Advances in Computer Science Research (ACRS), volume 54
International Conference on Computer Networks and Communication Technology (CNCT2016)

Research of Semantic Similarity Algorithm Based on the Knowledge of Medical Domain

Li-quan HAN[1,*], Zheng-chao XU[1] and Xiao-bo WANG[2]

[1]College of Computer Science and Engineering Changchun University of Technology
Changchun P.R. China

[2]Fuying Hospital of Siping P.R. China

*Corresponding author

Keywords: Semantic similarity, Semantic correlation, SNOMED CT, The knowledge of medical domain.

Abstract. With the development of medical information technology, the research and application of medical big data has become an important direction of data research. The semantic similarity evaluation between medical domain knowledge is an important part of the understanding of medical large data, which can effectively promote the processing, classification and structured processing of medical resources. In medical domain knowledge, the similarity calculation can improve the performance of information retrieval of medical resources and effectively promote the integration of heterogeneous clinical data. Based on the analysis of semantic similarity and semantic correlation algorithm and combined with the characteristics of medical psychology knowledge, the paper introduces the concept of weight value to simulate the characteristics of human psychological quantity, and gives the medical domain knowledge semantic similarity calculation method. Finally adding the semantic structure model with Oxford Centre for Tropical Forests(OCTF) similarity, constitute OCTF similarity calculation model based on the semantic, and formulas are given. By using Systematized Nomenclature of Medicine -- Clinical Terms(SNOMED CT) as the input ontology, the accuracy and usability of the algorithm are verified by the evaluation standard of medical terminology.

Introduction

Between the concept of language similarity and correlation calculation to determine the precise degree of semantic matching between concepts, that is the basis of semantic information retrieval and the fundamental. So the accuracy of semantic similarity between concepts and correlation calculation becomes the key to improve semantic information retrieval[1]. In the medical domain, with the rapid increase of electronic medical data, electronic medical records, medical treatment and scientific research papers have become important data resources for medical clinical research. Similarity computing can improve the performance of information retrieval for medical resources, and can effectively promote the integration of heterogeneous clinical data. Semantic similarity method to analyze the patient's medical records in a semantic way to identify patients with similar cases[2]. Analysis of related references[3-5], some research suggests a general knowledge ontology and semantic similarity between concepts of corpus research, while others are used to study the semantic similarity method of ontology and domain specific corpus. According to the theoretical basis, the different methods are defined based on the analysis of the geometrical structure, the concept of information content and the analysis of semantic features. According to the use of the knowledge resources, taking into account the different knowledge resources, for example, ontology, classification and structure, domain corpus and Thesaurus etc.. Many studies use generic ontology resources, for example, domain independent label corpora and WordNet. However, the coverage of these resources provided by the medical vocabulary is limited.

Based on the analysis of semantic similarity and semantic correlation algorithm, the paper is introduced to simulate the amount of weight is combined with the characteristics of medical psychology knowledge, while the method for calculating the medical knowledge in the field of semantic similarity; Secondly, the semantic structure is added to the OCTF similarity computing model. Finally, the accuracy and usability of the algorithm are verified by using SNOMED CT as the input ontology and the evaluation standard of medical terminology.

Concept Weight and its Calculation Method

Definition of the Concept of Weight

Defination1: Concept Weight: for any one of the concept of A in Concept Relation Tbox T, exists and exists only with the concept of A in the definition of T in the numerical value, this value is called Concept Weight, denoted as Sal (A).

Definition1 indicates that the concept weight is only related to the definition of the concept, and does not need to consider the other attributes of the concept and the relationship between concepts.

Factors1: the more specific semantic concepts of medical knowledge, the greater the stimulus of the concept. For the user, the specific concept can be given the amount of stimulus is greater than the abstract concept of the amount of sensory stimulation. When it comes to the cold, I believe that everyone's mind will show a sneeze, fever, dizziness and other symptoms; but when it comes to the increase of white blood cells, it is not all people can understand.

Factors2: the greater the use of medical concepts, the lower the feeling of the lower limit: Users have their own unique tendency to many concepts in ontology. Due to the user's tendency to weaken the sense of valve limit, in the concept of similarity measurement, the tendency of the concept of a large amount of psychological.

