Multi-dimensional Disease Semantic Similarity Research Postprint

Authors: Zhang Junliang

Date: 2023-04-01T16:15:55+00:00

Abstract

[Purpose/Significance] To address diverse representations of disease knowledge, this paper proposes a comprehensive semantic similarity calculation scheme that integrates multiple dimensions of diseases. [Method/Process] Based on integrating the respective characteristics of disease ontologies and medical encyclopedias, a comprehensive semantic similarity model composed of ontology-based semantic similarity and medical encyclopedia-based disease semantic similarity is designed. Specifically, graph theory is employed to calculate ontology-based semantic similarity, while LDA, set theory, and vector space models are utilized to compute medical encyclopedia-based disease semantic similarity. [Results/Conclusion] By comparing the proposed method with manual judgments by clinicians, the results demonstrate that the method can effectively reflect disease semantic similarity. The proposed method can provide a reference for further research on disease similarity.

Full Text

Research on Multi-dimensional Disease Semantic Similarity

Zhang Junliang1,2,3

1School of Management, Xinxiang Medical University, Xinxiang 453003
2Center for Health Information Resources Research, Xinxiang Medical University, Xinxiang 453003
3Institute of Health Central Plain, Xinxiang 453003

Abstract: *[Purpose/Significance]* Aiming at different expressions of disease knowledge, this paper proposes a comprehensive semantic similarity calculation scheme that integrates multiple dimensions of diseases. *[Method/Process]* Based on the integration of characteristics from disease ontology and medical encyclopedia, a comprehensive semantic similarity model is designed, consisting

of semantic similarity based on disease ontology and disease semantic similarity based on medical encyclopedia. Specifically, graph theory is employed to calculate semantic similarity based on disease ontology, while LDA, set theory, and vector space model are used to compute disease semantic similarity based on medical encyclopedia. *[Result/Conclusion]* Comparing the proposed method with manual judgments by clinicians, the results demonstrate that our approach can effectively reflect disease semantic similarity. This method provides a valuable reference for further research on disease similarity.

Keywords: semantic similarity; disease ontology; disease encyclopedia; similarity measure

---

Semantic similarity is used to reflect the similarity degree between conceptual terms or documents and has long been a hot topic and challenging problem in artificial intelligence, cognitive science, and natural language processing [**?**]. Semantic similarity finds extensive applications in information retrieval, service recommendation, and text clustering analysis [**?**]. Disease semantic similarity plays an active role in studying disease pathogenesis, diagnosis, and drug development, and has been widely applied in research on relationships among biomedical conceptual terms [**?**].

In the process of knowledge management for medical information, researchers have established various knowledge bases containing medical concept relationships from different perspectives, such as the International Classification of Diseases (ICD) developed by the World Health Organization based on disease etiology, pathology, clinical manifestations, and anatomical location, and the Medical Subject Headings (MeSH) established by the U.S. National Library of Medicine. In recent years, researchers have utilized ontologies to describe relationships among medical knowledge and built medical ontology knowledge bases to achieve semantic computability of biomedical terms. The BioPortal [**?**] maintained by the National Center for Biomedical Computing has compiled nearly a thousand medical ontologies. L. Schriml [**?**] constructed a disease ontology to enable formal representation of human diseases. Domestic research on medical knowledge organization has also been actively conducted, with Zhu Ling et al. [**?**] conducting research on traditional Chinese medicine ontology construction based on Chinese medicine literature, and Academician Li Lanjuan's team building a hepatitis ontology [**?**].

With the development of medical science, people have accumulated vast amounts of medical knowledge. Medical encyclopedias represent an important form of knowledge representation. Experts both domestically and internationally have published numerous medical encyclopedias, such as the *Merck Manual* collaboratively written by hundreds of medical experts worldwide, an independent peer-review editorial board, and professional medical authors, and the *Chinese Medical Encyclopedia* completed under government leadership in China. The Internet era has promoted the development of encyclopedias, forming platforms

like Wikipedia and Baidu Baike, which have also given rise to medical Internet encyclopedias such as MedlinePlus established by the U.S. National Library of Medicine and the medical encyclopedia in Baike Mingyi led by China's National Health Commission.

