Research on Natural Language Processing and Semantic Analysis Model Application Based on Conceptual Graphs

Duanming Shen*, Qing Li, Dexin Qiao, Xingwen Zhou
Computer Application Technology Department, China.
PetroChina Research Institute of Petroleum Exploration & Development, Beijing 100083, China.

* Corresponding author. Tel.:010-83595942; email: shenduanming@petrochina.com.cn
Manuscript submitted December 28, 2019; accepted January 20, 2020.
doi: 10.17706/jsw.15.2.45-52

Abstract: With the rapid development and allround popularization of artificial intelligence, all walks of life are also trying their best to promote the cross integration of information in different fields, using the Internet to promote industrial transformation, and promoting the transformation of industrial economy to information economy. Therefore, semantic understanding and text analysis are more and more indepth research in enterprise information intelligence. In order to meet the needs of effective analysis and processing a large amount of information data in key construction projects of large-scale energy enterprises, this paper proposes a natural language processing and semantic analysis model based on conceptual graphs design. The content of the unstructured data collected is studied by unsupervised machine learning according to the organization and representation knowledge in the conceptual graphs model, and then it is selfcontained the function of dynamic recognition of text semantics through text analyzer, and output the corresponding learning feedback results. The practical application results show that the design model is feasible, which significantly improves the learning effect and the accuracy of information screening, and also provides strong support for the follow-up big data analysis..

Key words: Machine reading comprehension, natural language processing, semantic analysis.

1. Research background

In the current high-level and deep integration of informatization and industrialization, enterprises are more and more dependent on their own processing and analysis of industrial information, so as to create a new development path of promoting industrialization with informatization and promoting informatization with industrialization. Therefore, how to use information retrieval to manage the knowledge and data of the whole enterprise becomes particularly important and urgent. The focus of semantic analysis is to evaluate the sensitivity and importance of key words of data information, analyze the potential intention and target range of statements, and accurately split the statements. In the industry, the innovation and exploration of information retrieval has been constantly improved, and achieved good results. In terms of analysis efficiency, it greatly improves the time and cost of machine learning. However, how to better identify the intention of user retrieval and the decomposition of articles in professional fields is a problem that needs to be solved at this stage. It includes two levels of knowledge representation and cognitive schema, elements of knowledge node, knowledge association, cognitive state and learning path [1]; it model attributes of peoples' sensory, motor, social, emotional, and cognitive experiences with words using behavioral ratings[2]; The MRC network based on semantic conflict detection (SCDNET) proposed by SCDNET aims to detect the

problem raised by semantic conflict detection network, SCDNET predicts Na probability by checking whether the article contains the overall semantics of the problem [3]. However, none of the above methods deal with the noise and external influence from the vocabulary itself, and cannot effectively separate and analyze the data content according to the actual needs and introduce external influence, so that the computer can effectively understand the key intention of user retrieval.

Therefore, this paper proposes a natural language processing and semantic analysis application technology based on conceptual graphs design, which has a more efficient and accurate semantic analysis ability, and with the help of unsupervised machine learning, the model is deployed in a distributed environment for continuous training, which solves the problems of semantic understanding, statement splitting, part of speech judgment and future information analysis. The problem of deep mining of all data realizes the ability to upgrade the learning prototype in the later stage. The practical application results show that the construction scheme of the semantic analysis processing is feasible, which significantly improves the machine's ability to understand and analyze all kinds of information data.

2. Language Preprocessing Model Based on Conceptual Graphs Design

In 1984, John Sowa first proposed a conceptual graphs concept, which is a limited, connected and two-part map [4]. It can be understood by people and machines to realize the communication between people and machines. The design principle of conceptual graphs emphasizes the requirements of cognitive representation: smooth mapping with natural language, "icon" structure, which is used to represent the perception mode in visual and tactile images; and cognitive reality operation for perception, reasoning and language understanding. conceptual graphs (CGS) is inspired by existence map (EGS) in semiotics, which is developed by John on the basis of Pearce's EGS theory CGS has been widely used in the field of computer and artificial intelligence [5]. At the same time, the international organization for standardization has designed some standards for CGS, which can prove the importance of CGS in this field. His concept graph theory has been widely used in the research and application of artificial intelligence.

There is a very simple example to explain the conceptual graphs: if we want to use CGS to express a sentence: "John is going to Boston by bus", then we will use Fig. 1 to show the process.

