Copyright © 2018 Tech Science Press CMC, vol.57, no.3, pp.603-619, 2018

Natural Language Semantic Construction Based on Cloud Database

Suzhen Wang[1], Lu Zhang[1], Yanpiao Zhang[1], Jieli Sun[1], Chaoyi Pang[2], Gang Tian[3] and Ning Cao[4, *]

Abstract: Natural language semantic construction improves natural language comprehension ability and analytical skills of the machine. It is the basis for realizing the information exchange in the intelligent cloud-computing environment. This paper proposes a natural language semantic construction method based on cloud database, mainly including two parts: natural language cloud database construction and natural language semantic construction. Natural Language cloud database is established on the CloudStack cloud-computing environment, which is composed by corpus, thesaurus, word vector library and ontology knowledge base. In this section, we concentrate on the pretreatment of corpus and the presentation of background knowledge ontology, and then put forward a TF-IDF and word vector distance based algorithm for duplicated webpages (TWDW). It raises the recognition efficiency of repeated web pages. The part of natural language semantic construction mainly introduces the dynamic process of semantic construction and proposes a mapping algorithm based on semantic similarity (MBSS), which is a bridge between Predicate-Argument (PA) structure and background knowledge ontology. Experiments show that compared with the relevant algorithms, the precision and recall of both algorithms we propose have been significantly improved. The work in this paper improves the understanding of natural language semantics, and provides effective data support for the natural language interaction function of the cloud service.

Keywords: Natural language, cloud database, semantic construction.

1 Introduction

The Internet of things and big data have promoted the rapid development of the artificial intelligence industry. Natural Language Processing is a major branch of artificial intelligence. It simulates and processes natural language through the research results of computer technology, such as analyzing, understanding or generating natural language, so that human and machine can realize natural communication. With the development of mobile terminals and natural language processing technology, numerous applications

[1] Hebei University of Economics and Business, Shijiazhuang, Hebei, 050061, China.

[2] The Australian e-Health Research Centre, ICT Centre, CSIRO, Australia.

[3] Shandong University of Science and Technology, Qingdao, Shandong, 266590, China.

[4] University College Dublin, Belfield, Dublin 4, Ireland.

[*] Corresponding Author: Ning Cao. Email: ning.cao2008@hotmail.com.

CMC. doi:10.32604/cmc.2018.03884 www.techscience.com/cmc

based on Natural Language processing technology have emerged today. American Nuance company launched a virtual voice assistant Nina. It gave users a very high degree of perception in the interactive experience [Masterson (2012)]. In 2009, Google launched the voice search service, which was the first combination of cloud technology and voice recognition technology in the industry [Breitbach (2010)]. In 2014, Microsoft Asia Internet Engineering Institute released Xiaoice, it introduced emotional computing framework, and realized a simple human-computer natural interaction [Shum, He and Li (2016)]. The development of natural language technology is inseparable from the support of big data. Nowadays, intelligent interactive products receive more and more attention. However, these products are faced with a same problem: their performance of speech recognition is correct, but the understanding of natural language is not accurate, so it cannot respond well to natural language.

In recent years, researches on natural language semantic construction have been increasing. For example, Archetypes [Beale (2002)] put forward a two-layer models semantic construction method, it constructs information semantic level and knowledge semantic layer, and creates the general semantic model and a special model. In literature [Misra, Sung, Lee et al. (2014)], a semantic analysis model is proposed and established, which defines the energy functions based on natural language instruction and coding independence hypothesis. The literature [Zhao, Li and Jiang (2016)] discloses a method for natural language understanding of intention, using intentions tagging, text vectoring and support vector machine to understand natural language instruction. He et al. [He, Deng, Gao et al. (2016); He, Deng, Wang et al. (2017)] proposed the improved model approach to grammatical evolution, which provides a new idea for semantic construction. At present, there are some problems in the research of semantic construction, such as poor portability and high requirement for data.

Understanding the semantics of natural language is the key step to achieve intelligent language interaction. Therefore, from the perspective of semantic analysis, this paper proposes a method of Chinese natural language semantic construction based on cloud database, and focuses on the corpus, ontology construction and semantic construction base on cloud database. The contributions of this paper include:

(1) This paper proposes a natural language semantic construction method based on cloud database;

(2) This paper proposes a TF-IDF and word vector distance based webpage de-duplication algorithm and a mapping algorithm based on semantic similarity;

(3) Based on the above algorithm, corpus and ontology knowledge base are stored in HBase, thesaurus and word vector library are stored in MySQL;

(4) This paper uses the crawling data set for experimental test, and the experimental results show that algorithms proposed in this paper are effective.

2 Natural language cloud database construction

2.1 Natural language cloud database architecture

In this paper, a cloud database architecture for natural language is designed and proposed. As shown in Fig. 1, the architecture of natural language cloud database is composed of two parts: Platform support layer and database layer, and the database layer is built on platform support layer.

