IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Automatic Non-Taxonomic Relation Extraction from Big Data in Smart City

Jing Qiu[1], Yuhan Chai[1,2], Yan Liu[2], Zhaoquan Gu[1], Shudong Li[1], Zhihong Tian[1*]

[1]Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou 510006, China

[2] School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050026, China

Corresponding authors: Zhihong Tian (e-mail: tianzhihong@gzhu.edu.cn).

This work is supported in part by the National Natural Science Foundation of China (61871140, 61572153, U1636215, 61572492, 61672020), the National Key research and Development Plan (Grant No. 2018YFB0803504), and Open Fund of Beijing Key Laboratory of IOT Information Security Technology (J6V0011104). We thank HIT-SCIR for providing us with LTP modules.

ABSTRACT The explosive data growth in smart city is making domain big data a hot topic for knowledge extraction. Non-taxonomic relations refer to any relations between concept pairs except the is-a relation, which is an important part of Knowledge Graph. In this paper, toward big data in smart city, we present a multi-phase correlation search framework to automatically extract non-taxonomic relations from domain documents. Different kinds of semantic information are used to improve the performance of the system. Firstly, inspired by the works of network representation, we propose a Semantic Graph Based method to combine structure information of semantic graph and context information of terms together for non-taxonomic relationships identification. Secondly, different semantic types of verb sets are extracted based on dependency syntactic information, which are ranked to act as non-taxonomic relationship labels. Extensive experiments demonstrate the efficiency of the proposed framework. The F1 value reaches 81.4% for identification of non-taxonomic relationships. The total precision of the non-taxonomic relationship labels extraction is 73.4%, and 87.8% non-taxonomic relations can be provided with "good" labels. We hope this article can provide a useful way for domain big data knowledge extraction in smart city.

INDEX TERMS Non-taxonomic relations, Semantic Graph, Dependency relations, Smart City

I. INTRODUCTION

With the rapid development of the smart applications, a huge amount of data is gathered from various sources every day [40-43, 46]. Particularly, a huge amount of unstructured text information is produced with the prevalence of mobile devices, social media, the Internet, and so on. This explosive data growth is making a challenge to extract useful information from big data.

Nowadays, with the development of the research works and applications of Deep Learning [39], Knowledge Graph is one of the most popular knowledge representation methods in the field of big data. The ontology defines the data pattern of the knowledge graph, and the results of the ontology construction research assist the construction of the knowledge graph to a large extent. Ontology learning (OL) aims for automatic or semi-automatic ontology construction, which can conserve much time and resources compared to manual ontology building. Several subtasks are included: identifying the domain concepts, extracting taxonomic relations and non-taxonomic relations between concepts, and extracting axioms on the relations. Extracting the non-taxonomic relations is considered one of the most difficult tasks.

This paper proposes an automatic framework for the extraction of non-taxonomic relations from domain big data, with web news texts as the source information. Two tasks are considered to be performed for the discovery of the non-taxonomic relations: identification of non-taxonomic relationships; and identification of labels for the relations. For non-taxonomic relationships identification, we propose a novel method named SGNRI (Semantic Graph Based Non-Taxonomic Relationships Identification) in which the structure information of semantic graph and context information of terms is combined to identify non-taxonomic relationships efficiently. For relation label identification, dependency syntactic information combined with classic statistical information is used to find appropriate verbs that act as relation labels. Because we believe the verbs to act as labels should have semantic relationships with the two concepts instead of arbitrary verbs or verb phrases just located between two concepts.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2881422, IEEE Access

The rest of the paper is organized as follows. Section 2 reviews related work in the area of big data in smart city and non-taxonomic relation extraction. Section 3 introduces the workflow of our method and its implementation. Then, section 4 evaluates the performance of our method, which includes the performance of non-taxonomic relation identification and labeling. Finally, concluding remarks and future work are given in section 5.

To summarize, we make the following contributions:

(1) SGNRI is proposed to identify non-taxonomic relationships. Structure information of semantic graph and context information of terms is combined under the framework of network representation to identify non-taxonomic relationships. For this task, the F1 value reaches 81.4%.

(2) Dependency syntactic information combined with statistical information is proposed to extract labels for non-taxonomic relationships. The total precision of the non-taxonomic relationship labels extraction is achieved to 73.4%, and 87.8% non-taxonomic relations can be provided with "good" labels.

(3) Provide a useful way for data centers, which store and process the received big data, to extract and represent knowledge.

II. RELATED WORK

A. RELATED WORK ON BIG DATA IN SMART CITY

Toward big data in green city, sensor-cloud is investigated in [1]. The aim of the article is acting as a guidance for relative research. Three types of sensor-cloud for green city are presented. The participatory sensing, agent, and social network are incorporated respectively for sensing big data, transmitting big data, and sharing big data.

Multi-Method Data Delivery (MMDD) scheme for sensor-cloud users is presented in [2]. Depend on the analysis of the potential applications and recent work about sensor-cloud, two issues are summed up to be solved. One issue is about repeated data transmissions, the other issue is about simultaneous data transmissions. Both of the two issues can increase the requirement with respect to the energy and resources as well as the bandwidth of sensor-cloud. Four kinds of delivery strategies are incorporated in MMDD to solve these two issues. The evaluation results show that MMDD could achieve lower delivery cost or less delivery time for sensor-cloud users.

Trust-based communication [4] is widely used in various systems [44, 45]. Three types of trust-based communication mechanisms for sensor-cloud are proposed in [3] to explore the performance of Industrial Internet of Things (IIOT). Open research issues are presented at the end of the article, which regarding trust-based communication for Mobile Sensor-Cloud, Underwater Sensor-Cloud, Green Sensor-Cloud, and Social Sensor-Cloud.

Trust-assisted sensor cloud (TASC) is one kind of sensor-cloud for smart city. Secure multimedia big data application in TASC is investigated in [5]. Two types of TASC, TASC-S and TASC-M, are proposed to address the critical issues that affect the success of secure multimedia big data in TASC. Where the throughput of TASC-S and TASC-M can both be generally higher than that of sensor-cloud without trust assistance, and can trend with tuned trust value threshold and fluctuate with the same trust value thresholds respectively.

Five Sensor-Cloud Pricing Models (SCPMs) are introduced in [6]. The characteristics of these models are discussed and exhibited in the article. The case studies regarding the application of SCPMs are presented together with the review of user behavior study.

B. RELATED WORK ON NON-TAXONOMIC RELATION EXTRACTION

Extracting the non-taxonomic relations is considered one of the most difficult tasks and is often neglected. Although there have been related research works on extracting the non-taxonomic relations. These methods may have limited practical applications or neglect syntactic and semantic information. This may cause concept pairs of extraction to be irrelevant, as well as the verbs extracted for concept pairs may not be appropriate. There are two types of non-taxonomic relation extraction: the supervised non-taxonomic relation extraction and the unsupervised non-taxonomic relation extraction. Our research work is focusing on the unsupervised non-taxonomic relation extraction.