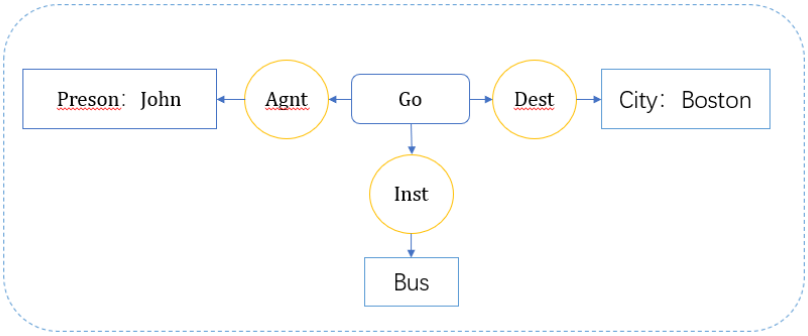

Fig. 1. The graph of “John is going to Boston by bus.”

Fig. 1 shows a conceptual graph with four concepts: [Go], [Person: John], [City: Boston], and [Bus]. It has three conceptual relations: (Agnt) relates [Go] to the agent John, (Dest) relates [Go] to the destination Boston, and (Inst) relates [Go] to the instrument bus.

What can we know from this graph? There is an intersection “GO”, so it will be showed:

[Go]-(Agnt)->[Person:John] (Dest)->[City: Boston] (Inst)->[Bus].

This example resembles frame notation, but linear form also permits co-reference labels to represent the cross references needed to represent arbitrary graphs.

Based on the prototype of CGs[6], John give a new method which called the Conceptual Graph Interchange Format (CGIF) [7], is one of the three standard dialects for Common Logic (ISO/IEC 24707). CGIF has a one-to-one mapping to and from the nodes of the display form. CGIF is designed for man-machine communication, so its syntactic forms are simpler and requirements of character set are stricter.

In the following CGIF representation for Figure 1, each concept has its own defining label:

[Go:*x] [Person: John *y] (Agnt?x?y) [City: Boston *z] (Dest ?x ?z) [Bus: *w] (Inst ?x ?z)

By nesting some of the concepts inside the relations, the CGIF form can be limited to just a single defining label *x and a bound label ?x inside each relation node:

[Go *x] (Agnt?x [Person: John]) (Dest?x [City: Boston]) (Inst ?x [Bus])

CGIF is used to deal with IT information exchange format, if it will exchange information with other system, so it will be swapped to Knowledge Interchange Format (KIF) [8]:

(exists ((?x Go) (?y Person) (?z City) (?w Bus)) (and (Name ?y John) (Name?z Boston) (Agnt?x?y) (Dest ?x ?z) (Inst ?x ?w)))

In order to solve the problem of single data acquisition and data collection for different dimensions, the system will include the router in the whole network fault detection conditions and use the remote monitoring mode to collect and analyze the data of router and load balancing.

As you can see, DF, LF, CGIF and KIF look different in format, but they are all based on the same logical concept, so their semantics are exactly the same, which means they can transform each other without losing information. The graph also has a highly regular structure, which can simplify many algorithms of search, pattern matching and reasoning. While no one knows how any information is represented in the brain, graphics minimize extraneous details: they display connections directly and avoid the order implied in strings or trees.

3. Semantic State Analysis

Conceptual graphs are not only used in some special languages, but also a general language structure according to the definition of linguistics. Every language is just an example. Natural language is a highly expressive system, which can express anything in any form of language or logic [9]. This huge expressive force makes it difficult or impossible for any formalism to express every feature of every natural language. Before that, we built a language preprocessing model based on conceptual graphs, which regards text parsing as a complete dynamic process, and any parsed words are generated based on conceptual graphs.

Whether DF or LF, they all involve the communication between human and machine, so they all involve natural language and need to deal with semantic relevance [10]. In this study, not only from "being" to "concept", but also expand and optimize the understanding of part of speech and the relevance between them, so that it has a multi-level semantic equivalent that can represent multiple modes, time and intention logic. What can we do in this mode? In natural language, a sentence may have different ways of expression for different people and places, with different contextual semantic connections and state related output phrases.