CloudStack is the platform supporting layer of cloud database. It is a highly available and extensible open source cloud computing platform. Based on CloudStack, it can quickly and easily create cloud services through the existing infrastructure. Cloud computing platform provides basic support for data analysis and database construction [Lin, Wu, Lin et al. (2017); Jhaveri, Patel, Zhong et al. (2018)].

The database layer is mainly composed of four parts: corpus, thesaurus, word vector library and ontology knowledge base. Corpus and ontology knowledge base are constructed by HBase, and thesaurus and word vector library are constructed by MySQL. The establishment of corpus, thesaurus and word vector library provides a basis for the construction of ontology knowledge base.

The construction technology of thesaurus and word vector library has been mature, so this section will focus on the corpus and ontology knowledge base construction.

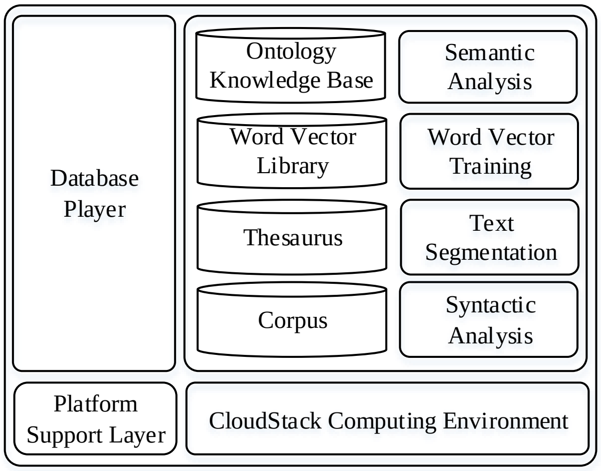

Figure 1: Architecture of natural language cloud database

2.2 Corpus dynamic construction

Corpus can sort, count and classify language data. In 1964, Francis and Kuceral established the first worldwide computer-readable corpus called Brown [Miller, Vandome and Mcbrewster (2010)]. Nowadays, researchers have paid more and more attentions in semantic corpus construction, there are several English corpus, such as LOB, Bank of English and ACL/DCI [Gacitua, Sawyer and Rayson (2008); Greenbaum and Nelson (2010); Davies and Mark (2014)]. Chinese commonly used corpus includes

People's Daily annotation corpus, ZW large General Chinese corpus system, and modern Chinese grammar research corpus, etc. This section will focus on the corpus construction process and the pre-processing of corpus.

2.2.1 Corpus construciton process

As shown in Fig. 2, corpus construction mainly includes 7 steps [Yuan, Chen, Li et al. (2014); Jiang and Wang (2016); Haveliwala (2003)], which are: (1) data acquisition; (2) web links removal; (3) webpage cleaning; (4) webpage parsing; (5) webpage content removal; (6) corpus storage; (7) text vectorization.

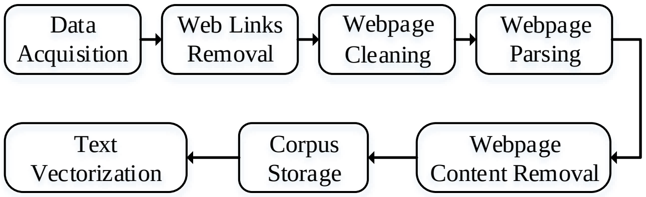

Figure 2: Corpus dynamic construction

2.2.2 TF-IDF and word vector distance based algorithm for duplicated webpages

Web content may be approximately repeated or completely repeated although these webpages have different URL addresses. According to the overall layout and contents of the webpage, duplication may in the following 4 cases: (1) Completely repeated pages: the overall layout and contents of the pages are exactly the same. (2) Content repeated pages: The overall layout of the pages is different, but the contents of the pages are same. (3) Local repeated pages: the overall layout of the pages is the same, but the contents of the pages are different. (4) Partially repeated pages: the overall layout of the pages is different; some webpage contents are the same.

However, web crawler cannot identify the repeated content automatically, so we need an algorithm to eliminate repeated web content. There exist several web de-duplication algorithms, such as: VSM algorithm [Huang (2016)], K-Shingle, Simhash and Minhash algorithm [Oprisa (2015)]. Based on the TF-IDF and the word vector distance, we propose a new web content de-duplication algorithm (TWDW) [Wang (2017)], which will be shown in Fig. 3.