Wong et al. [7] proposed a method to extract non-taxonomic relations from unstructured text. Correlated concept pairs are allowed to be located in different sentences. An association rule mining algorithm is used to identify potential concept pairs. Non-taxonomic relations are distinguished with taxonomic relations based on existing domain ontology. Relation labels are extracted by a pattern-based linguistic approach.

Serra et al. [8] presented a semi-automatic technique for the extraction of non-taxonomic relationships from an English corpus. Five NLP techniques are used for corpus annotation. Three extraction rules are used for the extraction of candidate relationships. Two statistical methods are used for relation refinement and label identification. With the use of extraction rules, the system receives better results compared to other techniques. However, no syntactic or semantic information is used in their solution.

Sanchez and Moreno [9] present an approach for learning and labeling non-taxonomic relations automatically. The Web is used as a data source for the purpose of constructing the domain ontology from scratch. First, a set of verbs that have high domain relevance is extracted. A statistical function is used to measure the degree of relationship between the domain and the verb. Then, the verbs combined with the domain key words are constructed to be the patterns used for non-taxonomic relation extraction, and the verbs are used as the labels of the relations. Because all the ontology concepts are identified from the Web automatically, it is difficult to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2881422, IEEE Access

compare their results with other techniques. However, they propose the method to evaluate the extraction results against WordNet.

Villaverde et al. [10] proposed a technique for discovering and labeling non-taxonomic relationships with a corpus of domain texts and a list of candidate concepts as input. Pairs of concepts combined with verbs between them are considered to be candidate non-taxonomic relation patterns. To ensure the precision of the system, the two concepts appear in the same sentence and are separated by no more than N terms, which are extracted with the verb between them. Association rules are used to suggest the existence of a relationship between a pair of concepts. Suitable labels are recommended and ordered by confidence. Weichselbraun et al. [11] presented a method for refining relation labels for non-taxonomic relations. A centroid function based on vector space is designed to train the model for relation label extraction. Structured information is used to remove invalid relation labels and improve the system performance.

Ferreira et al. [12] presented a method to extract non-taxonomic relations from a Brazilian Portuguese corpus. However, no new ideas are proposed to solve this task.

There are also many state-of-the-art ontology learning tools: Text-to-Onto [13], Text2Onto [14], ASIUM [16], Hasti [17], OntoLearn [22], RelExt [23], etc. Use of these tools can automatically construct an ontology, including the extraction of non-taxonomic relations. However, not all of these tools address the task of labeling the non-taxonomic relations.

There are many open information extraction systems in the open field relationship extraction in the English field, such as TextRunner [24], WOE [25], ReVerb [26] and OLLIE [27]. However, due to the differences in Chinese and English grammar and the high complexity of Chinese grammar, these systems may not be suitable for the Chinese domain. Most existing Multiple semantic information systems focus on English, and little research has been reported on Chinese. In addition, existing ORE (Open Relation Extraction) techniques are mainly concerned with the extraction of text relations, without trying to give semantic analysis, which is the advantage of traditional IE.

Tseng et al. [28] presents the Chinese Open Relation Extraction (CORE) system that is able to extract entity-relation triples from Chinese free texts based on a series of NLP techniques, i.e., word segmentation, POS tagging, syntactic parsing, and extraction rules. Qiu et al. [29] presents a syntax-based Chinese ORE system, ZORE, for extracting relations and semantic patterns from Chinese text. ZORE identifies relation candidates from automatically parsed dependency trees. A novel double propagation algorithm is used to extracts relations with their semantic patterns.

Knowledge representation learning aims to represent the relationship between existing knowledge, as well as conduct relational reasoning and so on. Network representation learning research aims to explore the ability to better study and analyze the connections between nodes in a complex information network. The task of identification of non-taxonomic relationships is to discover pairs of related concepts. Thus, network representation learning is suitable for finding associated concept pairs.

In recent years, there have been a large number of NE models proposed to learn efficient vertex embeddings, LINE [30] optimizes the joint and conditional probabilities of edges in large-scale networks to learn vertex representations. Node2vec [31] modifies the random walk strategy in DeepWalk into biased random walks to explore the network structure more efficiently. [32] introduce groupenhanced network embedding (GENE) to integrate existing group information in NE. [33] regard text content as a special kind of vertices, and propose context-enhanced network embedding (CENE) through leveraging both structural and textural information to learn network embeddings.

III. IDENTIFICATION OF NON-TAXONOMIC RELATIONSHIPS

In this section, a novel framework is proposed for non-taxonomic relation extraction from domain big data. To focus on the non-taxonomic relation extraction, we take a domain-specific text corpus and a set of domain concepts as input. We propose a novel method SGNRI based on semantic graph to identify non-taxonomic relationships.

With the widespread use of information technologies, information networks have increasingly become popular to capture complex relationships across various disciplines [37]. Analyzing information networks plays an important step to obtain knowledge from big data. However, network analytic tasks are computationally expensive, especially for large-scale networks with millions of vertices. Recently, network representation learning (NRL) has been proposed to learn latent and low-dimensional representations of network vertices. This makes network analytic tasks more easily and efficiently.

Depend on the idea of NRL, (1) we first construct the semantic graph (relation network) for all the concepts, where each node corresponds to one domain concept. (2) Then Context-Aware Network Embedding (CANE) [18] is used to fully utilize the context information of nodes (concepts) to learn context-relevant representations for each node (concept). This method makes representation of each node (concept) contains more semantic information, and helps to discover non-taxonomic relations in a deeper level. The similarities of concept pairs are calculated based on node embeddings, which help to identify the non-taxonomic relationships.

A. SEMANTIC GRAPH CONSTRUCTION

Two algorithms are used to calculate the similarities of concept pairs, Latent Dirichlet Allocation (LDA) [15] and Word2Vec [19-21]. Then the semantic graph is constructed based on the LDA-similarity and Word2Vec-similarity, respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2881422, IEEE Access

LDA-similarity Based Semantic Graph Construction. According to [15], the joint probability distribution of two words can be calculated based on Latent Dirichlet Allocation (LDA) theory. In this paper, value of joint probability is act as LDA-similarity for each concept pair.

The LDA algorithm assumes that each document is a mixture of a small number of latent topics and that each word creates a contribution to one topic. v_i represents a word in document d_m, z represents one topic, and K represents the number of topics. The LDA model considers three parameters: α, η and K. To obtain a word, the model chooses the topic distribution θ_m for document d_m through $P(\theta \mid \alpha)$ and chooses topic k through $P(z \mid \theta_m)$ and $\beta_k \sim Dirichlet(\eta)$. The distribution of each word given a topic z is $P(u_m \mid z, \beta_z)$. The LDA algorithm provides two output matrices $\Theta: P(u = v_i \mid z = k, \beta_k)$ and $\Phi: P(z = k \mid \theta_m)$, which represent the probabilities between topic-document and word-topic, respectively.