How to leverage different types of medical information resources to calculate disease semantic similarity, improve the comprehensiveness and accuracy of disease semantic similarity computation, and facilitate discovery services for medical information resources will provide support for the development of deeper smart medicine. Based on this, this study utilizes information from disease descriptions in disease ontologies and medical encyclopedias, designs a multi-dimensional disease semantic similarity calculation method, first analyzes commonly used calculation methods in domestic and international semantic similarity research, then studies the calculation method for multi-dimensional disease semantic similarity integrating disease ontology and encyclopedia, and finally analyzes the proposed method using specific examples.

2 Related Research

Semantic similarity has attracted widespread attention from researchers. According to differences in research objects and tasks, semantic similarity can be divided into concept (word) level and text (sentence, paragraph) level [**?**]. Concept semantic similarity measures the semantic relationships between words [**?**]. S. Spagnola et al. [**?**] utilized the shortest path of concepts in semantic networks, incorporating user ratings and other features to represent semantic similarity. R. Cilibrasi et al. [**?**] used the World Wide Web as a database and Google search engine as a foundation to construct a Google semantic similarity calculation method. Some scholars have also used semantic associations in existing semantic knowledge bases to calculate concept semantic similarity. Li Feng [**?**] and Liu Jie [**?**] studied Chinese concept semantic similarity based on HowNet-2000 and HowNet-2008 respectively. T. Nguyen [**?**] and X. Liu [**?**] used WordNet to calculate semantic similarity between words. Zhang Junliang et al. [**?**] used entry annotations in agricultural encyclopedias to calculate semantic similarity. Text semantic similarity calculates the semantic correlation degree between sentences or paragraphs [**?**]. I. Aainul et al. [**?**] used corpora and longest common subsequences to study semantic similarity between sentences or paragraphs. Q. Chen et al. [**?**] used LDA to calculate semantic similarity of short texts. M. Farouk [**?**] utilized word embedding vectors and WordNet to calculate semantic similarity between two sentences. Li Lin et al. [**?**] combined dependency parsing analysis and word embedding vectors to calculate sentence semantic similarity. Zhan Zhijian et al. [**?**] calculated short text semantic similarity based on complex network representation.

According to the implementation algorithms of semantic similarity, it can be divided into statistical methods, graph-based methods, and hybrid technology methods [**?**]. Statistical methods primarily use corpora as a foundation, employing word co-occurrence and contextual information to represent concepts

or texts, combined with mathematical operations to calculate semantic similarity. D. Bollegala et al. [**?**] used page counts and text snippets returned by Web search engines to calculate semantic similarity. Graph-based methods apply graph theory to explain semantic similarity based on existing knowledge bases [**?**]. R. Rada et al. [**?**] used the shortest path between two concepts to measure concept semantic similarity. A. Banu et al. [**?**] considered sub-concepts contained in concepts as influencing factors of semantic concepts. X. Zhu et al. [**?**] introduced local region density of graphs into semantic similarity calculation to improve similarity effects. Li Wenqing et al. [**?**] introduced information theory into concept semantic similarity calculation. Hybrid technology methods comprehensively apply multiple methods for semantic similarity calculation for multi-source information. L. Sahni et al. [**?**] integrated Web search engine similarity measures and word hierarchical structure similarity measures to achieve semantic similarity calculation. Y. Yang et al. [**?**] quantified concept semantic similarity by comprehensively considering semantic distance between concepts, concept hierarchy, and overlap between hypernym/hyponym sets.

