Journal of Computer and Communications, 2017, 5, 61-68
http://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Automatic Construction Method for Domain Concepts Based on Wikipedia Semantic Knowledge Base

Qiaoyan Zhang, Min Lin, Shujun Zhang

College of Computer and Information Engineering, Inner Mongolia Normal University, Hohhot, China
Email: 476588428@qq.com

How to cite this paper: Zhang, Q.Y., Lin, M. and Zhang, S.J. (2017) Automatic Construction Method for Domain Concepts Based on Wikipedia Semantic Knowledge Base. *Journal of Computer and Communications*, **5**, 61-68.
http://dx.doi.org/10.4236/jcc.2017.51006

Received: December 8, 2016
Accepted: January 15, 2017
Published: January 18, 2017

Abstract

This paper proposes a method to construct conceptual semantic knowledge base of software engineering domain based on Wikipedia. First, it takes the concept of SWEBOK V3 as the standard to extract the interpretation of the concept from the Wikipedia, and extracts the keywords as the concept of semantic; Second, through the conceptual semantic knowledge base, it is formed by the relationship between the hierarchical relationship concept and the other text interpretation concept in the Wikipedia. Finally, the semantic similarity between concepts is calculated by the random walk algorithm for the construction of the conceptual semantic knowledge base. The semantic similarity of knowledge base constructed by this method can reach more than 84%, and the effectiveness of the proposed method is verified.

Keywords

Wikipedia, Semantic Knowledge Base, Keywords Extraction, Semantic Similarity Computation, Random Walk

1. Introduction

In today's era of big data, the semantic content of the constructed artificial knowledge base is static and limited, and it is difficult to meet the needs of large-scale network text information retrieval. Xincai Wang and Jiayou Ding [1] used the visualization tool Gephi to generate the knowledge map, which explains the development, application, characteristics and influence of big data. Represented by the Wikipedia online, encyclopedia knowledge base is described in detail. The Wikipedia knowledge base which provides a favorable basic data resource for the construction of the semantic knowledge base that is faced with retrieving is described in detail, comprehensive and reliable. At the same time, the content is

updated rapidly. With the large-scale open-network courses MOOC (massive open online courses) as the representative of the rise of the network-based teaching model, constructing a domain-oriented semantic knowledge base for easy updating of content MOOC system is important to semantic retrieval of learning resources.

This paper puts forward the concept of the explained text keywords that represent the topic of the concept itself, and constructs the conceptual semantic knowledge base by combining with the semantic relationship among concepts, and constructs the knowledge map of the semantic relationship between the concepts of rules. Finally, the semantic similarity between the concepts is calculated by the random walk algorithm of graphs. The software engineering domain concept knowledge map is constructed. This method can support MOOC system of learning resources according to the semantic retrieval, and can also provide unsupervised topic model to prior knowledge instruction. Although the constructed semantic knowledge base is oriented to the software engineering domain, the method is not limited to the field, and it is also suitable for other fields.

2. Related Work

Nowadays, there is a lot of automatic and semantic knowledge base. For example, Xiaokang Su [2] proposed for the degree of contribution of each concept from his own interpretation of the text in the keyword extraction and calculated the contribution of keywords to the concept and formed the semantic meaning of the concept of semantic fingerprint. But there is no hierarchical relationship between the concepts. Based on the modeling of multi granularity topics in the text stream, a semantic description method based on the semantic hierarchy tree is proposed by Qian Chen *et al.* [3] the way that calculating the SIM divergence of the concept, which of fine granularity is represented. Qi Zhang *et al.* [4] puts forward the theme tree of four layers tree hierarchy structure，there are document, the root of the theme, the theme of the leaf and the concept layer from top-down, from the root to the leaf theme layer, each layer can be used to represent the distribution of a dirichlet. Chengzhi Wu [5] constructs the semantic relationship between the network structures of the concept, but the concept itself does not choose the keywords to describe its semantics. A graph model is used to represent the semantic relations between concepts by Yi Wan [6]. Tao Zhang *et al.* [7] constructed a knowledge map based on the concept of Wikipedia. Qiaoling Liu [8] proposed a semantic search based on Wikipedia and utilized Wikipedia's own characteristics to pluck the concept between the implicit semantic relationships. In essence, the knowledge map is a kind of semantic network, which has higher entity coverage than the traditional ontology and semantic network. By the way, the semantic relationship is more complex and comprehensive.