The joint probability between two words u_m and y_m in document d_m can be obtained as:

$$P(u_m, y_m) = \sum_{k=1}^{K} P(u_m, y_m \mid z = k, \beta_k) P(z = k \mid \theta_m)$$

For the convenience of calculation, an approximation for the above formula can be written as:

$$P(u_m, y_m) \approx \sum_{k=1}^{K} P(u_m \mid z = k, \beta_k) P(y_m \mid z = k, \beta_k) P(z = k \mid \theta_m)$$

Finally, the probability distribution of two words u and y in the entire corpus can be obtained as:

$$P(u, y) \approx \sum_{m=1}^{M} P(u_m, y_v) \delta_m$$

where δ_m is the prior probability for each document.

Domain-specific texts and domain-relative terms are collected as the input, we use the joint probability of two terms to construct the graph structure. Where edges are constructed between two terms only if the value of the similarity is higher than given threshold. Different numbers of topic are given from 10 to 100 (step is 5). Under each topic, threshold *l_t* is set from 0.01 to 0.09 to get corresponding semantic graph.

Word2vec-similarity Based Semantic Graph Construction. Word2vec is a tool based on deep learning and released by Google in 2013. This tool adopts two main model architectures, continuous bag-of-words (CBOW) model and continuous skip-gram model. They are used to learn the vector representations of words. The CBOW architecture predicts the current word based on the context, and the skip-gram predicts surrounding words given the current word. The Word2Vec model uses a word's context information to convert a word into a low-dimensional vector. The more similar words are in the space vector, the closer they are, because of its excellent ability to represent semantic relationships.

In this paper, we use domain relative corpus to train a Word2Vec model and then compute the similarity between terms to construct the semantic graph structure. After the similarity between terms computed, different threshold w_t are set from 0.01 to 0.59 to generate semantic structure graph.

B. SIMILARITY CALCULATION AND RELATIONSHIP IDENTIFICATION

CANE is used to learn node representation here, so knowledge base is required to provide context information for nodes (concepts). Two different methods are proposed to build the term's context information. One is the domain relative texts corpus itself, which is called Context Information 1. The sentences containing the terms are extracted out as context information of terms. The other is the encyclopedia source, which is called Context Information 2. The definitions of terms in encyclopedia source are collected as context information of terms. Context information of terms provides more background knowledge which could improve the performance and bring better node representations.

In this paper, we use domain relative corpus to train a Word2Vec model and then compute the similarity between terms to construct the semantic graph structure. After the similarity between terms computed, different threshold w_t are set from 0.01 to 0.59 to generate semantic structure graph.

The semantic graph that constructed in the previous subsection together with context information is served as input of CANE. Then context-aware embeddings for each term can be obtained. Representation of each term contains more semantic information which helps to discovered non-taxonomic relations in a deeper level.

In the vector space learned by CANE, cosine similarity are used to calculate the similarity between two terms, $Sim(A_{i,k}, A_{j,l})$ represents the similarity between terms, where $A_{i,k}$ and $A_{j,l}$ represent the vector representation of the term, respectively. $A_i = \{A_{i,k} \mid i = 0, \cdots, n; k = 0, \cdots, p\}$, $A_j = \{A_{j,l} \mid j = 1, \cdots, n; l = 0, \cdots, q\}$, where n represents the number of concepts, p and q represent the number of terms that concepts A_i and A_j contains.

$$\mathrm{Sim}(\mathrm{A_{i,k}}, \mathrm{A_{j,l}}) = \tfrac{1}{2} + \tfrac{1}{2} \times \frac{\sum_{k=0,l=0}^{p,q} (\mathrm{A_{i,k}} \times \mathrm{A_{j,l}})}{\sqrt{\sum_{k=0}^{p} (\mathrm{A_{i,k}})^2} \times \sqrt{\sum_{l=0}^{q} (\mathrm{A_{j,l}})^2}} (\mathrm{i} \neq \mathrm{j})$$

To increase the matching rates, we use the synonym set instead of each term. Each term is a word related to the target domain. Among those terms, synonym sets are used instead of terms. Lexical database and external semantic repository are used to find synonym sets. The terms that contain the same semantic sense are merged into a set, and each set is called a concept. Therefore, the concept is represented by a collection of terms.

$$\mathrm{Concept} = \{A_i, A_j \mid \mathrm{i} = 0, \cdots, \mathrm{n}; , \mathrm{j} = 1, \cdots, \mathrm{n}\}$$

Where n represents the number of concepts.

As the term is a collection of concepts, $Sim(A_i, A_j)$ represents the similarity between concepts.

$$Sim(A_i, A_j) = \sum Sim(A_{i,k}, A_{j,l})$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2881422, IEEE Access

Each combination of two concepts is constructed into a pair. Values of cosine similarities are computed as selection index. The concept pairs are added in the collection $R\text{elation}_{\text{identification}}$ if the values of the Cosine similarities are higher than the given threshold. $R\text{elation}_{\text{identification}}$ represents a collection of non-taxonomic relationship identifications. The range of threshold is from 0.001 to 0.999.

$$R\text{elation}_{identification} = \{(A_i, A_j) \mid Sim(A_i, A_j) > Threshold\}$$

IV. NON-TAXONOMIC RELATION LABELING

Lexical and syntactical information is combined with statistical information to label the non-taxonomic relationships. First, different verbs are extracted to label each relationship; second, an evaluation function is designed to rank the labels; finally, label sets are constructed, and several highest-confidence label sets are recommended to the user.

A. EXTRACTION OF VERBS

Verbs between two concepts are usually extracted to act as labels of the relationships. Some research works first identify the domain-relative verbs [9, 10] and then use these verbs combined with concepts to construct non-taxonomic patterns; finally, a data mining algorithm (such as the Association Rule Mining Algorithm) is used to refine the label of the relationship. Some research works first find concept pairs [7, 8], and then the frequency information is used to extract the verb labels. However, most works only focus on the statistical information to extract verb labels, and syntactic and semantic information is often neglected. It is crucial for the concept to extract suitable verbs. Appropriate verbs can well represent the non-taxonomic relationship between concept pairs.

In this paper, a dependency parser is used to parse the sentences and to find the dependency relationship between concept and verb. Dependency grammar can describe the relationship of two words directly. Each dependency relation has a relation type, which can be naturally mapped into a semantic expression.

For each concept pair, different types of verbs are defined as follows. Verbs are extracted from the corresponding sentences that contain both of the concepts of each concept pair and are constructed into different verb sets for each concept pair.

VerbsBetween (*VB*): All the verbs located between the two concepts of a concept pair.

OnlyVerbBetween (*OVB*): If the verb is the only verb between the two concepts.

CommonFaVerb (*CFV*): If the nearest common ancestor is a verb in the sub-dependency tree of the concept pair.

CommonFaVerbBetween (*CFVB*): If the nearest common ancestor is a verb and located between the two concepts in the sentence.

With observation of the Chinese online news corpus, we found that there are many long sentences containing several commas. Strings separated by commas are considered to be relatively independent semantic units of the sentence. Therefore, four other types of verbs are defined for pairs of concepts that are not separated by a comma. They are Nosep VerbsBetween (*NS_VB*), Nosep OnlyVerbBetween (*NS_OVB*), Nosep CommonFaVerb (*NS_CFV*), and Nosep CommonFaVerbBetween (*NS_CFVB*).