It is described as follows：

(1) Using TF-IDF to get the keywords

Assuming that there is a word *w* in document *d*, the word *w* is not a stop word and the frequency of occurrence in the document is high, then it is assumed that the word *w* is one of the keywords in document *d*. If the occurrence frequency of the word *w* in documents except the document *d* is very low, then the word *w* is easy to distinguish, that is, the importance of the word *w* in document *d* can be evaluated [Chen (2010)]. The main ideas of TF-IDF algorithm are as follows:

a) Firstly, segment the document and remove the meaningless stop words.

b) Calculating term frequency (*TF*) as follows.

$$TF(w,d) = \frac{count(w,d)}{size(d)} \tag{1}$$

$TF(w,d)$ is the term frequency, $count(w,d)$ is the frequency of words *w* in document *d*, $size(d)$ is the total number of words in document *d*.

c) Calculating the Inverse Document Frequency (*IDF*) as follows.

$$IDF(w) = \log(\frac{n}{docs(w,D)}) \tag{2}$$

Where *n* represents the total number of documents in the entire document collection, and $docs(w,D)$ represents the document number with word *w*'s appearance in the entire document set.

d) Calculating the product of *TF* and *IDF* as word *w*'s *TF-IDF*.

$$TF - IDF(w) = TF(w,d) \times IDF(w) \tag{3}$$

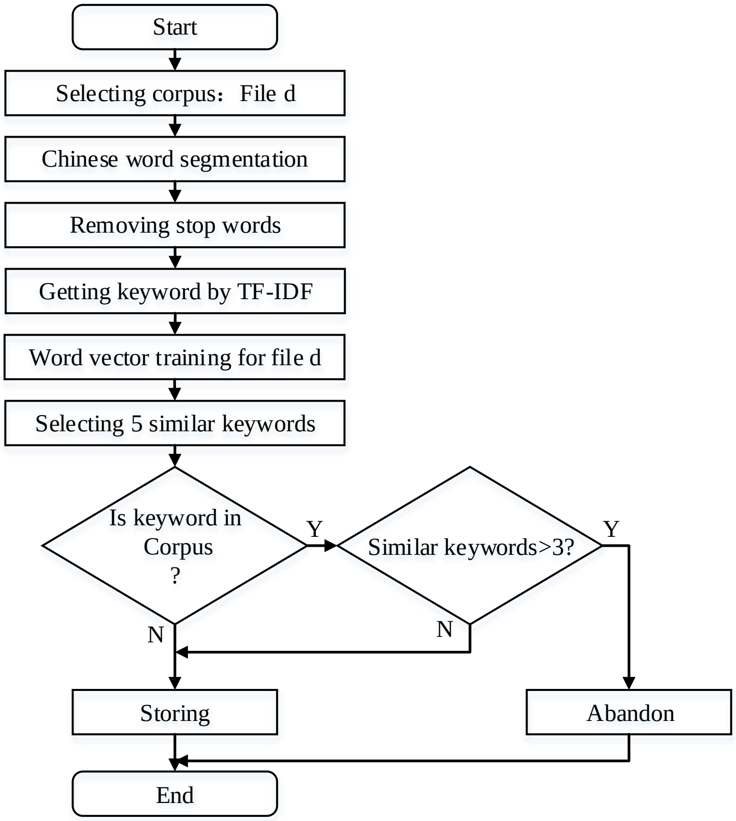
Start
Selecting corpus：File d
Chinese word segmentation
Removing stop words
Getting keyword by TF-IDF
Word vector training for file d
Selecting 5 similar keywords
Is keyword in Corpus ?
Y
Similar keywords>3?
Y
N
N
Storing
Abandon
End

Figure 3: TWDW flow chart

(2) Using Word2vec to train the words, which are similar to the keywords of the document will be extracted.

a) Using Word2vec to train the divided words into bin files.

b) Using the LoadModel method to load the relevant path under the bin file, and selecting the top-5 similar keywords.

c) Judging the repeated pages through word matching. If more than 3 keywords are similar, the web page is regarded as a repeated one, and should not be included.

2.3 Ontology knowledge base construction

Ontology [Greenbaum and Nelson (2010)] begins with the concept of philosophical, it is a formal and explicit statement of shared concepts. After the introduction of artificial intelligence in the early 1990s, ontology is seen as a conceptual modeling tool which can describe the information system at the semantic and knowledge level [Wang, Wu and Ren (2014)]. At present, there have been many research results based on semantic ontology in the field of Natural Language Processing, and the most prominent are WordNet and HowNet [Ren and Guo (2012)]. This paper discusses the construction principle of ontology knowledge base and the representation method of ontology.

2.3.1 Construction principle

In 1990s, Gildea et al. [Gildea (2000)] proposed shallow semantic analysis, which is based on "Predicate-Argument (PA) structure". Deep semantic analysis, based on shallow semantic analysis, implements a complete mapping from the PA structure to the semantic structure. The background knowledge ontology is the bridge between the PA structure and the semantic structure. Knowledge base develops continuously, and the semantic construction which seems it as background knowledge is more feasible [Yuan (2012)]. Therefore, the ontology knowledge base is used as the background knowledge ontology of the semantic construction, and the ontology knowledge base should be designed according to the characteristics of the PA structure. It mainly satisfies three principles: (1) The mapping of PA structure to knowledge structure is convenient; (2) Covering almost all PA structures; (3) It is easy to optimize the expression of knowledge in future.