Semantic similarity is also important in biomedical research processes, such as gene clustering, gene expression data analysis, and molecular interaction prediction. Semantic similarity in the biomedical field is primarily based on existing ontologies and medical knowledge bases [**?**]. J. Jeong et al. [**?**] and P. Dutta et al. [**?**] used gene ontology to study semantic similarity of genes and gene products. H. Al-Mubaid et al. [**?**] explored the feasibility of measuring semantic similarity between biomedical domain concepts under the UMLS framework using Medline as a standard corpus and grid ontology. Li Wenqing [**?**] proposed a medical semantic similarity algorithm using a method that compares all classification knowledge of concepts.

In summary, existing semantic similarity calculation methods are primarily based on concept hierarchical relationships in ontology knowledge and text similarity, but they have some problems: (1) Few studies integrate both approaches to achieve comprehensive semantic similarity calculation for the same semantic concept. Using different semantic calculation methods for different knowledge resources and effectively integrating them to calculate semantic similarity can comprehensively reflect semantic similarity between concepts; (2) In text similarity calculation, concept description texts are generally analyzed as a whole, with less attention paid to different description expressions for concepts. Using different calculation methods, decomposing text describing concepts according to content categories, and processing them separately based on different text descriptions can more scientifically and reasonably reflect concept semantic similarity.

Since disease semantic similarity research is of great significance, and diseases have multiple knowledge representation forms such as ontologies and medical encyclopedias, while disease concepts in medical encyclopedias consist of multiple parts including overview, symptoms, etiology, diagnosis, and treatment, this paper integrates disease ontology and medical encyclopedia information resources,

processes different descriptions of disease concepts in medical encyclopedia separately, and designs a multi-dimensional disease semantic similarity calculation method based on disease knowledge representation and content description to improve the comprehensiveness and rationality of disease semantic calculation.

3 Disease Comprehensive Semantic Similarity Calculation

Disease concepts have various representation forms. By combining characteristics of different representation forms, selecting appropriate similarity calculation methods, and effectively integrating different semantic similarities, more comprehensive and accurate semantic similarity can be obtained. This paper integrates content from disease ontology and medical encyclopedia disease entries to analyze disease semantic similarity. The specific scheme is shown in Figure 1 [Figure 1: see original paper]. The calculation process for disease comprehensive semantic similarity is as follows:

First, find the corresponding disease concepts for two disease terms in the disease ontology, and calculate the semantic similarity between diseases based on the Disease Ontology (DO), denoted as $S_o(w_1, w_2)$.

Then, locate the entries for the two diseases in the medical encyclopedia, and calculate the semantic similarity between diseases using the similarity of medical encyclopedia content, denoted as $S_d(w_1, w_2)$. This requires calculating definition semantic similarity, symptom semantic similarity, etiology semantic similarity, diagnosis semantic similarity, and treatment semantic similarity.

Finally, combine the DO-based disease semantic similarity and medical encyclopedia-based disease semantic similarity through formula (1):

$$S(w_1, w_2) = \alpha \cdot S_o(w_1, w_2) + \beta \cdot S_d(w_1, w_2) \quad \text{Formula (1)}$$

where $\alpha + \beta = 1$. The proportion between the two similarities is adjusted according to requirements, with the basic principle being to make the comprehensive disease semantic similarity as consistent as possible with manually judged disease semantic similarity.

3.1 DO-Based Disease Semantic Similarity Calculation

DO treats each disease concept as a node and establishes an ontology knowledge base through conceptual semantic associations, linking disease concept terms with knowledge bases such as MeSH, ICD, SNOMED, and OMIM. Figure 2 [Figure 2: see original paper] shows a partial structure of metabolic disease classes in DO:

The semantic similarity calculation method for disease ontology is computed through formula (2):

$$S_o(w_1, w_2) = \frac{depth(NCW(w_1, w_2))}{depth(w_1) + depth(w_2) - depth(NCW(w_1, w_2))} \quad \text{Formula (2)}$$

The value of $S_o(w_1, w_2)$ is in $[0, 1]$, where a larger value indicates greater similarity between the two disease concepts in the disease ontology.