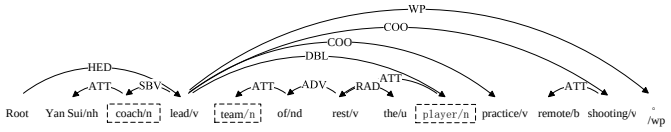


FIGURE 1: The dependency tree

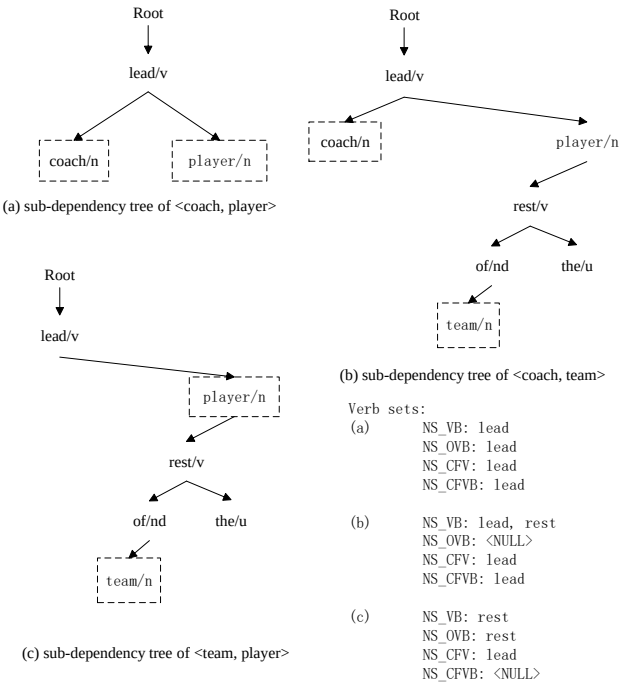


FIGURE 2: The sub-tree

Figure.1 is the dependency parse tree of an example sentence. The head of the sentence is “lead”. To show the dependency structure clearly, the Chinese sentence is translated into English word by word, without thinking about grammar. There are three concepts marked by a dotted box in the sentence: “coach”, “player”, and “team”.

Figure.2 shows the sub-dependency trees and corresponding verb sets for three concept pairs: <coach, player>, <coach, team>, and <team, player>. It is easy and intuitive to reach the conclusion that “lead” will have a higher probability of acting as the non-taxonomic relation label of the concept pair <coach, player> than the other two concept pairs.

B. SELECTION OF LABELS

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2881422, IEEE Access

Corresponding verb sets can be obtained according to the different verb definitions above. The next step is to identify the most appropriate verb that can act as the label of the non-taxonomic relationship. Actually, to improve the performance, we provide several labels for the user to choose from.

First, the noisy verbs that contain less semantic information need to be removed from the list, such as "is", "do" and "can". Second, a function is defined to calculate the confidence levels of verbs. Finally, the semantic relation distance between each pair of verbs is computed by HowNet, which is a lexical database and semantic repository for the Chinese language. The verbs that contain the same semantic sense are merged into a label set. The sum of all element scores in the set is computed to act as the final score of this label set, and the values are used to rank the list of them.

Let $P=\{P_1,P_2,...,P_n\}$ be the set of concept pairs, where P_i is a concept pair. $V=\{V_1,V_2,...,V_n\}$ is the set of verbs, where V_i is the verb set that contains all the extracted verbs for P_i. For each P_i, there are eight different verb type sets that can be extracted, and four are corresponding to the concept pairs that appear in the same sentence but may be separated by commas: VB, OVB, CFV, and $CFVB$; another four sets are corresponding to the non-separated concept pairs that appear in the same sentence and are not separated by commas: NS_VB, NS_OVB, NS_CFV, and NS_CFVB. Therefore, we define S_i as the collection of these eight verb type sets of concept pair P_i, and V_i as the union of these ten sets.

$$S_i=\{VB_i,OVB_i,CFV_i,CFVB_i,NB\_VB_i,$$
$$NB\_OVB_i,NB\_CFV_i,NB\_CFVB_i\}$$

$$V_i=\bigcup S_i=\{v_{i1},v_{i2},...,v_{im}\}$$

where v_{ij} is the jth verb in set. The confidence level score for each verb is computed in the following manner:

$$Score(v_{ij})=\sum_{SET\in S_i}F_{SET}(v_{ij})$$

where

$$F_{SET}(v_{ij})=\begin{cases}\dfrac{feq(v_{ij})}{\max_{feq}(SET)}, & \text{if } v_{ij}\in SET\\ 0, & \text{if } v_{ij}\notin SET\end{cases}$$

In this function, $feq(v_{ij})$ is the frequency of v_{ij}, and $\max_{feq}(SET)$ is the max verb frequency in set SET.

After calculating each verb's confidence level score (verb score for short), verbs that contain the same semantic sense are collected into a label set. Therefore, verb set V_i can be expressed as a set of label set L_{ik}.

$$V_i=\{v_{i1},v_{i2},...,v_{im}\}=\{L_{i1},L_{i2},...,L_{ir}\}$$

Label sets of each concept pair P_i are ranked according to L_Score, and the top 5 label sets will be returned to users. The verb that has the highest verb score in the label set is used as the tag of this label set.

V. EXPERIMENTS AND RESULTS

A. EXPERIMENTAL SETUP

The following experiments have been performed on a set of Chinese news texts, which is relative with the domain of football competition.

We collected 2600 documents regarding the competition news of the China Football Association Super League (CSL). 498 terms are used to construct semantic graph, where 24 domain concepts with 74 non-taxonomic relationships are labeled manually for test.

The Language Technology Platform (LTP), which is developed by HIT-SCIR [34], is used as the dependency parser. HITSCIR Tongyici Cilin (Extended) [34] and Hownet [35, 36] are used together to find synonym sets of concepts and label sets that contain the same semantic sense.

For non-taxonomic relation identification, values of precision, recall, and F1 are used to measure the performances of the system with different thresholds. For non-taxonomic relation labeling, the top 5 label sets are provided as candidate relation labels for each concept pair. Then, the performances are evaluated by domain experts.

B. NON-TAXONOMIC RELATION IDENTIFICATION RESULTS

Baselines

We employ the following three methods as baselines:

Apriori algorithm. Apriori algorithm is a frequent itemsets algorithm for mining association rules. Its core idea is to mine frequent itemset. Frequent itemset-2 is extracted out as the non-taxonomic relation set. Experiment results are shown in Table I. Best extraction results are obtained when support and confidence values are 0.00032 and 0.01respectively. 16 concept pairs can be extracted out with only 6 pairs are correct, which corresponds to the best F1 is 0.133.