2.3.2 Ontology description method

As shown in Fig. 4, PA structure is composed of predicate, argument and argument semantic role. Predicate is the core vocabulary in an arbitrary sentence, which can express the main idea of the sentence. Argument is the modification and supplement of predicate, and semantic role argument mainly determines the semantic relationship between predicate and argument, a semantic role may contain more than one argument.

The structure of background knowledge ontology is similar with PA structure, it is represented with the Event-Property (EP) structure [Liu (2013)]. In Fig. 5, E represents the event, P represents the property of the event, and Ele represents the element of the event. Background knowledge ontology including event ontology and argument ontology. Event ontology corresponds to the predicate in PA structure, which describes the concept of verb. There is a generic event class in the event ontology, and the other events are its subclass. Argument ontology corresponds to the argument in PA structure, it describes nominal concept. According to the semantic characteristics of PA structure, the argument class ontology can be divided into subclass: such as time, place, personal pronoun and so on.

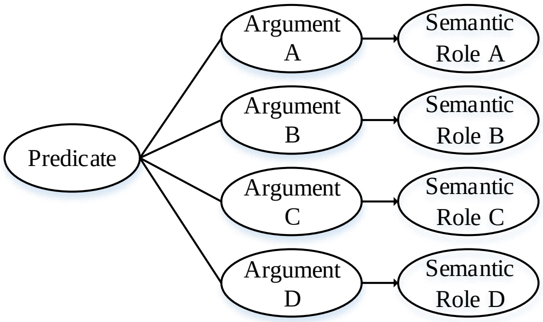

Figure 4: PA structure

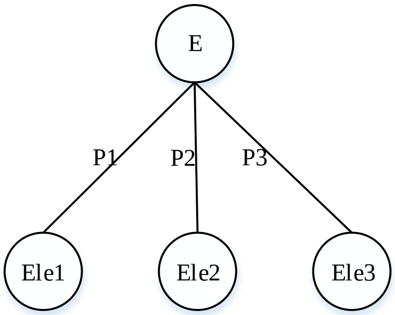

Figure 5: EP structure

Background knowledge ontology is described in the RDF(S) language, which is defined as follows:

(1) Generic event class definition.

<event,rdf:type,rdfs:Class>

(2) Predicate P and its six core semantic roles definition.

<P, rdf:type, rdf:property>
<A0, rdf:type, rdf:property>
<A1, rdf:type, rdf:property>
……
<A5, rdf:type, rdf:property>

(3) 15 additional semantics definition.

<ADV, rdf:type, rdf:property>
<BNE, rdf:type, rdf:property>
<CDN, rdf:type, rdf:property>

……

<PSE, rdf:type, rdf:property>

(4) Semantic constraints definition. Using TMP semantics as an example.

<TMP, rdf:domain, event>

<TMP, rdf:range, Date>

3 Dynamic construction of natural language semantics

3.1 Semantic construction process

Dynamically natural language semantic construction is based on ontology knowledge base. Fig. 6 represents the detailed process of semantic dynamic construction. First, using syntactic analysis to get the syntactic structure of sentences; Secondly, according to the semantic role labeling method, the syntactic structure is transformed into PA structure; Then, transforming the PA structure into semantic structure based on the background knowledge ontology, and finally the user evaluation is carried out. This process uses unsupervised semantic role training to label large-scale corpus, and implements supervised semantic role labeling for newly added text, it is based on trained PA structure and semantic annotation result of text set. The process realizes the dynamic incremental construction of the natural language semantics, and effectively improves the efficiency of semantic extraction. This paper will focus on the establishment of the mapping between PA structure and background knowledge ontology.

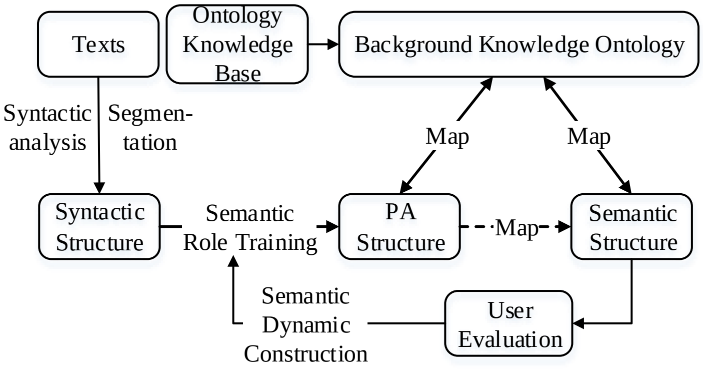

Figure 6: Semantic dynamic construction

3.2 Establishment of mapping process

3.2.1 Mapping process

The establishment of mapping relationship between PA structure and background knowledge ontology is important for the construction of natural language semantics [Liu (2013)]. The center point of the PA structure is the predicate P, and the argument element is mainly the modification and supplement of the predicate. The "event" is the basic unit of the background knowledge ontology, and each event contains the corresponding relation of elements and attributes. The ontology structure of background knowledge is very similar to PA structure. Therefore, predicates in PA structure can be mapped to event in background knowledge ontology, and then the argument is mapped to the elements contained in the event. The mapping process between PA structure and EP structure is shown in Fig. 7.