Definition 1: $depth(w)$ represents the depth of concept w to the root node, i.e., the distance from the node to the root node. For example, in Figure 2, the distance from "paramyloidosis" to the root node "metabolic disease" is 3, so $depth(\text{paramyloidosis}) = 3$.

Definition 2: $NCW(w_1, w_2)$ represents the nearest common ancestor concept between concept w_1 and concept w_2. In Figure 2 [Figure 2: see original paper], the nearest common ancestor node for "hereditary fructose intolerance syndrome" (w_1) and "glycerol kinase deficiency" (w_2) is "hereditary metabolic disease", so $NCW(w_1, w_2) =$ hereditary metabolic disease.

For example, in Figure 2, the semantic similarity between "hereditary fructose intolerance syndrome" (w_1) and "glycerol kinase deficiency" (w_2) is calculated as $\frac{2+3-1}{2+3-1} = 0.25$.

3.2 Medical Encyclopedia-Based Disease Semantic Similarity Calculation

In medical encyclopedias, disease entries provide relatively complete explanations of disease knowledge from overview, symptoms (including clinical manifestations, etc.), etiology, diagnosis (including examinations, etc.), and treatment (including prevention, etc.). Through analysis of entries, the overview section condenses the essential basic knowledge of diseases; the symptoms section describes disease manifestations; the etiology section details disease causes; the diagnosis section describes the diagnostic process; and the treatment section elaborates on treatment plans.

Since disease entries differ in text length for overview, symptoms, etiology, diagnosis, and treatment sections, and there are differences in semantic density of descriptive language and terms across sections—for instance, overview content is relatively short with concentrated semantic density, symptoms contain relatively more medical terminology, and etiology, diagnosis, and treatment sections are relatively longer—this paper designs different semantic similarity calculation methods for the descriptive characteristics of each section. The overview section has relatively few words with high correlation between them, and LDA can identify latent topic information in texts. Therefore, an LDA-based similarity calculation method is designed for the disease overview section. Symptom terms mostly describe clinical manifestations and patients' abnormal feelings or objective pathological changes, which can be understood as a set of terms. Thus, a set-based similarity calculation method is designed for the symptom section,

referencing J. Zhang's [**?**] concept tree structure similarity. The etiology, diagnosis, and treatment sections are similar to general text content, but there are differences in term frequency and other relevant features across sections. Therefore, vector space-based similarity calculation methods are designed for etiology, diagnosis, and treatment sections separately.

3.2.1 LDA-Based Similarity Calculation In 2003, D. Blei et al. proposed the Latent Dirichlet Allocation (LDA) model [**?**] based on word co-occurrence and the "word-document-topic" relationship. As an unsupervised machine learning method, LDA has been widely applied in text information analysis [**?**]. This paper treats the overview section of disease entries in medical encyclopedias as disease definitions, uses the LDA model to obtain the topic distribution of each disease, and employs relative entropy [**?**] for similarity calculation since topic models exist in probability form. The steps for disease definition similarity calculation are:

(1) Segment words in the overview section of encyclopedia disease entries and extract medical terms and nouns to form a dataset;
(2) Use the LDA model algorithm to analyze the training dataset and obtain the "topic-word" LDA model;
(3) Calculate the topic distribution T_w of disease w using the LDA model, and compute disease similarity using Definition 3.

Definition 3: T_w is the topic distribution of disease definition, $(t_1, t_2, ..., t_n)$. The definition similarity between disease w_1 and disease w_2 is:

$$S_{de}(w_1, w_2) = 1 + \sum_{i=1}^{n} t_{1i} \ln \frac{t_{1i}}{t_{2i}} + \sum_{i=1}^{n} t_{2i} \ln \frac{t_{2i}}{t_{1i}} \quad \text{Formula (3)}$$

where n is the number of topics in the LDA model, and t_{1i} and t_{2i} are the probabilities of the i-th topic for w_1 and w_2 respectively. The value of $S_{de}(w_1, w_2)$ is in $[0, 1]$, where a larger value indicates greater similarity between the two disease concepts.