TABLE I
RESULTS BASE ON APRIORI ALGORITHM

(support, confidence)	#Identified	#Correct	P	R	F1
(0.0004,0.01)	3	3	1.000	0.041	0.078
(0.00038,0.01)	4	3	0.750	0.041	0.077
(0.00038,0.5)	4	3	0.750	0.041	0.077
(0.00037,0.01)	6	4	0.670	0.054	0.100
(0.00035,0.01)	9	4	0.440	0.054	0.096
(0.00032,0.01)	**16**	**6**	**0.375**	**0.081**	**0.133**
(0.000304,0.01)	19	4	0.210	0.054	0.086

Word2Vec based method. Use the distances between word representations to identify non-taxonomic relations. The dimension is set to 400. Non-Taxonomic relations are identified when the distances between the two concepts are higher than the given identification threshold. Figure. 3 shows

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2881422, IEEE Access

the experimental results of this method with different threshold values. Best F1 value is 0.59 when threshold is set to be 0.058.

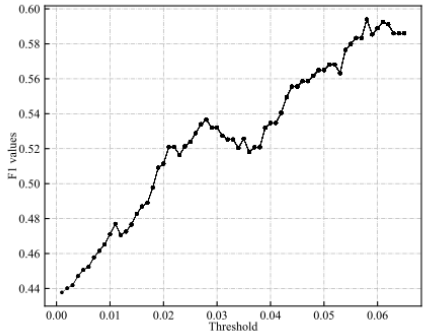


FIGURE 3: Results based on Word2Vec

LDA based method. Use joint probabilities of two concepts to identify non-taxonomic relations. Where joint probabilities can be calculated based on LDA topic-document matrix and word-topic matrix. Different topics are given from 10 to 100, step is 5. Different experimental results are calculated according to different topic numbers and identification threshold values, which are shown in Figure. 4. Better F1 values are achieved when threshold values are around 0.02 with arbitrary topic numbers. Best F1 value reaches to 0.739 when topic number is 30 and threshold value is 0.022.

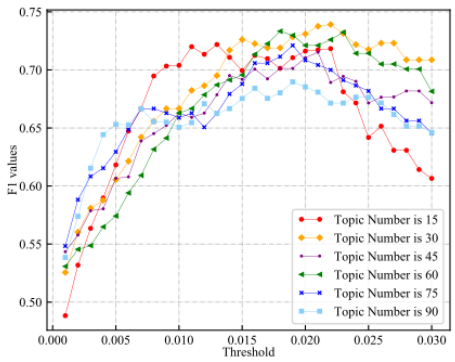


FIGURE 4: Results based on LDA

SGNRI Results and Analysis

Figure.5 – Figure.10 and Table II show the SGNRI evaluation on same dataset. Two types of context information (as mentioned in previous section) are used respectively to compare the experimental results. From these Figures and the table, we have the following observations:

(1) Figure.5 and Figure.6 show the performances of SGNRI[1](LDA) and SGNRI[2](LDA). SGNRI[1](LDA) means the semantic graph is constructed based on LDA-similarity, and Context Information 1 is used as the context information for network representation learning. SGNRI[2](LDA) means the semantic graph is constructed based on LDA-similarity, and Context Information 2 is used as the context information. Threshold for semantic graph construction l_t is set as 0.01, parameters in CANE model is $\alpha = 1.0$, $\beta = 0.2$, $\gamma = 0.2$.

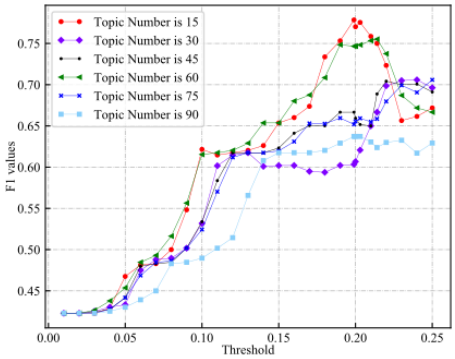


FIGURE 5: Performances of SGNRI[1](LDA)

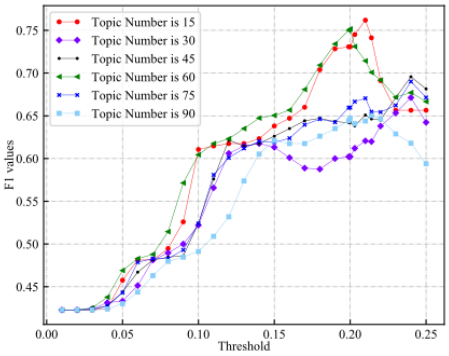


FIGURE 6: Performances of SGNRI[2](LDA)

Best F1 values are captured when topic number is 15 for both of the SGNRI1(LDA) and SGNRI2(LDA). Highest value of SGNRI1(LDA) is 0.079 when the identification threshold value is 0.199. Highest value of SGNRI2(LDA) is 0.0762 when the identification threshold value is 0.21.

The best performances of these two models are compared in Figure. 7 which shows Context Information 1 is more helpful for relation identification.

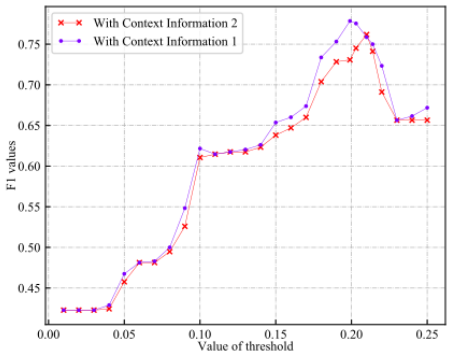


FIGURE 7: Performances of SGNRI[1](LDA)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2881422, IEEE Access

(2) Figure.8 and Figure.9 show the performances of SGNRI[1](Word2Vec) and SGNRI[2](LDA) with different semantic graph construction threshold values. SGNRI[1](Word2Vec) and SGNRI[2](LDA) means the semantic graph is constructed based on Word2Vec-similarity with Context Information 1 and Context Information 2 as context information respectively. Threshold for semantic graph construction w_t is set as 0.15, parameters in CANE model is $\alpha = 1.0$, $\beta = 0.2$, $\gamma = 0.2$.

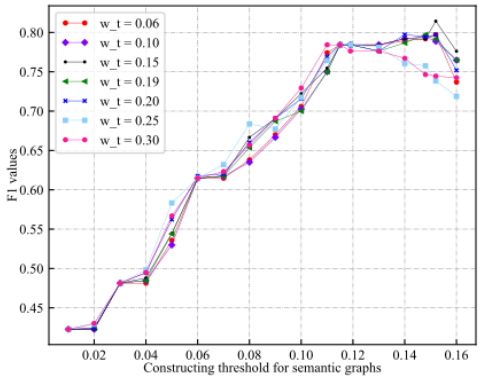


FIGURE 8: Performances of SGNRI[1](Word2Vec)

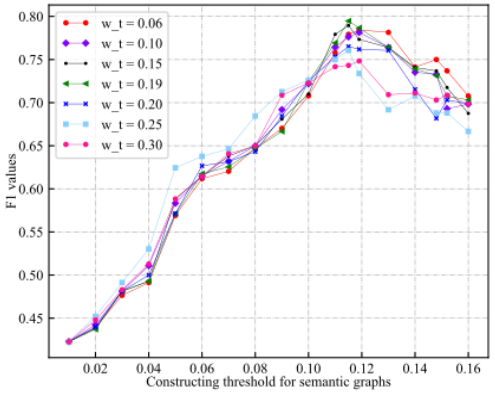


FIGURE 9: Performances of SGNRI[2](Word2Vec)