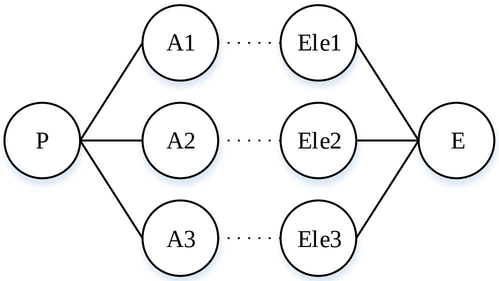

Figure 7: Mapping relation

3.2.2 Mapping algorithm based on semantic similarity

Mapping between the PA structure and background knowledge ontology is mainly divided into two steps [Liu (2013); Chen, Chen, Nagiza et al. (2014); Zhang, Yang, Xing et al. (2017)]: one is the semantic similarity calculation of predicate P and all of the events, which can get a matching set; two is to match all elements of matching set and argument corresponding to predicate P, and find the most matching events.

This paper optimizes and improves the mapping algorithm provided by document [Liu (2013)], and proposes a mapping algorithm based on semantic similarity (MBSS) to construct mapping relationship between PA structure and background knowledge ontology, which will be shown in Fig. 8.

The main idea are as follows:

(1) Set representation

PA structure is represented with two-tuple $PA=<P, S_p>$, P represents a predicate, S_p represents all the semantic roles contained in the predicate, each semantic role contains a collection of argument, denoted by A_s; EP structure is shown in two-tuple $ED=<e, D_e>$, E on behalf of the event, D_e on behalf of the event elements.

(2) Calculating the matching event set of predicate *P*

An important part of establishing the mapping relation between PA structure and background knowledge ontology is to find the matching event set E_p of predicate P. This paper uses vector space model to calculate semantic similarity of predicate P and event E, If the similarity is greater than 0.5, then adding the event E to the set of events set E_p. The detailed are as follows:

a) Using CBOW model to train the word vector of the large-scale corpus, and obtaining the word vector of predicate P and event E, the predicate word vector is shown as $P=\{P1，P2,\cdots，Pn\}$, event word vector is shown as $E=\{E1，E2,\cdots，Em\}$, CBOW training formula is shown as Eq. (4).

$$P(w(t)) = P(w \mid w(t-n), w(t-n+1), \cdots, w(t+n-1), w(t+n)) \tag{4}$$

b) Computing the word vector distance *d*(*P*, *E*).

$$d(P,E) = \sum_{j=1}^{m} \sqrt{(P_i - E_j)^2} \tag{5}$$

c) Computing the semantic similarity *sim*(*P*, *E*) between predicate *P* and event *E*, as shown in Eq. (6). If *sim*(*P, E*)*>0.5*, then the event *E* is placed in the matching event set E_p, and the semantic similarity value between the event and the predicate is placed in the similarity set S_{sim}.

$$sim(P,E) = \begin{cases} \dfrac{\alpha}{d(P,E)+\alpha}, & d(P,E) > 0 \\ 0, & \text{others} \end{cases} \tag{6}$$

Where α is seemed as an adjustable parameter.

(3) Getting the most matching event e_{max} of the predicate *P*

Finding the most matching event e_{max} by calculating the semantic similarity of two-word clusters. Matching rules [Liu (2013)] are shown in Tab. 1.

Table 1: Matching rules of argument and element

Rules	Rule description
Rule 1	If a_i or its synonym is the same as the name of *d*, then a_i is the matching argument of *d*.
Rule 2	If a_i or its synonym is the same as the names of any descendants of *d* (judged by subClassOf), then a_i is a matching argument of *d*.
Rule 3	If a_i or its synonym is the same as any name in the instance set I_d of *d*, then a_i is a matching argument of *d*.

Calculating process are as follows:

a) Calculating the ratio of matching argument, that is calculating the proportion of all argument under each semantic role and matching event element.

$$M_{s \to d} = \frac{\left|\{a_i \mid a_i \in A_s, Match(a_i) \in d\}\right|}{|A_s|} \tag{7}$$

$Match(a_i) \in d$ represents that a_i is the matching argument of event element *d*.

b) Calculating the matching degree between the whole PA structure and the event propery structure. The calculation formula is as follows.

$$M_{PA \to ED =} \frac{\sum_s \max_d M_{s \to d} \times |A_s|}{\sum_s |A_s|} \times \beta \tag{8}$$

$$\beta = \frac{\mathrm{sim}_i}{\sum_{i=1}^{k} \mathrm{sim}_i} \tag{9}$$

β is the similarity weight, sim_i represents the semantic similarity between the event i and the predicate P, k is the number of events in event set that matching the predicate P.