Concept Weight Implication Priority Algorithm

Defination2: for the concept of Tbox T, the weight of the implication relation is conducted to the concept A, which comes from the concept of explicit implication $D_i \in D_r(A)$, implicit implication concept $D'_i \in D'_r(A)$, and the degree of interest $E(A)$. Remember as:

$$Sal_L(A) = Sal_{LP}(A) + (1 - Sal_{LP}(A)) \times E(A) \tag{1}$$

Where: $Sal_{LP} = \max(ST(D_1),...,ST(D_m),ST(D_1'),...,ST(D_m'))$

$$D_1,...,D_m \in D_T(A), D_1^{'},...,D_m^{'} \in D_T^{'}(A)$$

In conceptual relationship Tbox T, the Sal (C) of $\forall\, C \in Cr$ are in the [0, 1] range, the SalL (A) is also in the [0, 1] range. Combined with the breadth first traversal algorithm, it is easy to calculate the weight of the conduction along the implication relation.

Defination3: for the concept of the relationship between Tbox T in the role of the connection to the concept of A weights from the role of the connection between the concept $C \in Cr(A)$ of connection. Remember as: $Sal_R(A)=\max(RT_1(C_1),\ldots,RT_n(C_n))$ Where: $C_1, ..., C_n \in Cr(A)$

Set the concept of the relationship between Tbox T Sal(C) are in the [0,1] range, the SalR (A) is also in the [0,1] range.

Defination4: for the concept of relationship T Tbox, set the concept of Hi implicit concept A form Ri.Ci, define the concept of Ci to the concept of the role of A connection weights of the first k of the transfer function:

$$RT_i^k(C_i) = \begin{cases} r_i \times Sal^{k-1}(C_i) & H_i = \exists R_i.C_i \\ 1 - r_i + r_i \times Sal^{k-1}(C_i) & H_i = \forall R_i.C_i \\ 0 & H_i others \end{cases} \tag{2}$$

Where: ri is the transmission coefficient of $R_i \in R_T$, and $1 \leq r_i \leq 1$, $Sal^{k-1}(C_i)$ for the concept of the concept of the k-1 Ci weight, k>0.

Defination5: for the concept of the relationship Tbox T, the concept of the K calculation of the weight of the A:

$$Sal_k(A) = \mu \times Sal_L(A) + v \times Sal_R^k(A) \tag{3}$$

Where: μ is the entailment relation coefficient, ν is the role of connection coefficients, and μ+ν=1.

$$Sal_R^k(A) = \max(RT_1^k(C_1), \ldots, RF_n^k(C_n)) \;,\; k \geq 1 \qquad Sal_R^0(A) = 0 \tag{4}$$

Defination6: for the concept of the relationship Tbox T, naming the concept of A weights:

$$Sal(A) = \lim_{k \to \infty} Sal^k(A) \tag{5}$$

By definition 6 it is known that the iterative calculation method can be used to determine the weights in a finite number of times.

Medical Knowledge Semantic Similarity Computation Algorithm

Semantic Similarity Computing Model

In the method, the smaller the semantic distance of medical concept knowledge is, the closer it is to the request of the user. When the semantic distance is equal to 0, the current concept knowledge is the data that the user requests; When the semantic distance is greater than a certain value, the medical concept knowledge is not related to the user's query, and it can not be returned to the requesting user as a result set.

All the medical concept knowledge is organized into the semantic concept tree according to the related semantic description and semantic relation. According to the basic principle and operation method of the tree in the discrete mathematics, the distance and the concept weight of the concept in the semantic concept tree can be calculated directly according to the concept of computational medicine. Then the semantic distance converter is converted to semantic similarity, which can complete the measurement and calculation of the semantic similarity of medical concept ontology knowledge. Concept similarity calculation method generally includes 4 types: Based on the name logo, based on corpus statistics, based on thesaurus and based on graph structure.

In the paper, we use the tree structure to store and store the relevant concept knowledge. Ontology of medical knowledge ontology semantic concept tree representation of medical ontology. Tree structure is a kind of special graph structure, so it is important to introduce the concept similarity measurement and calculation method based on graph structure to complete the semantic similarity computation of medical knowledge ontology. Simple calculation procedure and calculation method are given here: Suppose that the two conceptual graph (tree) of G1 and G2, their intersection is G=G1∩G2, as shown in figure 1. The concept graph (tree) similarity S can be defined as two components: the concept similarity Sc and the relational similarity Sr, which are defined as follows:

$$S_c = \frac{2N(G)}{N(G_1)+N(G_2)} \qquad S_r = \frac{2E(G)}{E_G(G_1)+E_G(G_2)} \tag{6}$$