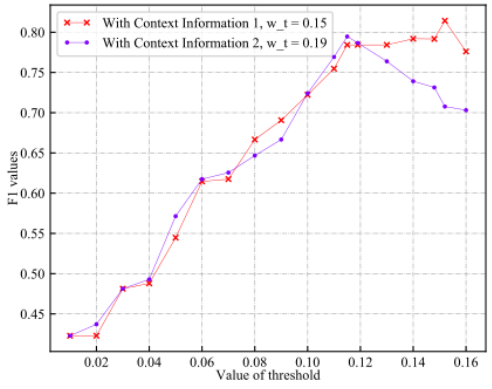


FIGURE 10: Best performances of SGNRI[1](Word2Vec) and SGNRI[2](Word2Vec)

Highest $F1$ value of SGNRI[1](Word2Vec) is 0.814 when the identification threshold value is 0.152. Highest $F1$ value of SGNRI[2](Word2Vec) is 0.795 when the identification threshold value is 0.115.

The best performances of these two models are compared in Figure. 10 which also shows Context Information 1 is more helpful for relation identification.

(3) As we can see from Table II, among the models that are compared, the association rules achieved the lowest $F1$ values. The performances of Word2Vec based model is lower than LDA based model. SGNRI model achieves better performance in any cases, which illustrates network structure can provide more useful information and capture complex relationships.

TABLE II
RESULTS OF DIFFERENT METHODS FOR IDENTIFY NON-TAXONOMIC RELATIONS

Method	P	R	F1
Apriori-based model	37.5%	8.1%	13.3%
Word2vec-based model	53.9%	66.2%	59.4%
LDA-based model	79.7%	68.9%	73.9%
SGNRI[1](LDA)	76.7%	75.7%	**76.2%**
SGNRI[2](LDA)	77.3%	78.4%	**77.9%**
SGNRI[1](Word2Vec)	77.9%	81.1%	**79.5%**
SGNRI[2](Word2Vec)	86.4%	77.0%	**81.4%**

Word2Vec based model not achieve good performance maybe because Word2Vec is good at finding semantically similar word pairs but not the relationships, since two words that have relationship between them not always have the similar semantics. Instead, the similarity information provided by LDA is more helpful for relation identification.

However, the performances of SGNRI(Word2Vec), which use Word2Vec-based similarity to construct semantic graph for network representation-based relation identification, is higher than SGNRI(LDA) model. This illustrates when constructing information network based on word semantic similarities, it is helpful to find indirect semantic relations due to the transitivity of the network structure.

Context information 1 plays better than Context information 2, which shows definitions of terms is not good at finding non-taxonomic relations. To a large extent due to the definitions of terms contains fewer common words or phases between each other, however, Context information 1 can avoid this problem naturally since it's all come from domain texts which contains more common key words and context.

C. NON-TAXONOMIC RELATION LABELING RESULTS

For each non-taxonomic relation, the system provides the top 5 label sets as candidate relation labels. Therefore, a total of 370 (74*5) label sets is found and presented to the user. A domain expert was asked to rate the candidate label sets as "good" or "bad".

Verbs are extracted from the corresponding sentences to act as labels of non-taxonomic relations. Two methods are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2881422, IEEE Access

used to extract verbs. In the first method, sentences are considered to be analysis units. In the other method, sentences are separated into smaller semantic units by commas.

TABLE III

EXTRACTION RESULTS FOR THE CONCEPT PAIR <COACH, TEAM>)

Concept Pair	Label Set	Common Sense	Score	Label Rating
"Jiaolian, Qiudui" <Coach, Team>	"danren" ("served as") "danren" ("served as") "jianren" ("concurrently served as") 兼 ("concurrently served as")	"danren" ("served as")	8.07 5.20 2.34 0.53	Good
	"zhijiao" ("coaching") "zhijiao" ("coaching") "canjiao"("participate") "canyu" ("participate")	"congshi" ("engaged in")	3.73 2.33 1.02 0.38	Good
	"jiashang"("add") "jiashang" ("add")	"zengjia" ("increase")	3.35 3.55	Bad
	"renming" ("appointment") "renming" ("appointment")	"shjzhishi" ("let it become")	2.34 2.34	Bad
	"dai"("lead") "dai" ("lead") "lingdao" ("lead") "dailing" ("lead")	"yindao" ("guide")	2.25 1.76 0.37 0.12	Good

TABLE IV

EXTRACTION RESULTS FOR THE CONCEPT PAIR <TEAM, PLAYER>

Concept Pair	Label Set	Common Sense	Score	Label Rating
"Qiudui, Qiuyuan" <Team, Player>	"chengwei" ("become") "chengwei" ("become")	"chengwei" ("become")	3.569 3.569	Bad
	"qianding" ("signed") "qianding" ("signed") "qianyue" ("signed") "qian" ("signed") "qianding" ("signed")	"yueidng" ("signed")	2.99 1.21 0.91 0.53 0.34	Good
	"yinjin" ("introduce") "yinjin" ("introduce")	"tichu" ("propose")	2.68 2.68	Good
	"tixing"("remind") "tixing" ("remind") "zhonggao" ("advise")	"quanshuo" ("persuade")	2.65 1.36 1.29	Bad
	"baokuo" ("contain") "baokuo" ("contain")	"baokuo" ("contain")	2.26 2.26	Good

TABLE V

EXTRACTION RESULTS FOR THE CONCEPT PAIR <REFEREE, GAME>)

Concept Pair	Label Set	Common Sense	Score	Label Rating
"Caipan, Bisai" <Referee, Game>	"jieshu" ("end") "jiechu" ("end") "zhongzhi" ("terminate") "zhongzhi" ("terminate")	"tingzuo" ("stop")	6.24 3.90 1.66 0.68	Good
	"chui" ("blow the whistle") "chui" ("blow the whistle")	"chui" ("blow the whistle")	5.18 5.18	Good
	"jinxing" ("carry on") "jinxing" ("carry on") "zhifa" ("perform") "zuowei" ("act as")	"shishi" ("conduct")	4.48 2.24 1.64 0.60	Bad
	"chuixiang" ("blow") "chuixiang"("blow")	"chuixiang" ("blow")	2.7 2.7	Good
	"zhu" ("call the shots") "zhu" ("call the shots")	"jueding" ("decide")	2.06 2.06	Bad

When using the first method, for the 370 total label sets, there are 272 "good" and 98 "bad" with a precision of 73.4%. Tables III-V show a portion of the extracted Chinese non-taxonomic concept pairs and the top 5 label sets found for each concept pair, where Pinyin is used to represent Chinese words. We can observe that, for each concept pair, there is usually more than one "good" label set. For the 74 total non-taxonomic relations, 65 non-taxonomic relations can be provided with "good" labels (87.8%).

When using the second method, the precision is 71.1%. This may be because of the complexity of Chinese grammar. When just considering the situation that the concept pairs are non-separated by a comma, many high-frequency domain-relative verbs will be ignored.