3.2.2 Set-Based Similarity Calculation Disease symptom similarity is represented by calculating symptoms described for different diseases, specifically through Definition 4.

Definition 4: Disease symptoms are represented by $set(w)$, i.e., the set of terms describing disease symptoms. The symptom similarity between disease w_1 and disease w_2 is:

$$S_{sy}(w_1, w_2) = \frac{|set(w_1) \cap set(w_2)|}{|set(w_1) \cup set(w_2)|} \quad \text{Formula (4)}$$

where ∩ represents set intersection and ∪ represents set union. The value of $S_{sy}(w_1, w_2)$ is in $[0, 1]$, where a larger value indicates more shared symptoms and greater similarity between the two disease concepts.

3.2.3 Vector Space-Based Similarity Calculation The vector space model vectorizes text content, enabling vectorized processing and representing semantic similarity through spatial similarity, which has been widely applied in text information processing. The content on etiology, diagnosis, and treatment in disease encyclopedias is relatively rich, so this paper adopts vector space-based similarity calculation methods. The specific implementation steps are: first, vectorize the text of etiology, diagnosis, and treatment sections; then calculate similarity using Definitions 5, 6, and 7 respectively; finally, compute disease similarity based on medical encyclopedia through Definition 8.

Definition 5: The word vector for disease etiology w is defined as w_s. The etiology similarity between disease w_1 and disease w_2 is:

$$S_{et}(w_1, w_2) = \frac{w_{s1} \cdot w_{s2}}{\|w_{s1}\| \cdot \|w_{s2}\|} \quad \text{Formula (5)}$$

where w_{s1} and w_{s2} are text vectors for the etiology sections of diseases w_1 and w_2, $\cdot$ represents vector dot product, and $\|\|$ represents vector norm operation.

Definition 6: The word vector for disease diagnosis w is defined as w_d. The diagnosis similarity between disease w_1 and disease w_2 is:

$$S_{di}(w_1, w_2) = \frac{w_{d1} \cdot w_{d2}}{\|w_{d1}\| \cdot \|w_{d2}\|} \quad \text{Formula (6)}$$

where w_{d1} and w_{d2} are text vectors for the diagnosis sections of diseases w_1 and w_2.

Definition 7: The word vector for disease treatment w is defined as w_t. The treatment similarity between disease w_1 and disease w_2 is:

$$S_{tr}(w_1, w_2) = \frac{w_{t1} \cdot w_{t2}}{\|w_{t1}\| \cdot \|w_{t2}\|} \quad \text{Formula (7)}$$

where w_{t1} and w_{t2} are text vectors for the treatment sections of diseases w_1 and w_2.

The values of $S_{et}(w_1, w_2)$, $S_{di}(w_1, w_2)$, and $S_{tr}(w_1, w_2)$ are in $[0, 1]$, where larger values indicate greater similarity between the two diseases in etiology, diagnosis, and treatment aspects.

Definition 8: The semantic similarity based on disease description is:

$$S_d(w_1, w_2) = \gamma_1 \cdot S_{de}(w_1, w_2) + \gamma_2 \cdot S_{sy}(w_1, w_2) + \gamma_3 \cdot S_{et}(w_1, w_2) + \gamma_4 \cdot S_{di}(w_1, w_2) + \gamma_5 \cdot S_{tr}(w_1, w_2) \quad \text{Formula (8)}$$

where $\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5$ are weights for each semantic similarity, with $\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 + \gamma_5 = 1$. These weights are adjusted and set through expert evaluation and experiments, with the basic principle being to set them according to the degree to which content reflects disease semantics. The overview section provides a relatively comprehensive summary of diseases and is assigned a higher weight; etiology, symptoms, diagnosis, and treatment describe different aspects of disease semantics, and their impact on disease semantic similarity is considered equal.

