linguistics, **(2004)**.
[21] G. N. Stephen and J. Nocedal, "A numerical study of the limited memory BFGS method and the truncated-Newton method for large scale optimization", SIAM Journal on Optimization vol. 1, no. 3 **(1991)**, pp. 358-372.

Authors

Jingtai Zhang, (1991), male, born in Jiangsu, master candidate, specialized in knowledge graph and natural language processing research.

Jin Liu, (1975), male, born in Sichuan, associate professor, specialized in web data mining, NLP and software engineering.

Xiaofeng Wang, (1958), male, born in Liaoning, professor, doctoral advisor, specialized in data mining and artificial intelligence.