Table III shows the extraction results of the concept <Coach, Team>. The highest confidence label set contains three verbs, and we use "served as" as the tag of this label set because it receives the highest verb score of 5.20 among the three verbs. The second label set is tagged as "coaching", which is more suitable to be the label of this non-taxonomic relationship as the formal expression.

Table IV shows the extraction results of the concept <Team, Player>. The second label set is constructed by four verbs, and their verb scores of these four are not very high. However, when we use Hownet to collected them into a set, it receives a high label score.

Table V shows the extraction results of the concept <Referee, Game>. The second and fourth label sets have the same meaning, and there is another extracted verb "blow the whistle" that has the same semantic meaning as them. However, they do not have same sense records in the Hownet, so they are not collected into a label set and instead need to be ranked independently. As a result, these three verbs cannot obtain the best position in the list, and the verb "blow the whistle" is not even ranked in the top 5.

VI. CONCLUSIONS

Targeted to extract more useful information and knowledge from domain big data in smart city, this article has investigated non-taxonomic relation extraction, which is the basis of knowledge representation. The recent work about big data in smart city and relation extraction is reviewed first. Further, a new framework for extraction of non-taxonomic relations from domain big data in smart city is proposed. Structure information of semantic graph and context information of terms is combined to identify non-taxonomic relationships. Statistical information combined with dependency syntactic information is used to extract labels for non-taxonomic relations. Verbs that contain the same semantic senses are constructed as a label set to label the non-taxonomic relations.

Based on the analysis of the experimental results, following conclusions can be obtained:

1.Combining semantic graph as structure information with context information could help to identify more effective non-taxonomic relationships.

2.There are more informal expressions than formal expressions in web news texts. However, the formal expressions are more suitable to act as labels of the non-taxonomic relations. Thus, how to extract good formal verb expressions is a task in the next step.

3.Verbs that contain the same semantic sense and are collected into a label set can help to obtain good extraction performance. However, only basing the task on a dictionary, such as HowNet, is not sufficient to obtain good results. Thus, how to cluster verbs with the same meaning is another task to be addressed in the future.

4. With the explosive growth of data from a variety of sources, multimedia big data is utilized to describe the huge amounts of multimedia data produced by different devices in smart city. Multi-elements and multi-information should be considered to improve the system performance in the future.

ACKNOWLEDGMENT

The authors express their sincere appreciation to the editors and the anonymous reviewers for their helpful comments.

REFERENCES

1. C. Zhu, H. Zhou, Victor C. M. Leung, K. Wang, Y. Zhang, and Laurence T. Yang. Toward Big Data in Green City. IEEE Communications Magazine, 55 (11), 2017, pp. 14-18.
2. C. Zhu , Joel J. P. C. Rodrigues , Victor C. M. Leung, L. Shu , and Laurence T. Yang. Trust-Based Communication for the Industrial Internet of Things. IEEE Communications Magazine, 56 (2), 2018, pp. 16-22.
3. C. Zhu, L. Shu, Victor C. M. Leung, S Guo, Y. Zhang, and Laurence T. Yang. Secure Multimedia Big Data in Trust-Assisted Sensor-Cloud for Smart City. IEEE Communications Magazine, 55(12), 2017, pp. 24-30.
4. H. Yu, Z. Shen, C. Miao, C. Leung, D. Niyato. A Survey of Trust and Reputation Management Systems in Wireless Communications. Proc. IEEE, vol. 98, no. 10, Oct. 2010, pp. 1755–72.
5. C. Zhu , Victor C. M. Leung , K Wang , Laurence T. Yang , Y Zhang. Multi-Method Data Delivery for Green Sensor-Cloud. IEEE Communications Magazine, 55 (5), 2017, pp. 176-182.
6. C. Zhu, X. Li, Victor C. M. Leung, Laurence T. Yang, C. H. Ngai, L. Shu. Towards Pricing for Sensor-Cloud. IEEE Transactions on Cloud Computing, 2017, DOI: 10.1109/TCC.2017.2649525.
7. M. K. Wong, S. S. R. Abidi, I. D. Jonsen. A multi-phase correlation search framework for mining non-taxonomic relations from unstructured text. Jounal of Knowledge and Information System, 38(3), 2014, pp. 641–667.

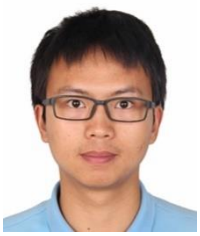
8. I. Serra, R. Girardi, P. Novais. PARNT: A Statistic based Approach to Extract Non-Taxonomic Relationships of Ontologies from Text. In Proceedings of 10th International Conference on Information Technology-New Generations, 2013, pp. 561-566.
9. D. Sanchez and A. Moreno. Learning non-taxonomic relationships from web documents for domain ontology construction. Jounal of Data & Knowledge Engineering, 64(3), 2008, pp. 600–623.
10. J. Villaverde, A. Persson, D. Godoy, A. Amandi. Supporting the discovery and labeling of non-taxonomic relationships in ontology learning. Jounal of Expert Systems with Applications, 36(7), 2009, pp. 102880–10294.
11. A. Weichselbraun, G. Wohlgenannt, and A. Scharl. Refining non-taxonomic relation labels with external structured data to support ontology learning. Jounal of Data & Knowledge Engineering, 69(8), 2010, pp. 763–778.
12. V. H. Ferreira, L. Lopes, R. Vieira, and M. J. Finatto. Automatic Extraction of Domain Specific Non-Taxonomic Relations from Portuguese Corpora. In Proceedings of 12th IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology, 2013, pp. 135-138.
13. A. Maedche, S. Staab. The text-to-onto ontology learning environment. In Proceedings of Software Demonstration at the 8th International Conference on Conceptual Structures, 2000.
14. P. Cimiano, J. Volker. Text2Onto: A framework for ontology learning and data-driven change discovery. In Proceedings of 10th International Conference on Applications of Natural Language to Information Systems, 2005, pp. 227-238.
15. D. M. Blei, A. Y. Ng, M. I. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 2003, pp. 933–1022.
16. C. Nedellec. Corpus-based learning of semantic relations by the ILP system, Asium. Learning language in logic, 2000, pp. 259–278.
17. M. Shamsfard, A. A. Barforoush. Learning ontologies from natural language texts. International Journal of Human-Computer Studies, 60(1), 2004, pp. 17–63.
18. C. Tu, H. Liu, Z. Liu, M. Sun. CANE: Context-Aware Network Embedding for Relation Modeling. The 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), 2017, pp. 1722-1731.
19. T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient estimation of word representations in vector space. In Proceedings of workshop at ICLR, 2013.
20. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean. Distributed representations of words and phrases and their compositionality, In Proceedings of neural information processing systems, 2013, pp. 3111-3119.
21. T. Mikolov, W. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies. Association for Computational Linguistics, 2013, pp. 746-751.
22. P. Velardi, R. Navigli, A. Cucchiarelli, F. NERI. Evaluation of OntoLearn, a methodology for automatic learning of domain ontologies. Ontology learning from text: methods, applications and evaluation, 123(92), 2003, pp. 92–106.
23. A. Schutz, P. Buitelaar. RelExt: A tool for relation extraction from text in ontology extension. In Proceedings of 4th International Semantic Web Conference, 2005, pp. 593-606.
24. M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, Oren Etzioni. Open information extraction from the web. Proceedings of IJCAI'07, 2007, pp. 2670-2676.
25. F. Wu and D. S. Weld. Open information extraction using Wikipedia. Proceedings of ACL'10, 2010, pp. 118-127.
26. A. Fader, S. Soderland, O. Etzioni. Identifying relations for open information extraction. Proceedings of EMNLP'11, 2011, pp. 1535-1545.
27. Mausam, M. Schmitz, R. Bart, S. Soderland, O. Etzioni. Open language learning for information extraction. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012, pp. 523–534.
28. Y. H. Tseng, L. H. Lee, S. Y. Lin, B. S. Liao, M. J. Liu, H. H. Chen, O. Etzioni, A. Fader. Chinese open relation extraction for knowledge acquisition. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers , 2014, pp. 12-16.
29. L. Qiu, Y. Zhang. ZORE: A syntax-based system for Chinese open relation extraction. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1870-1880.
30. J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, Q. Z. Mei. Line: Large-scale information network embedding. In Proceedings of WWW, 2015, pp. 1067–1077.
31. A. Grover and J. Leskovec. Node2vec: Scalable feature learning for networks. In Proceedings of KDD, 2016, pp. 855-864.
32. J. Chen, Q. Zhang, X. Huang. Incorporate group information to enhance network embedding. In Proceedings of CIKM, 2016, pp. 1901-1904.
33. X. Sun, J. Guo, X. Ding, T. Liu. A general framework for content-enhanced network representation learning. arXiv preprint arXiv:1610.02906 , 2016.