3. Construction Conceptual Semantic Knowledge Base

We firstly take the concept of SWEBOK V3 as the standard. The concept of software engineering field is selected from SWEBOK V3 and JWPL (Java Wiki-

pedia Library) [9] is used to extract the explanatory text from Wikipedia. At the same time, we explain the concept of the semantic of the concept itself, and construct the concept map according to the established rules.

The construction process is shown in **Figure 1**.

3.1. Concept Semantic Knowledge Base Construction Process

The concept of the relationship between the original level of Wikipedia and the concept of new join each other to explain the relationship between the constructions of the concept of the concept of semantic map in this paper. Specific composition rules are as follows:

1) The concept of the existence of the upper to the lower side of the relationship between Wikipedia and the edge of the weight value is initialized to 0.9;

2) The concept of the existence of the lower to the upper side of the relationship between Wikipedia and the edge of the weight value is initialized to 1;

3) The concept of B appear in the interpretation of the concept of A text keywords and the concept of A in the concept of B value is assigned to the concept of B to the edge of the concept of A;

4) If the concept of A and the concept of B's interpretation of the text have the same keyword C, then the concept of A to the concept of the edge of the B value is assigned to $\left(w_1 + w_2\right)/2$;

Among them, w_1 is used to express the keywords C in the interpretation of the concept A as the keywords in the text keywords; w_2 represents the keywords C in the interpretation of the concept of B as the keywords in the text.

We take into account that the concept of the lower level can be more specific to the upper concept, so the upper to the lower edge of the concept of the weight value is defined as 0.9; the lower to the upper edge of the concept of weight value is defined as 1.

With priority from high to low are 1) → 2) → 3) → 4);

Construct the directed graph $G = (V, E)$, where V represents the concept set, E is the edge set $E \subseteq V * V$.

According to the above four rules to build the concept of knowledge map as shown in **Figure 2**, where CR represents the root concept node; ▭ represents that the concept has a lower concept and ◯ represents leaf node.

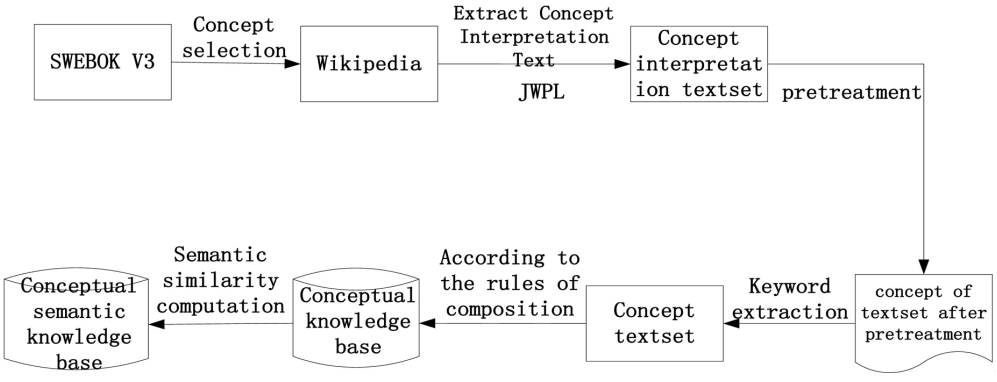

Figure 1. Construction process of semantic knowledge base.

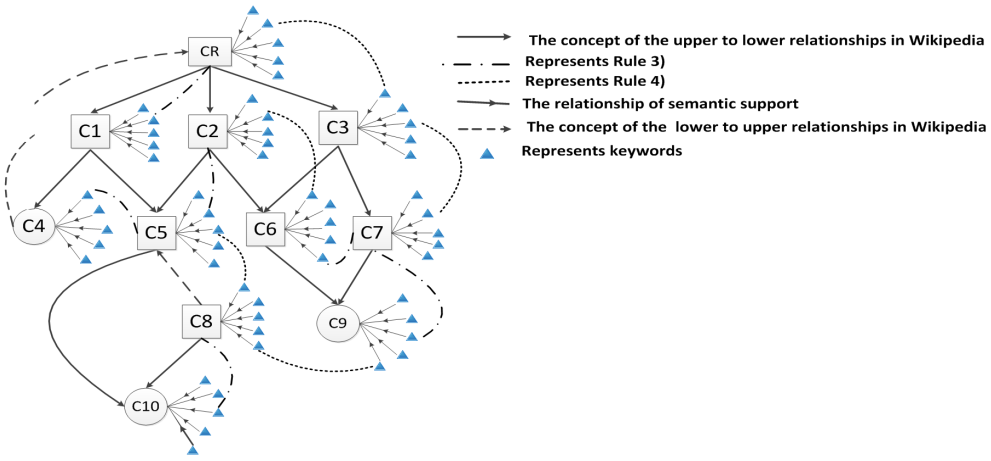

Figure 2. The structure of conceptual semantic knowledge base.

