Research of Chinese Word Knowledge Graph Based on SLPA Algorithm

Xicheng Qian\textsuperscript{1,a}, Yuyang Hu\textsuperscript{1,b} and Jing-Chang Pan\textsuperscript{1,c,*}

\textsuperscript{1}School of Mechanical, Electrical & Information Engineering, Shandong University (Weihai) 264209, China
\textsuperscript{a}xichengqian@gmail.com, \textsuperscript{b}huyuyang_2015@outlook.com, \textsuperscript{c}pjc@sdu.edu.cn

*Corresponding author

Keywords: Chinese words; Knowledge graph; Attribute relationship

Abstract. The different semantics of Chinese words will make a variety of relations between Chinese words. In the known relationship, synonymy, antisense, hypernym, etc. are the most common relationship. Because of the intricate relationship between Chinese words, so that words and other words form a series of knowledge network, which contains a variety of useful information. Based on these knowledge networks, this paper establishes the knowledge graph of Chinese words and analyzes the inherent characteristics of Chinese words. Establish the semantic database of words, and improve the semantic library structure, including the relationship between the number of semantic library words and words (synonymy, hypernym, etc). Based on the semantic database, the semantic network is formed, and community discovery is carried out for the collection of synonyms in the semantic network, and the community properties are extracted to form a complete semantic relation network.

1. Introduction

In 1985, George A. Miller, a professor of psychology at the University of Princeton, established a relatively complete English dictionary, Wordnet, which contains most of the semantic information of English words. It groups them based on the different meaning of each word, in which a number of words with the same meaning of the word group called a synset.

In Wordnet, according to the degree of semantic similarity of these words, a number of synonym collections are formed. Each set of synonyms represents a basic semantic concept.

In addition to the synonyms, the relationship between words is the upper and lower relationship\textsuperscript{[2]}. The upper word refers to a more generalized word in concept. The concept of the upper word generally contains the concept of the lower word. Because of these relationships, synonyms in Wordnet will have another level of relationship between the collection. About 117,000 synonym collections in Wordnet are associated with other synonym collections through these "conceptual relationships\textsuperscript{[3]}", and the association between these collections and collections allows Wordnet to form a large word network.

In order to study the inherent relationship between Chinese words, this paper constructs the knowledge graph of Chinese words base on Wordnet framework. The results can be used not only for search, but also for e-commerce, news and other recommendations.

2. Relevant astronomical background knowledge

According to "Simplified Chinese Words Dictionary" records, Chinese people can use the words about 380,000, of which the most commonly used about 100,000.

There is a series of inherent relationships between Chinese words, such as synonyms, the upper relationship, type, word class, field, side by side.,

CSC Chinese semantic thesaurus, which is a self-developed software system, which contains the daily life of the various terms used, such as synonymy, upper, lower and so on.
In this paper, we use the words "Simplified Chinese Words Dictionary" and the Chinese word network to form the word list, query CSC Chinese semantic thesaurus words in the word to form the lexicon used in this study. The storage of the thesaurus uses the database storage method. The database storage has relatively good stability relative to XML and other file class storage.

3. Community Detection

Community. There are many communities in the network of Chinese words. The connections between the nodes and nodes in the same community are very close, and the connection between the community and the community is sparse or not directly related. In essence, the community is the subnet of the entire network and is a small complex network with vertices and edges. If there is no intersection between the community nodes, known as non-overlapping communities, otherwise known as overlapping communities. As shown in Figure 1.

![Non-overlapping communities and overlapping communities](image)

**Community Detection.** Community Detection is to analyze the composition of the community in a complex network. In the network of Chinese word formation, the community structure is also objective, these community structure expresses the relationship between words and words. The mathematical description of the Community Detection as follows:

Assume graph \( G = (V, E) \), and \( n (n \geq 1) \) communities are determined in graph \( G \)

\[ T = \{T_1, T_2, \ldots, T_n\} \]

so that \( T \) can cover all the \( V \)s.

**Label Propagation.** Label Propagation Algorithm (LPA) is a graph-based semi-supervised learning method proposed by Zhu et al. In 2002. The core idea is to label the unlabeled nodes with labels that have been labeled.

**LPA Algorithm Description.** In the graph \( G = (V, E) \), let \( V_1 = \{v_1, v_2, \ldots, v_t\} \) be the marker data, \( L_t = \{l_1, l_2, \ldots, l_t\} \) \( \in \{1,2,\ldots,C\} \), where \( C \) is known, make \( V_2 = \{v_{t+1}, v_{t+2}, \ldots, v_{t+m}\} \) is not marked data. \( L = \{l_{t+1}, l_{t+2}, \ldots, l_{t+m}\} \) is unknown, \( t \ll m \), and \( V = V_1 + V_2 \).

All nodes (labeled and not labeled) form a complete graph with the weight of the formula as follows:

\[
\omega_{ij} = \exp \left( -\frac{v_i^2}{\sigma^2} \right) = \exp \left( -\frac{\sum_{k=1}^{p}(x_i^k-x_j^k)^2}{\sigma^2} \right)
\]

(1)

Where: \( v_{ij} \) represents the Euclidean distance between any two points.