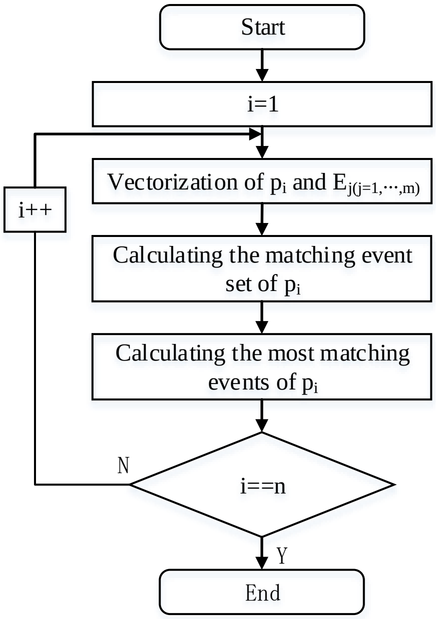
Start
i=1
Vectorization of p_i and $E_{j(j=1,\cdots,m)}$
i++
Calculating the matching event set of p_i
Calculating the most matching events of p_i
N
i==n
Y
End

Figure 8: MBSS flow chart

4 Experimental verification and analysis

This section provides a scientific assessment of the proposed TWDW algorithm and MBSS algorithm. The main criteria are Precision and Recall [Zhou (2012)]. A good algorithm often has both high precision and recall ratios, while these two standards are mutually constrained and researchers can only take an appropriate method to keep both at a relatively high level.

4.1 Experiment of TWDW algorithm

Experiment mainly includes three aspects, first, verifying the selection of keyword extraction algorithm, second, verifying the selection number of repeated keywords, and then verifying the effectiveness of webpages removal algorithm.

Experimental data sets are crawled from webpages by WebMagic crawler, including: (a) Crawled 500 texts, containing 50 texts with approximately duplicated content, (b) Crawled 500 texts, including 25 texts with completely duplicate content, and 25 texts with similar content.

4.1.1. Verifying the validity of the keyword selection algorithm

The commonly used keyword extraction algorithms mainly include the following three: TextRank algorithm [Mihalcea (2004)], Latent Dirichlet allocation [Blei, Ng and Jordan (2003)] and TF-IDF statistical method. In this paper, three algorithms are used to extract keywords from a section of text. The text is as follows:

Natural language processing is an important direction in the field of computer science and artificial intelligence. It studies various theories and methods that enable effective communication between people and computers using natural language. Natural language processing is a science that integrates linguistics, computer science, and mathematics.

Table 2: Experimental results of keywords extraction

	TextRank		LDA		TF-IDF	
	Keyword	Weight	Keyword	Weight	Keyword	Weight
1	science	97	natural language processing	0.1302	natural language processing	0.0986
2	computer science	7	computer science	0.0739	computer science	0.0872
3	linguistics	7	artificial intelligence	0.0739	artificial intelligence	0.0872
4	artificial intelligence	7	study	0.0739	linguistics	0.0872
5	mathematic s	7	computer	0.0739	mathematics	0.0872
6	theory	5	natural language	0.0739	computer	0.0706
7	study	5	communication	0.0739	natural language	0.0706
8	natural language processing	5	linguistics	0.0739	communication	0.0651
9	natural language	3	mathematics	0.0739	study	0.0651
10	computer	3	science	0.0739	field	0.0651

In this paper, the experimental results are arranged in descending order. The top 10 keywords obtained by each algorithm are shown in Tab. 2. The decimal numbers in the table indicate the corresponding weight of keywords in each method.

Tab. 2 shows that the TF-IDF and TextRank algorithm are better than the LDA model in weight discrimination. Compared with the TextRank algorithm, the TF-IDF is more in line with human judgment. Therefore, using the TF-IDF algorithm to extract keywords is more appropriate.

4.1.2 Verifying the selection number of repeated keywords

Using data (a) as a data set, when the number of matches is 2, 3, and 4, the precision and recall of the algorithm are shown in Tab. 3.

Table 3: Experimental results of matching numbers

	Matching numbers	Actually repeated documents	Detected repeated documents	Detected correctly repeated documents	Precision (%)	Recall (%)
1	2	50	49	41	83.6	82
2	3	50	45	42	93.3	84
3	4	50	42	39	92.9	78

It can be seen from the experiment in Tab. 3 that, when matching number is 3, the precision and recall of the document are relatively high and stable, so it is appropriate to set the matching number to be 3.

4.1.3 Verifying the effectiveness of webpages removal algorithm

Using data (b) as a data set to compare Simhash algorithm and TWDW algorithm, the precision and recall of them are shown in Tab. 4.