3.2. Semantic Similarity Computation between Concepts

This paper by using random walk of a graph [10] to calculate the semantic similarity between concepts, random walk algorithm is used to capture the similarity between two nodes in a graph. According to the rule, there are some similarities between the direct upper and lower relation and the two nodes with indirect links, as shown in **Figure 2**.

We use A constructed mapping to map the matrix to form a $m*m$ probability transfer matrix $\boldsymbol{P}$, m represents the number of nodes in a graph; Each element of the matrix $\boldsymbol{P}$ is normalized, that is, the sum of all the elements in which each element is divided by the number of rows that are not zero.

Before the concept similarity is calculated by random walk of a graph, it is needed to establish the initial distribution of the concept map. Assuming that the random walk is started from the node i, the initial distribution of the concept map is Formula (1)

$$v_0\left(x_i\right)=\begin{cases}1, & \text{if a random walk start from node } i\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

Given the initial distribution, according to the random walk of a graph algorithm to walk, through multiple iterations can get a stable probability distribution, stable probability distribution can be expressed semantic association strength between the initial node and other nodes.

Random walk of a graph specific algorithm process is as follows:

1) Given initialization matrix $\boldsymbol{v}_0$, and let $v=v_0$;
2) According to the transition probability between concepts, generator matrix $\boldsymbol{P}$;
3) $v_{\text{new}}=\alpha*p^{\mathrm{T}}*v+\left(1-\alpha\right)*v_0$;
4) $v=v_{\text{new}}$;
5) Repeat step 3), 4) until v_{new} reaches a stable state or the number of iterations exceeds a certain threshold.

Starting from the node i random walk, after reaching a stable state, In the graph each node has a probability value, which reflects the importance of the

semantic similarity between the node and the node *i*. In 20 iterations each node a probability value reaches a steady state proved by experiments, being same with the literature [11] α values of 0.5.

Figure 3 is part of the experimental results. If randomly walk from the node 1 of Figure 3 "software engineering", according to the calculation formula of the random walk process 3, the results of the first iteration are as follows:

$$\begin{bmatrix}0.500\\0.000\\0.000\\0.250\\0.000\\0.250\end{bmatrix} = 0.5 * \begin{bmatrix}0.000 & 1.000 & 0.000 & 0.000 & 1.000 & 0.000\\0.000 & 0.000 & 0.757 & 0.797 & 0.000 & 0.000\\0.000 & 0.000 & 0.243 & 0.000 & 0.000 & 0.000\\0.500 & 0.000 & 0.000 & 0.203 & 0.000 & 0.000\\0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.826\\0.500 & 0.000 & 0.000 & 0.000 & 0.000 & 0.174\end{bmatrix} * \begin{bmatrix}1\\0\\0\\0\\0\\0\end{bmatrix} + 0.5 * \begin{bmatrix}1\\0\\0\\0\\0\\0\end{bmatrix}$$

After 20 iterations, the value of each node of the graph reaches a steady state, and the probability distribution satisfies the following relationship:

$$\begin{bmatrix}0.563\\0.062\\0.000\\0.157\\0.064\\0.154\end{bmatrix} = 0.5 * \begin{bmatrix}0.000 & 1.000 & 0.000 & 0.000 & 1.000 & 0.000\\0.000 & 0.000 & 0.757 & 0.797 & 0.000 & 0.000\\0.000 & 0.000 & 0.243 & 0.000 & 0.000 & 0.000\\0.500 & 0.000 & 0.000 & 0.203 & 0.000 & 0.000\\0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.826\\0.500 & 0.000 & 0.000 & 0.000 & 0.000 & 0.174\end{bmatrix} * \begin{bmatrix}0.563\\0.062\\0.000\\0.157\\0.064\\0.154\end{bmatrix} + 0.5 * \begin{bmatrix}1\\0\\0\\0\\0\\0\end{bmatrix}$$

Similarly, if a random walk from Node 3 "formal methods", after 20 iterations to reach the steady state, the probability distribution satisfies the following relations:

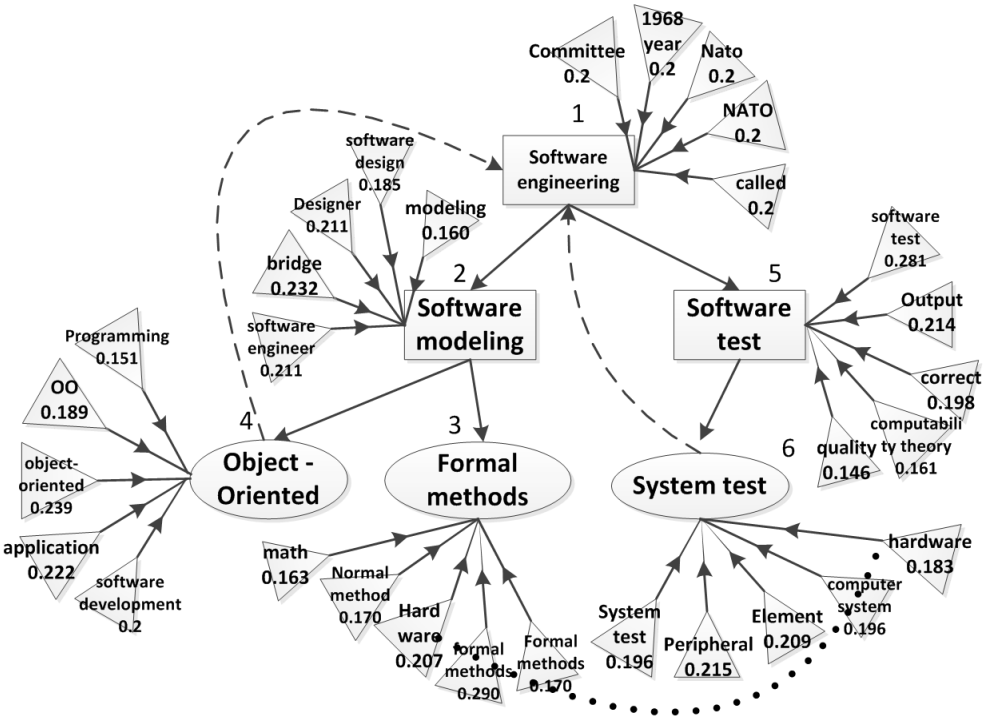

Figure 3. Part of the semantic relationships between concepts.

$$\begin{bmatrix}0.121\\0.229\\0.569\\0.033\\0.014\\0.033\end{bmatrix}=0.5*\begin{bmatrix}0.000&1.000&0.000&0.000&1.000&0.000\\0.000&0.000&0.757&0.797&0.000&0.000\\0.000&0.000&0.243&0.000&0.000&0.000\\0.500&0.000&0.000&0.203&0.000&0.000\\0.000&0.000&0.000&0.000&0.000&0.826\\0.500&0.000&0.000&0.000&0.000&0.174\end{bmatrix}*\begin{bmatrix}0.121\\0.229\\0.569\\0.033\\0.014\\0.033\end{bmatrix}+0.5*\begin{bmatrix}0\\0\\1\\0\\0\\0\end{bmatrix}$$

According to the probability values obtained by the 20 iteration, Random walk from node 1, the "Object-Oriented" to "software engineering" semantic similarity is 0.157. However, the "formal methods" not to the "software engineering" path, so the probability of semantic similarity value of 0, Random walk from the node 3, "software engineering" semantic similarity to "object-oriented" is 0.121. It is concluded that the concept of the lower concept to the concept of the upper level is more than the concept of the upper concept to the lower concept of semantic similarity. The rest of the node has to "formalism" path, and "software modeling" is "formalism" recent upper concept, so in addition to the "formalism" itself "software modeling" to "formalism" semantic similarity is one of the biggest. With the path which is from the rest of the concept to the "formal method" is becoming longer, the semantic similarity probability value is becoming smaller. This result is consistent with the original rule of the concept semantic map.

4. Experimental Results and Analysis

4.1. Experimental Process

Experiments to download the interpretation of the concept of Wikipedia page document content, The relationship between classification and The internal link between the concept page 3 data packets，Using JWPL (Java Wikipedia Library) of Data Machine parse import MySQL database to obtain 11 jwpl_tables.sql table. According to the SWEBOK V3 standard, two hundred core concepts of software engineering have been selected. In addition, the experiment of semantic similarity among concepts is compared with the results obtained by using the symmetric measure sym-KL of literature [3] and the random walk algorithm in the same data set. The comparison of the above two methods with manual standard results are shown in **Table 1**.