3.1. The Algorithm of Correlation Smoothing

In the daily semantic analysis, in order to more approximate the original meaning of the sentence, we need to carry out "Semantic Smoothing" processing on the above model. The purpose of smoothing is to show that some words do not have their own meaning, but also appear frequently, and have nothing to do with the meaning of the discourse itself. First, we introduce a term P(w) into the log linear model, where P(w) is the relevance probability of words (in the whole corpus) is scalar, defining the variable parameter

TimeL in the time period, the purpose of this parameter is to indicate that the correlation V_w between the context and the word is captured in the interior between the word vector C_t and the word vector in a certain period of time, which improves the common occurrence probability of the words with high component C_t. Specific formulas are as follows:

$$P[w\ at\ time\ t\ (C_t)] \propto TimeL\ (C_t, V_w) \tag{1}$$

In the process of execution, word vector C_t will change the overall relevance of the sentence due to the change of time, which means that these key words, such as time, place, person, event and other words, will lead to vector offset and semantic change of relevant words according to the change of external environment. However, we assume that when inputting words in a sentence, the utterance vector C has no actual meaning and is used frequently, resulting in little change in the sentence. Therefore, we can use a single vector C_s to replace all C_t in the sentence. C_s stands for a set of words with clear nature and state after filtering meaningless words or words that need to be excluded in a certain period of time This paper defines a common word vector C_0, which is used to correct deviation as the most frequently used word correction term related to grammar. Therefore, the expression can be expressed as:

$$\check{C}_S = \alpha C_0 + (1-\alpha) C_s \tag{2}$$

In this paper, we introduce x into y to further correct the semantic deviation caused by time deviation, which makes the high relevance words appear, even though their vectors have a very low internal vocabulary frequency, but after the correction, they still maintain a high relevance impact. Specifically, given the word vector C_s, the probability scalar P(w) of relevance caused by word w in a sentence is calculated as follows:

$$P[w\ in\ sentence\ s\ (C_s)] = \beta p(w) + (1-\alpha)\frac{TimeL\ (\check{C}_s, V_w)}{Z_{\check{C}_s}} \tag{3}$$

α and β are constant weight parameters, which are set according to the specific use scenario. $Z_{\check{C}_s}$ is the relevance evaluation value of the word vector $\check{C}_S$ in the whole history accumulation for a particular length of time.

3.2. Lexical State Assignment Model

In the process of smoothing the part of speech relations, the above Association optimization algorithm inherits the filtering and filtering method of words independent of itself in a linear form. On this basis, according to the needs of the actual scene processing, the lexical state model designs some semantic understanding processing methods under special scenes. For example, each term refers to the description of general concept types, and each sentence semantics is composed of multiple associated $P[w_{n\in v}]$. Each concept relationship determines whether these concept types can be combined. Therefore, in this logic method, complex sentences can be effectively identified by machine splitting and combining.

The semantic evaluation ass, f_t is the most influential word in a certain period of time. Combined with the relevance probability P(w) caused by the word in the context, we can calculate the real nature and semantics of a document in the current cycle of the word:

$$Ass(w\ in\ document) = \sum_x exp\ (P_w f_t(\underset{T,T>0}{max}, Proc_y)) \tag{4}$$

In the task of word segmentation, when each word is input, the algorithm will determine whether the

word is already a saved word in the thesaurus or start to register a new word. At the same time, we will assign initial vector value to the term according to the periodic search heat and current semantics to ensure that it is within the controllable range of machine understanding.

Input: raw sentence sent – a list of words
Initialization: set a single of sentence S, Word splits $\{P_w: w \in S\}$, Semantic value $\{Ass_w: w \in S\}$, set src = [],
Variables: candidate sentence items – a list of relationship words $\{List_{1\to n}= [P_w, P_{w+1},..., P_{w+n}]\}$

```
Output：
for all sentence S in document D :
  var char = S[index]
   foreach itemWord in S[index]{ List_{1→n}};
      if append as a new word to the candidate:
          dictionaryWords.append(itemWord)；
      then src.insert(itemWord)
     end for
 For all sentence S in document D :
      Form a matrix D_M which input {src_w, Ass_w};
End for
```

Through the matrix D_M, we can know the weighted average value of the word in the sentence. For more frequent words, P(w) will also be reduced and controlled correspondingly. With the change of time, the weight will also be reduced correspondingly. This mode not only speeds up the effectiveness of training, but also speeds up the accuracy of the machine in vocabulary semantic understanding and learning, and also has good recognition for derivative words.