34. W. Che, Z. Li, T. Liu. LTP: A Chinese Language Technology Platform. Journal of Chinese Information Processing, 2(6), 2010, pp. 13–16.
35. Z. Dong, Q. Dong. Hownet Knowledge Database. Available from http://www.keenage.com/, 2007.
36. Q. Liu, S. Li. Word similarity computing based on How-net. Computational Linguistics and Chinese Language Processing, 7(2), 2002, pp. 59–76.
37. F. Colace, M. D. Santo, L. Greco, F. Amato, V. Moscato, A. Picariello. Terminological ontology learning and population using latent Dirichlet allocation. Journal of Visual Languages and Computing, 25(6), 2014, pp. 818–826.
38. D. Zhang, J. Yin, X. Zhu, C. Zhang. Network Representation Learning: A Survey. IEEE transactions on Big Data, 2017, 10.1109/TBDATA.2018.2850013.
39. Zhihong Tian, Feng Jiang, Xiang Yu, Yunsheng Fu, Xiang Cui, Lihua Yin. Insider Threat Detection Using Deep Learning and Dempster-Shafer Theory. Cluster Computing.
40. Xiang Yu, Zhihong Tian, Jing Qiu, Feng Jiang. A Data Leakage Prevention Method Based on the Reduction of Confidential and Context Terms for Smart Mobile Devices. Wireless Communications and Mobile Computing.
41. Yuhang Wang, Zhihong Tian, Hongli Zhang, Shen Su and Wei Shi. A Privacy Preserving Scheme for Nearest Neighbor Query. Journal of Sensors. 2018; 18(8):2440. https://doi.org/10.3390/s18082440.
42. Zhihong Tian, Yu Cui, Lun An, Shen Su, Xiaoxia Yin, Lihua Yin and Xiang Cui. A Real-Time Correlation of Host-Level Events in Cyber Range Service for Smart Campus. IEEE Access. vol. 6, pp. 35355-35364, 2018. DOI: 10.1109/ACCESS.2018.2846590.
43. Qingfeng Tan, Yue Gao, Jinqiao Shi, Xuebin Wang, Binxing Fang, and ZhiHong Tian. Towards a Comprehensive Insight into the Eclipse Attacks of Tor Hidden Services. IEEE Internet of Things Journal. 2018. DOI: 10.1109/JIOT.2018.2846624.
44. Juan Chen, Zhihong Tian, Xiang Cui, Lihua Yin, Xianzhi Wang. Trust Architecture and Reputation Evaluation for Internet of Things. Journal of Ambient Intelligence & Humanized Computing, 2018(2):1-9.
45. Feng Jiang, Yunsheng Fu, Brij B. Gupta, Fang Lou, Seungmin Rho, Fanzhi Meng, Zhihong Tian. Deep Learning based Multi-channel intelligent attack detection for Data Security. IEEE Transactions on Sustainable Computing.
46. Xiang Yu, Zhihong Tian, Jing Qiu, Feng Jiang. A Data Leakage Prevention Method Based on the Reduction of Confidential and Context Terms for Smart Mobile Devices. Wireless Communications and Mobile Computing, https://doi.org/10.1155/2018/5823439.

Jing Qiu, received the Ph.D. degree in computer applications technology from Beijing Institute of Technology. She was a Visiting Scholar with the University of Southern California, LA, USA, under the supervision of Professor Craig A. Knoblock. Her current research interest is Information Extraction, Network Representation, and Big Data Analysis. E-mail: qiujing@gzhu.edu.cn.

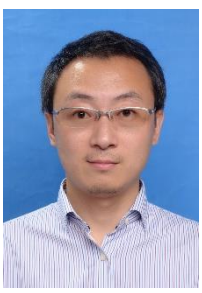
Yuhan Chai, born in 1993, Master Candidate, Hebei University of Science and Technology. Her current research interest is information extraction. E-mail: chyuhan_smile@163.com.

Yan Liu, born in 1992, Master Candidate, Hebei University of Science and Technology. Her current research interest is Deep Learning and Nerual Machine Translation. E-mail: apillowhill@gmail.com.



Zhaoquan Gu, received his bachelor degree in Computer Science from Tsinghua University (2011) and PhD degree in Computer Science from Tsinghua University (2015). He is currently a Professor in the Cyberspace Institute of Advanced Technology (CIAT), Guangzhou University, China. His research includes wireless networks, distributed computing, and big data analysis. E-mail: zqgu@gzhu.edu.cn.

Shudong Li, received M.S. degree in applied mathematics from Tongji University (China) in June 2005 and his Ph.D. degree in Information Security at Beijing University of Posts and Telecommunications (China) in July 2012. He is currently the postdoc of National University of Defense Technology, Changsha, China. His current research interest includes social network analysis, Big Data and its security, information security and cryptography, the robustness of complex networks. E-mail: lishudong@gzhu.edu.cn.



Zhihong Tian, Ph.D., professor, PHD supervisor, Dean of cyberspace institute of advanced technology, Guangzhou University. Standing director of CyberSecurity Association of China. Member of China Computer Federation. From 2003 to 2016, he worked at Harbin Institute of Technology. His current research interest is computer network and network security. E-mail: tianzhihong@gzhu.edu.cn.