4 Experiments

To verify the effectiveness of the proposed disease semantic similarity method, this paper calculated similarity for 20 pairs of diseases using our method and simultaneously organized clinicians to conduct similarity judgments on the same 20 pairs of diseases, comparing the correlation between our method and manual judgments.

4.1 Experimental Environment

The dataset primarily includes disease ontology and disease encyclopedia. Disease ontology data was sourced from Disease Ontology [**?**] from the Northwestern University Center for Genetic Medicine and the Institute for Genome Sciences at the University of Maryland School of Medicine. Disease concept descriptions utilized the disease encyclopedia from Baike Mingyi [**?**] constructed by the National Health Commission, collecting 7,808 disease concepts. This paper used medical terms from Baike Mingyi's drug and test encyclopedias, along with collected medical symptom terms, to form a medical dictionary used for segmenting disease concept descriptions.

Clinicians are frontline practitioners in clinical treatment and medical research who best understand semantic associations between diseases. Therefore, this study organized clinicians from the First Affiliated Hospital of Xinxiang Medical University (a tertiary Grade A hospital) to judge disease similarity. Five clinicians (including 1 chief physician, 2 associate chief physicians, and 2 attending physicians) from the endocrinology and neurology departments independently evaluated the similarity of 20 disease pairs on a scale of 0 to 9, where 0 indicates completely dissimilar and 9 indicates completely similar. The manual similarity was calculated using formula (9):

$$sp = \frac{1}{5} \sum_{i=1}^{5} sp_i \quad \text{Formula (9)}$$

where sp_i is the similarity score given by the i-th expert for a disease pair. The value of sp is in $[0, 1]$, where 0 indicates clinicians consider the two disease concepts completely dissimilar and 1 indicates completely similar. To verify consistency among the five clinicians' disease semantic judgments, Cronbach's Alpha [**?**] was used for consistency testing, yielding a result of 0.977, indicating high consistency among doctors' disease semantic similarity judgments.

The programming environment used in the experiments was Python 3.6 64-bit system [**?**], with natural language processing tools HanLP [**?**], mathematical computation library NumPy [**?**], and topic analysis using gensim's LDA and TF-IDF [**?**].

4.2 Evaluation Methods

To evaluate the effectiveness of the proposed algorithm, Spearman correlation coefficient and Pearson correlation coefficient were employed. For two random variables X (disease comprehensive semantic similarity) and Y (manually judged disease semantic similarity) with the same number of elements, X_i and Y_i are the i-th elements in X and Y respectively. Elements in X and Y are sorted in ascending or descending order, with x_i and y_i representing the ranking positions of X_i and Y_i. The difference between corresponding elements in sets X and Y is calculated as $d_i = x_i - y_i$. The Spearman correlation coefficient [**?**] is calculated as:

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \quad \text{Formula (10)}$$

Its value ranges between $[-1, 1]$, with larger values indicating stronger correlation. This paper uses the Spearman correlation coefficient to reflect the correlation between the proposed disease comprehensive semantic similarity calculation method and clinicians' judgments. A Spearman correlation coefficient closer to 1 indicates that the designed disease semantic similarity is more consistent with doctors' cognitive judgments.

The Pearson correlation coefficient [**?**] is calculated as:

$$r = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}} \quad \text{Formula (11)}$$

Its value also ranges between $[-1, 1]$. The larger the absolute value of the Pearson correlation coefficient, the stronger the correlation. A coefficient closer to 1 or -1 indicates stronger correlation, while a coefficient closer to 0 indicates weaker correlation. Like the Spearman correlation coefficient, a Pearson correlation coefficient closer to 1 indicates that the designed disease semantic similarity is more consistent with doctors' cognitive judgments.