Table 4: Experimental results of algorithm comparison

	Algorithm	Actually repeated documents	Detected repeated documents	Detected correctly repeated documents	Precision (%)	Recall (%)
1	Simhash	50	41	35	85.3	70
2	TWDW	50	48	46	95.8	92

It can be seen from Tab. 4 that the TWDW algorithm performs obviously better in the precision and recall, compared with Simhash algorithm. At the same time, comparison between Tab. 3 and Tab. 4 shows that, precision and recall rates of TWDW algorithm are higher when the texts are similar.

4.2 Experiment of MBSS algorithm

Experiment mainly includes two aspects, first, verifying the performance of improved semantic similarity calculation method of predicate and event, and then verifying the effectiveness of optimized algorithm.

Experimental data set: (a) PA structure of unsupervised semantic role training for the corpus constructed in this paper, (b) EP structure constructed in this paper, (c) SemLink obtained from network.

4.2.1 Verifying the performance of semantic similarity algorithm

Using semantic similarity algorithm provided by MBSS(SS-MBSS) and semantic similarity algorithm provided by literature [Liu (2013)] (SS-[Liu (2013)]) to calculate the matching event set of predicate P. As shown in Tab. 5, one of the experimental results is selected in this paper.

Table 5: Experimental results of semantic similarity calculation

	Predicate	Event	SS-MBSS algorithm	SS-[Liu (2013)] algorithm
1	research	explore	0.855	1
2	research	query	0.927	1
3	research	dig	0.911	1
4	research	discuss	0.855	1
5	research	deliberate	0.713	1
6	research	elaborate	0.654	0
7	research	seek	0.601	0
8	research	search	0.496	0
9	research	hammer	0.513	1
10	research	debate	0.374	0

It can be seen from Tab. 5 that the result of SS-[Liu (2013)] algorithm is relatively rough, and it only determines whether events match with predicates. The result of SS-MBSS algorithm is more accurate, so the SS-MBSS algorithm is more reasonable than SS-[Liu (2013)] algorithm.

4.2.2 Verifying the performance of MBSS algorithm and algorithm provided by literature

Using MBSS algorithm and algorithm provided by literature [Liu (2013)] (A-[Liu (2013)]) to construct the mapping relationship between PA structure and background knowledge ontology, and the precision and recall are calculated by referencing to the SemLink. In this experiment, data (b) and 500 PA structure from data (a) are taken as an example, and the results are shown in Tab. 6.

Table 6: Experimental results of Mapping algorithm comparison

	Algorithm	Correct EP structure	All found EP structure	Correctly found EP structure	Precision (%)	Recall (%)
1	MBSS	500	491	391	79.63	78.2
2	A-[Liu (2013)]	500	480	354	73.75	70.8

As shown in Tab. 6, the precision and recall of MBSS algorithm are significantly improved compared with algorithm provided by literature [Liu (2013)]. Therefore, the MBSS algorithm is more accurate while constructing the mapping relationship.

5 Conclusions

Understanding semantics correctly is the basis for the realization of natural language interaction. This paper proposes a method of natural language semantic construction based on cloud database, analyzes and optimizes the construction technology of natural language cloud database and semantic structure. Meanwhile, this paper proposes TF-IDF and word vector distance based webpages de-duplication algorithm and mapping algorithm based on semantic similarity, both of them are proved effective. The work of this paper improves the understanding of natural language semantics, and provides a good data support for the natural language interaction function of the cloud service. However, our work also has some shortcomings, such as the semantic construction method proposed in this paper is more applicable to the verb structures, which needs to be optimized in the future.

Acknowledgement: This paper is partially supported by the Natural Science Foundation of Hebei Province (No. F2015207009) and the Hebei higher education research project (No. BJ2016019, QN2016179), Research project of Hebei University of Economics and Business (No. 2016KYZ05) and Education technology research Foundation of the Ministry of Education (No. 2017A01020). At the same time, the paper is also supported by the National Natural Science Foundation of China under grant No. 61702305, the China Postdoctoral Science Foundation under grant No. 2017M622234, the Qingdao city Postdoctoral Researchers Applied Research Projects, University Science and Technology Program of Shandong Province under the grant No. J16LN08, and the Shandong Province Key Laboratory of Wisdom Mine Information Technology foundation under the grant No. WMIT201601.

References

Beale, T. (2002): Archetypes constraint-based domain models for futureproof information systems. *Oopsla Workshop on Behavioural Semantics*, pp. 1-18.

Blei, D. M.; Ng, A. Y.; Jordan, M. I. (2003): Latent dirichlet allocation. *Machine Learning Research Archive*, vol. 3, pp. 993-1022.

Breitbach, W. (2010): Google voice. *Charleston Advisor*, vol. 12, no. 1, pp. 20-22.

Chen, H. L. (2010): *News Retrieval System Based on Automatic Web Page Classification (Ph.D. Thesis)*. Harbin Institute of Technology, China.

Chen, W. B.; Chen, Z. Z.; Nagiza, F.; Samatov.; Peng, L. X. et al. (2014): Solving the maximum duo-preservation string mapping problem with linear programming. *Theoretical Computer Science*, vol. 530, pp. 1-11.