4. Results and Analysis

First, the distributed virtual machine environment is deployed based on Windows 2012. Initialize 1200 sample files of 4 file categories and their subclasses, and import them into database and analyzer model. Secondly, the text is disassembled and analyzed. The specific machine information is as follows:

Table 1. Performance Test Environment Configuration Table

Configuration item	parameter
CPU	8 Core
Memory	8G
Hard disk	500G
operating system	Windows Server2012 R
RabbitMQ	V 3.6.10

Based on the above-mentioned test environment, we will input 1000 learning samples for analysis in this study. First, we will calculate the four types of data models with the most features after extraction and analysis, and then we will talk about their relevance in PPPP, get the impact factors of this kind of vocabulary in the time cycle, and prove the interference of external environment on the data. Secondly, we input all the data into ass for optimization and correction, and compare it with the uncorrected data to determine the optimal solution. See Table 2 for details.

Table 2. Optimal Results Experimental Scheme

NO	Test Case	Content	Purpose
1	Input 1000 sample documents, including information in energy, finance, education, oil and gas exploration and other fields	Test the semantic relevance in the change samples after time	Find the fluctuation trend of semantics in time change
2	The entered information is input into Ass（w）for calculation and compared with the unoptimized results	Draw the result scatter diagram after calculation	Analyze the actual changes of the modification of the optimization algorithm and get the optimal solution

This test will deploy five virtual machines as a cluster environment, start different number of persistent queues in each virtual machine, link a total of 4 different document types, a total of 20 producers, 24 consumers, one channel in each producer connection, and two channels in each consumer connection, to ensure the stability of data analysis.

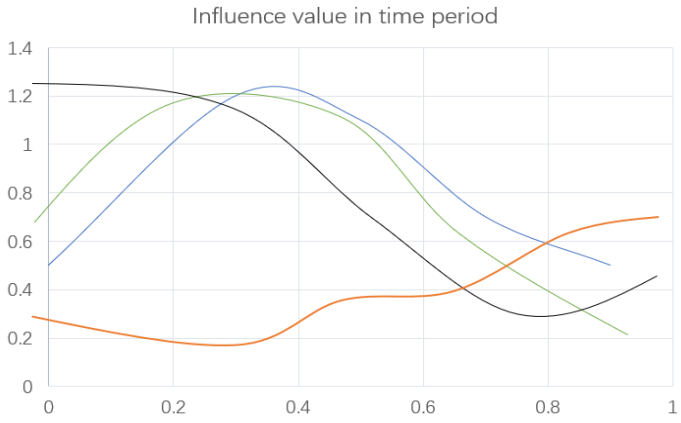

Fig. 2. The graph of Influence value in time period.

As can be seen from the Fig. 2, with the change of time, the semantics and influence of vocabulary are decreasing day by day in a cycle time. At the same time, the diversity of some semantic definitions makes the part of speech association that needs to be concerned more and more complex for machine understanding, and the performance and efficiency of understanding processing mechanism need to be stronger and stronger. Using the optimized method, we can split the required information text, and then carry out association analysis, and make a reverse comparison of the data in the existing semantic database to ensure that the text data can be executed according to the expectation of the current semantics.

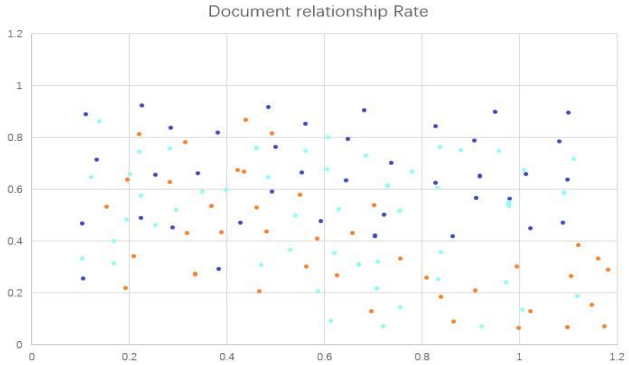

Fig. 3. The graph of Document relationship Rate.