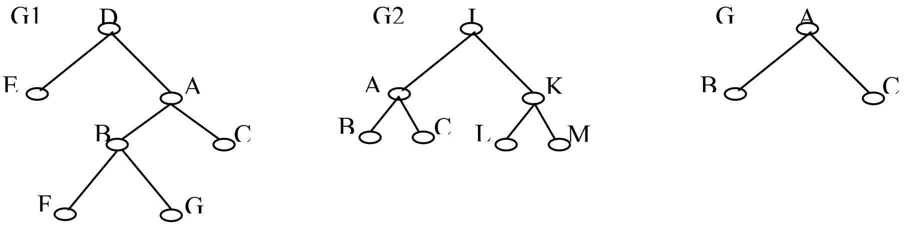

Fig. 1 Concept graph (tree) similarity calculation method

Where: N(G), N(G1), N(G2) respectively of the graph (tree); G, G1 and G2 are the concept of knot points.

Where: E(G) represents the number of edges in a graph EG, EG(G1), EG(G2) respectively the figure G1 and G2 in at least one end and figure G connected to the number of edges.

Through the calculation process and calculation formula, we can know: The concept similarity measurement and calculation method based on graph structure is simple and clear, and it is easy to be operated and implemented.

Improved OCTF Similarity Calculation Model

BACH T-L algorithm focuses on constructing the concept similarity algorithm, which can extract the latent semantic information, and pay attention to the latent semantic. In this paper, the OWL description method is adopted in the process of knowledge ontology acquisition in medical field, In the process of ontology concept similarity calculation, based on the BACH T-L algorithm, the measurement method of the formal concept analysis is added to improve the accuracy of semantic similarity algorithm through the concept feature relation layer. The specific processing flow Figure 2.

(1) the concept structure of the relationship between the characteristics of computing

The concept of two medical concepts (E1, I1) and (E2, I2) is defined as the relationship between the concept of structural similarity:

$$Sim((E_1,I_1),(E_2,I_2)) = [\frac{|E_1 \cap E_2|}{m} \times w + \frac{|I_1 \cap I_2|}{n} \times (1-w)] \times (1+c)^{(l_1+l_2)} \tag{7}$$

M is the larger base of E1 and E2, W is the weights in [0, 1]. N is the base of the set I2. I1 and I2 represent the number of layers in the medical semantic concept tree (E1, I1) and the medical concept node (E2, I2), C is the correction of the volume, according to the concept of depth with respect to the impact of semantic similarity of the impact of modification, this amount is also a result of a number of experiments.

(2) DL OWL language description of the ontology elements to the RDF (RDFs) three sets of the model conversion

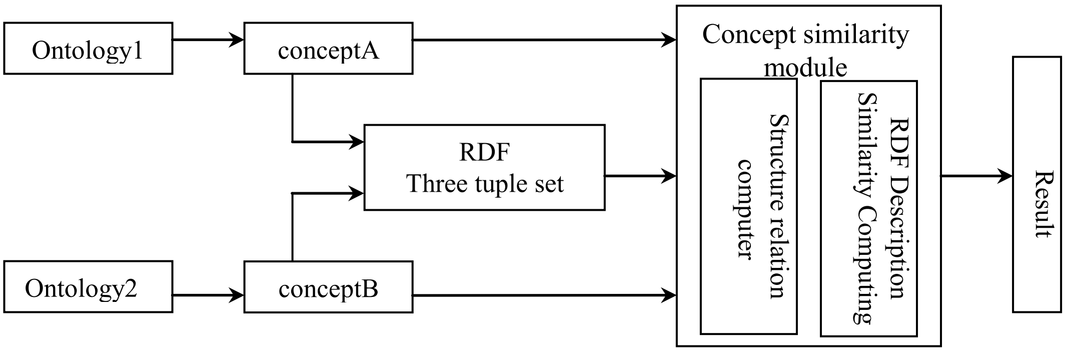
Ontology1
conceptA
RDF
Three tuple set
Ontology2
conceptB
Concept similarity module
Structure relation computer
RDF Description Similarity Computing
Result

Fig. 2 concept similarity calculation process

- O={(S, P, O)} represents a body. (S, P, O) represents a RDF three tuple. S is the subject, P is the predicate, and the O is the object.
- $P(s)=\{p|\exists o,(s, p, o)\in T(s)\}$ The predicate set of the three tuple of the concept S.
- $PC_1=\{p|p\in p(s_1)\wedge(\exists q\in p(s_2), \mathrm{SimPred}(p,q) > 0)\}$ Pc_1 for the medical ontology concept S1 and S2 have the similarity of predicate sets. $PC_2=\{p|p\in p(s_2)\wedge(\exists q\in p(s_1), \mathrm{SimPred}(p,q) > 0)\}$ Pc_2 for the medical ontology concept S1 and S2 have the similarity of predicate sets. $So, Pc = pc_1 \cup pc_2$