The propagation probability between a node and a node also has a certain probability, and the probability of propagation between any two points is defined as a probability matrix of \((t + m) \times (t + m)\), as follows:

\[
P_{ij} = P(j \rightarrow i) = \frac{\omega_{ij}}{\sum_{k=1}^{t+m} \omega_{kj}}
\]

(2)

Where: \( P_{ij} \) is the probability that the node \( j \) label propagates to \( i \).

Suppose there are \( C \) classes and \( t \) labeled data, define a \( t \times C \) label matrix \( Y_t \), and the i-th
row represents the label of the i-th data, that is, if the label of the i-th data is j, then the j elements are 1, the other is 0. The remaining m unlabeled data also form a m × C matrix Y_m, which is combined to obtain a matrix Y = [Y_t; Y_m], its size is (t + m) × C. Where Y_m has no data and can be set to a value.

**LPA algorithm process.**
(1) the data is initialized, calculat the weight matrix \( \omega_{ij} \) according to the formula(1);  
(2) Calculate the propagation probability between any two points according to the formula (2) from \( \omega_{ij} \) obtained by (1), and obtain the probability matrix P;  
(3) Define the label matrix \( Y_t \) and the unlabeled matrix \( Y_m \), into the matrix Y;  
(4) To propagate \( Y = PY \);  
(5) The \( Y_t \) in the reset label matrix is the original \( Y_t \);  
(6) Repeat steps (4) and (5) until Y converges.

In the process of algorithm execution, the matrix P and Y are multiplied by (4), and each node propagates the propagation probability of its own label to other nodes. If the similarity between the two nodes is higher, The label of the node is easier to spread by its own label. In step (5), the \( Y_t \) in the Y matrix is reset because the tag of the node in \( Y_t \) is determined and it can not be affected by the label of the other node, so it is reset to the original node information.

**SLPA.** SLPA\(^{[7]}\) (Speaker-listener Label Propagation Algorithm) is an overlapping community detection algorithm.SLPA algorithm introduced Speaker and Listener two concepts, in the process of refresh the node label, select a node for the Listener, and its associated node is its Speaker, usually Speaker more than one, in the transmission, the label of each node is a sequence of labels that records all historical labels in the process of node label propagation. For example, if the iteration is n times, each node will have a label sequence of length n. After the end of the iteration, the frequency of the label in the history label sequence of each node is counted, the label with the smaller frequency is filtered according to the size of the threshold \( \gamma \), and the remaining labels are the label of the node (usually more than one). Through the appropriate value of \( \gamma \), the community can be degraded into non-overlapping communities. The algorithm is shown in Figure 2.

---

**The Result and Analysis of Community Detection Experiment of Synonym Collection in Chinese Words**

In the Chinese word network, there is also the same synonym collection with Wordnet, through the SLPA algorithm to explore the synonym collection, and finally found in the results of the commonly used Chinese word synonym collection of a total of 25,735 synonym collections. The size of the synonym collection shown in Figure 3.
The results show that the number of synonyms with the scale of 2 is the largest, about 13645. This also shows that in the daily use of Chinese words, most of the synonyms are only one, which also conforms to the true relationship of Chinese semantics.

Words also form 164 isolated nodes, which form a collection themselves, and these nodes do not have synonyms in the lexicon formed in this paper, which may be due to the lack of the thesaurus.

In addition to the collection of synonyms of scale 1 and 2, there are 50 other sets of synonyms of other scales, which have multiple synonyms, which are caused by the ambiguity of Chinese words. The meaning of the different, often there will be a number of different synonyms, this ambiguity will lead to the Chinese word network synonyms in the collection of concentrated phenomenon.

In the Chinese word network, there will be overlapping words between the Chinese word synonyms, for example, the index finger means the meaning of the finger, and the two fingers are synonyms, in the name of the time and Guo Lusheng (a poet, also known as the index finger) is synonymous. It can be seen that the network formed by Chinese words will appear overlapping communities in the community detection. Figure 5 shows the distribution of the number of synonyms obtained by taking different values for $\gamma$ when overlapping the word network.
The value of $\gamma$ will have an effect on the collection of synonyms, but the range of influence will not be large. On the whole, the size of the synonym collection is not very different. When $\gamma$ changes, there will be an influence between the collection and the collection rather than the number of collections.

4. Summary

The Chinese word lexicon constructed in this paper can provide a complete Chinese word knowledge base for Chinese researchers. This knowledge base includes most of the semantic information in Chinese words, and because of its extensibility, it can expand the semantic library information in later studies.

The establishment of Chinese word knowledge map enriches the Chinese word network, which makes the Chinese lexicon more widely used, such as the word association function of input method, recommendation of similar information and so on.

Acknowledgements

This work was financially supported by National Natural Science Foundation (U1431102).

References