From the output data on the graph 3, the accuracy will decline when the initial data increases, but with

the growth of time, because the algorithm considers the context and external influence, the blue part of the data will constantly modify the actual semantic target and the text matching degree, while the text information (yellow and blue parts) which is not optimized by the algorithm is parsed out The result shows the intention of breaking away from the original words. Therefore, it is proved that the algorithm can guarantee the correctness and stability of the semantic understanding results, and provide a satisfactory parsing list with the change of time and data correlation.

5. Conclusion

In order to achieve faster and more accurate machine identification and information data understanding ability, we build a language preprocessing model based on concept map design, and then use unsupervised learning method to machine learn the initialization prototype of discrete sub nodes. This method uses semantic state analysis technology to build semantic correlation resource pool, and effectively associates the basic semantics, external information and periodic influence of text data with the results of algorithm output, thus forming information association network. Finally, it realizes effective recognition and efficient semantic understanding of information data.

From the perspective of context effect, the reasoning rules defined by Pierce are clear and easy to understand, especially for graphic logic reasoning. Because CGS is superior to EG in information processing and its concept is more complex, the algorithm rules are more extensive in linguistics. In addition, in order to express personalized sentences, CG is used in some context sensitive special methods, such as consistent operation and personalized symbols. Therefore, the research results of linguistics and computer technology play a very important role in the formation of conceptual graphs.

At the same time, we have tried to use more abundant information to help correct word segmentation, for example, to input some specific nouns, to prove that the design prototype is feasible in the actual retrieval and deep mining of the existing huge portal system files. Compared with the existing general semantic understanding model, it is superior to the latter in speed and feedback results, and at the same time, it also provides a better Strong ability of resource integration, deep processing and document classification can improve the utilization and overall value of data resources. A model of data dynamic management is established, which includes collaborative work, automatic analysis, professional classification and optimization of scoring. It improves the ability and level of information retrieval in professional field.

References

[1] Li, Z., & Zhou, D. D. (2019). Research on conceptual model and construction method of educational knowledge graph. *E-Education Research.*

[2] Anderson, A. J., Binder, J. R, Fernandino, L., Humphries, C. J., Conant, L. L, Raizada, R. D. S, Lin, F., & Lalor, E. C. (2019). An integrated neural decoder of linguistic and experiential meaning. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*.

[3] Liu, Y. B., Wang, X. J., Yuan, C. X., & Yi, L. (2019). Semantic conflicts detection-based machine reading comprehension neural network. *Journal of Beijing University of Posts and Telecommunications*.

[4] Sowa, J. F. (1984). *Conceptual Structures: Information Processing in Mind and Machine*, Addison-Wesley, Reading, MA.

[5] *Peirce, Charles Sanders*. The Writing of C.S.Peirce Volume, Indiana University Press, Bloomington, 1986.

[6] Sowa, J. F. (1992). Conceptual graphs for representing conceptual structures. *Conceptual Structures in Practice*.

[7] *Peirce, Charles Sanders*, Reasoning and the Logic of Things, The Cambridge Conferences Lectures of 1898.

[8] Dau, F. (2003). The logic system of concept graphs with negation and its relationship to predicate logic. *Lecture Notes in Computer Science*.
[9] Sowa, J. F. (2000). *Knowledge Representation: Logical, Philosophical, and Computational Foundations*.
[10] Sanjeev, A., Liang, Y. G., & Ma, T. Y, (2017). *A Simple but Tough to Beat Baseline for Sentence Embeddings*,

Duanming Shen has completed his master's degree in oil and gas information engineering from 2012-2015 in China. Following this his worked as a software development engineer at Computer Application Technology Department from 2015-2019. Currently he is pursuing his career as algorithm engineer at PetroChina Research Institute of Petroleum Exploration & Development, China.

Qing Li has completed her master's degree in information and network security from 2013-2015 in University of Limerick. Following this she worked as a software development engineer at Computer Application Technology Department from 2015-2019. Currently she is pursuing her career as algorithm engineer at PetroChina Research Institute of Petroleum Exploration & Development, China.

Dexin Qiao is a senior lecturer and IT program leader at Computer Application Technology Department, PetroChina Research Institute of Petroleum Exploration & Development, Beijing, China. He has published number of research papers on ensemble learning. His expertise and research interests include software design, ensemble learning and knowledge discovery.

Xingwen Zhou has completed his bachelors in computer science from 2008-2012 in China. He worked as software development engineer at Computer Application Technology Department from 2012-2019. His research interests include software design, ensemble learning and software development.