4.3 Experimental Process and Results

4.3.1 Experimental Process The experiments include two parts: calculating semantic similarity based on disease ontology and calculating semantic similarity based on disease encyclopedia. The ontology-based semantic similarity is calculated using the path from disease to root node. The encyclopedia-based semantic similarity requires first calculating the LDA model for overview and TF-IDF models for etiology, diagnosis, and treatment. The model implementation process is shown in Figure 3 [Figure 3: see original paper].

The main preprocessing work in the experiments involves collecting disease entry-related content from the Internet, comprehensively using the content structure to segment disease entries according to overview, etiology, symptoms, diagnosis, and treatment sections, and cleaning the content to provide raw materials for constructing datasets for each section in the next step.

The word segmentation and feature selection process in the experiments first adds medical-related terms to the medical dictionary and tags their part-of-speech as "nh", while also adding a Chinese stopword dictionary. Then, HanLP's segmentation tool is used to segment different sections of diseases separately. Finally, feature words are selected based on part-of-speech, with nouns selected for overview, etiology, diagnosis, and treatment sections, and medical-related words with part-of-speech "nh" primarily selected for the symptoms section.

The TF-IDF model implementation process in the experiments uses feature words from disease etiology, diagnosis, and treatment sections to build respective corpora, then uses gensim's TfidfModel module to construct respective TF-IDF models.

The LDA model implementation process in the experiments uses feature words from disease overview sections to build a corpus, then uses gensim's LdaModel to construct the LDA model. In establishing the LDA model for disease overview, determining the number of topics is crucial for model application. This paper uses perplexity [**?**] to determine the number of topics in experiments, with perplexity values for different topic numbers shown in Figure 4 [Figure 4: see original paper].

Figure 4 shows that when the number of topics is set to 60 and the number of iterations is set to 1000, the LDA model's perplexity is minimized. Therefore, this paper selects 60 topics and 1000 iterations for the LDA model.

The ontology-based semantic similarity process in the experiments first obtains the nearest common ancestor node of two concepts, then retrieves the distances from the three nodes to the root node, and finally calculates similarity using formula (2). For encyclopedia-based disease semantic similarity calculation, the LDA-based similarity process first imports the LdaModel, then uses disease overview segmentation and feature extraction feature words with the LdaModel to calculate disease topic distribution, and finally uses gensim's similarities to calculate similarity. The set-based similarity calculation process computes sim-

ilarity between feature words from symptom sections of two diseases using formula (4). The vector space-based similarity calculation process first imports corresponding TF-IDF models, then calculates corresponding sections for diseases, and finally uses gensim's similarities to calculate similarity.

4.3.2 Results Analysis In calculating comprehensive semantic similarity of diseases, the weights α and β between DO-based disease semantic similarity and encyclopedia-based disease semantic similarity affect the comprehensive semantic similarity calculation. The experiments designed three different comprehensive semantic similarities: Comprehensive 1, Comprehensive 2, and Comprehensive 3. For Comprehensive 1: $\alpha = 0.5$, $\beta = 0.5$; Comprehensive 2: $\alpha = 0.6$, $\beta = 0.4$; Comprehensive 3: $\alpha = 0.4$, $\beta = 0.6$.

In encyclopedia-based disease semantic similarity calculation, the weights $\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5$ for semantic similarities of overview, etiology, symptoms, diagnosis, and treatment are involved. Since the overview section provides a relatively comprehensive expression of diseases, it is assigned a higher weight. Etiology, symptoms, diagnosis, and treatment describe different aspects of disease semantics, and their impact on disease semantic similarity is considered equal. Therefore, the four weights are the same. In the experiments, encyclopedia-based disease similarity is defined as "Comprehensive 4", with weights set as: $\gamma_1 = 0.4$, $\gamma_2 = \gamma_3 = \gamma_4 = \gamma_5 = 0.15$.

Using the proposed method, semantic similarity experiments were conducted on 20 pairs of diseases. The experimental results for ontology similarity, definition similarity, etiology similarity, symptom similarity, diagnosis similarity, treatment similarity, encyclopedia semantic similarity (Comprehensive 4), and comprehensive semantic similarity (Comprehensive 1, 2, 3) are shown in Table 1 .