Davies, M. (2014): The corpus of contemporary American English (COCA). *Canadian Journal of Chemistry*, vol. 92, no. 1, pp. 9-15.

Gacitua, R.; Sawyer, P.; Rayson, P. (2008): A flexible framework to experiment with ontology learning techniques. *Knowledge-Based Systems*, vol. 21, no. 3, pp. 192-199.

Gildea, D.; Jurafsky, D. (2000): Automatic labeling of semantic roles. *Computational Linguistics*, vol. 28, no. 28, pp. 245-288.

Greenbaum, S.; Nelson, G. (2010): The international corpus of English (ICE) project. *World Englishes*, vol. 15, no. 1, pp. 3-15.

Haveliwala, T. H. (2003): Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 784-796.

He, P.; Deng, Z. L.; Gao, C. Z.; Wang, X. N.; Li, J. (2017): Model approach to grammatical evolution: deep-structured analyzing of model and representation. *Soft Computing*, vol. 21, no. 18, pp. 5413-5423.

He, P.; Deng, Z. L.; Wang, H. F.; Liu, Z. F. (2016): Model approach to grammatical evolution: theory and case study. *Soft Computing*, vol. 20, no. 9, pp. 3537-3548.

Huang, J. (2016): An assignment duplicate-checking algorithm based on semantic vector space model. *Electronic Science & Technology*, vol. 3, no. 6, pp. 786-789.

Jhaveri, R.; Patel, N.; Zhong, Y.; Sangaiah, A. (2018): Sensitivity analysis of an attack-pattern discovery based trusted routing scheme for mobile Ad-Hoc networks in industrial IoT. *IEEE Acess*, pp.1-1.

Jiang, L.; Wang, D. B. (2016): Automatic extraction of domain terms using continuous bag-of-words model. *New Technology of Library & Information Service*, vol. 2, pp. 9-15.

Lin, W. W.; Wu, Z. M.; Lin, L. X.; Wen, A. Z.; Li, J. (2017): An ensemble random forest algorithm for insurance big data analysis. *IEEE Access*, vol. 5, pp. 16568-16575.

Liu, F. (2013): *Research on Semantic Information Extraction Based on Predicate-Argument Structure (Ph.D. Thesis).* Ocean University of China, China.

Masterson, M. (2012): Siri, meet Nina. *Speech Technology Magazine*, vol. 17, no. 5, pp. 6-17.

Mihalcea, R.; Tarau, P. (2004): TextRank: Bringing order into texts. *Conference on Empirical Methods in Natural Language Processing*, vol. 85, pp. 404-411.

Miller, F. P.; Vandome, A. F.; Mcbrewster, J. (2010): *Brown Corpus*. Alphascript Publishing.

Misra, D. K.; Sung, J.; Lee, K.; Saxena, A. (2014): Tell me Dave: context-sensitive grounding of natural language to manipulation instructions. *International Journal of Robotics Research*, vol. 35, pp. 1-3.

Oprisa, C. (2015): A MinHash approach for clustering large collections of binary programs. *International Conference on Control Systems and Computer Science*, vol. 6, pp. 157-163.

Ren, W. L.; Guo, J. J. (2012): Word similarity algorithm based on WordNet and HowNet. *Applied Mechanics & Materials*, vol. 155-156, pp. 375-380.

Shum, H. Y.; He, X. D.; Li, D. (2018): From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 10-26.

Wang, H. Y.; Wu, X. S.; Ren, J. Z.; Zhao, Y. L. (2014): The research and implementation of the MinHash algorithm in distributed platform. *Intelligent Computer and Applications*, vol. 4, no. 6, pp. 44-46.

Wang, S. Z.; Zhang, Q. C.; Zhang, L. (2017): Natural language semantic corpus construction based on cloud service platform. *Proceedings of 2017 International Conference on Machine Learning and Cybernetics*, vol. 2, pp. 670-674.

Yuan, D. (2012): *The Research of Semantic Construction Method Based on Massive Text (Ph.D. Thesis).* Ocean University of China, China.

Yuan, Y. W.; Chen, D.; Li, Y.; Yu, D. J.; Yan, L. M. et al. (2014): The improved shark search approach for crawling large-scale web data. *International Journal of Multimedia & Ubiquitous Engineering*, vol. 9, no. 8, pp. 251-260.

Zhang, S. H.; Yang, Z. B.; Xing, X. F.; Gao, Y.; Xie, D. Q. et al. (2017): Generalized pair-counting similarity measures for clustering and cluster ensembles. *IEEE Access*, vol. 5, pp. 16904-16918.

Zhao, Q. F.; Li, H. Q.; Jiang, T. F. (2016): *One Kind of Human-Computer Interaction in Natural Language Understanding Intent Method.*

Zhou, Y. (2012): Research on web pages removal algorithm based on pre-classification of text length and key sentences. *Software Guide*, vol. 10, pp. 48-50.