The *sym-KL* and *KL* formulas of the literature [3] are as follows: Formula (2) and Formula (3):

$$sym-KL(p\|q)=\frac{1}{2}\left(KL(p\|q)+\left(KL(q\|p)\right)\right) \tag{2}$$

$$KL(p\|q)=\sum_i p(x_i)\log\frac{p(x_i)}{q(x_i)} \tag{3}$$

4.2. Analysis of Experimental Results

The reason why the random walk algorithm is used in this paper is that the random walk algorithm would obtain a stable value after several iterations. In the *Sym-KL* method, when the two random distributions are the same, their relative

Table 1. Random walk, *Sym-KL* and artificial marked contrast to results.

Experimental method	Completely consistent	Basically consistent	Not quite consistent	Completely inconsistent
Random walk	32%	52%	15%	1%
Sym-KL	3%	12%	53%	32%

entropy is zero. When the difference of two random distribution increases, their relative entropy also increases. This is the reason why the sym-KL algorithm is not as accurate as the random walk algorithm. What's more, in order to transfer the relationship among nodes, the graph random walk algorithm accomplishes the random walk according to the connectivity and the transfer probability between nodes. After several iterations, the concept of the semantic relationship between the probability values can be more accurate representation.

5. Conclusions

This paper proposes a method for constructing a teaching-oriented concept of art based on Wikipedia semantic knowledge base, through the figure of random walk algorithm for building the concept of semantic knowledge base of calculating the semantic similarity of concept semantic knowledge base. This method can provide technical support for the semantic retrieval of MOOC system and non-supervised topic model.

At present, it constructs the concept semantic knowledge base in the field of software engineering. Although the random walk algorithm is very effective, there are still many disadvantages: 1) selecting the concept of less, so that the result is not accurate enough; 2) in the composition rules, only the above-mentioned four rules also have certain shortcomings. In the follow-up research work it will gradually solve the above problems, for example, considering the dynamic update of the semantic knowledge base, combined with the word2vec to do in-depth study. Although this paper takes the field of software engineering as an example, the method is not limited to the field, and also can be applied to other fields. Follow-up can also be considered with similar areas of software engineering, such as computer science and other fields.

Acknowledgements

Thanks to Professor Lin Min for his instructions, and thanks to my parents for their support.

References

[1] Wang, X.-C. and Ding, J.Y. (2013) Big Data Knowledgebase Map: Concepts, Characteristics, Application and Effect. *Journal of Intelligence Science*, **9**, 10-14.

[2] Su, X.K. (2010) Research on the Construction of Semantic Knowledge Base and Its Application in the Field of Text Categorization Based on Wikipedia. Central China Normal University, Wuhan.

[3] Chen, Q., Guo, X. and Wang, S. (2015) Topic Structure Modeling of Multi-Granu-

larity Text Flow. *Journal of Chinese Information Processing*, **29**, 118-125.

[4] Zhang, Q., Chen, Q. and Guo, X. (2013) Topic Detection Technology Based on Topic Ontology Tree. *Microelectronics and Computer*, **7**, 60-63.

[5] Wu, C.Z. (2012) Research and Implementation of Knowledge Search System Based on Wikipedia. South China University of Technology, Guangzhou.

[6] Wan, Y. (2014) Research on the Modeling and Application of the Concept Map Based on Wikipedia. Central China Normal University, Wuhan.

[7] Zhang, T., Liu, K. and Zhao, J. (2015) A Graph Model Based Method for Computing the Similarity of Wikipedia Concepts and Its Application in the Entity Linking System. *Journal of Chinese Information Processing*, **29**, 58-67.

[8] Liu, Q.L. (2009) Semantic Search on Wikipedia. Shanghai Jiaotong University, Shanghai.

[9] Fan, Y.J. and Liu, H.L. (2013) Research on Chinese Short Text Classification Based on Wikipedia. Xidian University, Xidian.

[10] Pearson, K. (1905) The Problem of The Random Walk. *Nature*, **72**, 318. https://doi.org/10.1038/072294b0

[11] Hu, J., Wang, G., Lochovsky, F., *et al.* (2009) Understanding User's Query Intent with Wikipedia. *WWW* 2009, 20-24 April 2009, Madrid, 471-480.