(3) Similarity computation based on RDF

By the first (2) step calculation results can be obtained using RDF description of the concept of knowledge of medical ontology structure. Based on the computational method of the semantic distance of the medical ontology knowledge concept, the concept of similarity and attribute similarity of all RDF three tuple containing the same predicate is calculated.

$$Sim_{rdf}(s_1,s_2)=\frac{|p_c|}{\max(|p(s_1)|,|p(s_2)|)}+\frac{|p_x|}{\min(|p(s_1)|,|p(s_2)|)} \tag{8}$$

$$p_x=\{p\mid p\in pc_1\wedge[(\exists q\in pc_2)\wedge(q\neq p),\ \underset{\substack{o1\in o(s_1,p)\\ o2\in o(s_2,q)}}{Sim}(o_1,o_2)>0.6]\} \tag{9}$$

Pc for the concept of S1 and S2 have the same predicate.
To sum up (1) (2) (3) in the whole process we can know: In the above formula, the semantic relations between medical knowledge and RDF are considered; Secondly, the attribute of the concept feature is added to the structure level, and the calculation process of the semantic distance is completed; At last, the concept attribute is added to calculate the similarity of RDF, and further enrich the semantic information of the concept of medical knowledge.

Conclusions and Future Works

SNOMED CT is a set of concept system compositional, Through the mutual combination of other concepts, the concept of the concept to be a special case of. We describe the standard clinical terminology in the field of medicine through the concept, each SNOMED CT concept is a clinical term, and each concept has a unique SNOMED CT concept to identify. UMLS(Unified Medical Language System) provides a free knowledge representation framework, UMLS supports a wide range of medical field research inquiries. It contains more than 100 medical terminology and classification systems with different semantic and syntactic structures.

During the research, other semantic similarity algorithms based on ontology are analyzed, such as semantic similarifiy and relatedness algorithms based on attribute and hybrid semantic similarity and relatedness algorithms. In addition, the algorithms mentioned in this article are based on medical domian ontology or corpus, which can be evaluated and applicated in general domain ontology and corpus.

References

[1] Ming-Yen Chen, Hui-Chuan Chu, Yuh-Min Chen. Developing a semantic-enable information retrieval mechanism[J]. Expert System with Applications.2010, 37(9):322-340

[2] Sachin Mathur, Dinakarpandian. Finding disease similarity based on implicit semantic similarity[J]. Journal of Biomedical Informatics. 2012, 45(9): 363-371.

[3] Liu Hong-zhe, Xu De.Ontology Based Semantic Similarity and Relatedness Measures Review.Computer Science, 2012, 39(2): 8-13.

[4] S.Ravi, M.Rada: Unsupervised graph-based word sense disambiguation using measures of word semantic similarity[C]. In Proceedings of the International Conference on Semantic Computing. IEEE Computer Society Washington, DC, USA, 2007:66-89.

[5] Aseervatham S, Bennani Y. Semi-structured document categorization with a semantic kernel[J]. Pattern Recognition, 2009, 42(9): 2067-2076.

[6] Jiang Yuncheng, Zhang Xiaopei, Tang Yong, et al. Feature-based approaches to semantic similarity assessment of concepts using Wikipedia[J]. Information Processing & Management, 2015, 51(3):215-234.

[7] Mabotuwana T, Lee C M, Cohen-Solal E V. An ontology-based similarity measure for biomedical data-application to radiology reports[J]. Journal of Biomedical Informatics, 2013, 46(5):857-868.

[8] Allones J L, Taboada M, Martinez D, et al. SNOMED CT module-driven clinical archetype management[J]. Journal of Biomedical Informatics, 2013, 46(3):388-400.

[9] Zhang Zhi-Qiang, Peng Qing-Qing, Xie Xiao-Qin, Feng Xiao-Ning. Information Retrieval Techniques Based Query Intention[J]. Journal of Software, 2013, 24(2):162-177

[10] Tian Xuan, Du Xiao-Yong, Li Hai-Hua. Computing Term-Concept Association in Semantic-Based Query Expansion[J]. Journal of Software, 2008,8(19):2043-2053

[11] Liu Hongzhe, Bao Hong, Xu de. Conceptvector for similarity messurement based on hieratchical domain structure[J]. Computing and Informatics, 2011(30):1001-1021.