In Table 1, Type 2 diabetes and diabetes have the highest semantic similarity. Since Type 2 diabetes is a type of diabetes, their semantic similarity is highest. Type 2 diabetes and diabetic retinopathy belong to metabolic diseases and anatomical diseases respectively in the ontology, with ontology semantic similarity of 0. However, diabetic retinopathy is actually caused by diabetes, so there is similarity between them. Similarly, Ebola hemorrhagic fever, viral meningitis, and viral pneumonia have high similarity in definition description and etiology, but ontology similarity is 0. Table 1 demonstrates that the multi-dimensional semantic similarity calculation in this paper can reflect disease semantic similarity.

To analyze the correlation between various similarity measures (ontology similarity, definition similarity, etiology similarity, symptom similarity, diagnosis similarity, treatment similarity, encyclopedia semantic similarity (Comprehensive 4), and comprehensive semantic similarity (Comprehensive 1, 2, 3)) and manual similarity, Spearman correlation coefficient formula (10) and Pearson correlation coefficient formula (11) were used. The results are shown in Figure

5 [Figure 5: see original paper].

In Figure 5, overall, disease comprehensive semantic similarity (Comprehensive 1, 2, 3), ontology similarity, definition similarity, etiology similarity, symptom similarity, diagnosis similarity, treatment similarity, and encyclopedia semantic similarity (Comprehensive 4) all have high correlation with manually judged disease similarity. From the Spearman correlation coefficient perspective, Comprehensive 2 has the highest correlation with manual judgment (Spearman correlation coefficient of 0.935), followed by Comprehensive 1 (0.923), then encyclopedia semantic similarity (Comprehensive 4, 0.900), Comprehensive 3 (0.874), with symptom similarity being the lowest (0.221). From the Pearson correlation coefficient perspective, the results are generally consistent with Spearman correlation coefficients: Comprehensive 2 has the highest correlation (0.948), followed by Comprehensive 1 (0.941), then Comprehensive 3 (0.925), with symptom similarity being the lowest (0.471).

Experiments comparing with manually judged disease semantic similarity show that: (1) Disease semantic similarity calculation combining disease ontology and encyclopedia outperforms using either disease ontology or encyclopedia alone; (2) Adjusting weights between ontology-based semantic similarity and encyclopedia-based semantic similarity can improve comprehensive semantic similarity results; (3) Encyclopedia-based comprehensive semantic similarity significantly outperforms calculating definition, etiology, symptoms, diagnosis, and treatment similarities separately; (4) The correlation coefficients of Comprehensive 1, 2, and 3 with manual judgment indicate that the weight of DO-based disease semantic similarity is relatively higher than that of encyclopedia-based disease semantic similarity. Overall, the proposed multi-dimensional semantic similarity calculation method meets the requirements of manual disease semantic similarity judgment and demonstrates good effectiveness.

5 Conclusion

This paper designs a multi-dimensional disease semantic similarity calculation scheme using disease ontology and encyclopedia. For the disease overview section in medical encyclopedias, an LDA-based semantic similarity calculation method is designed; for the disease symptoms section, a set-based semantic similarity calculation method is designed; for etiology, diagnosis, and treatment sections, vector space-based semantic similarity calculation methods are designed. All similarities are fused together based on expert evaluation and experiments. The weights between DO-based disease semantic similarity and encyclopedia-based disease semantic similarity are explored using three different value sets in experiments, showing that the weight of DO-based disease semantic similarity is relatively higher than that of encyclopedia-based disease semantic similarity. The proposed method includes both disease concept relationship similarity and disease semantic description, enabling multi-dimensional measurement of disease semantic similarity, which demonstrates better effectiveness compared to single-dimensional semantic similarity. Future work will continue to study

incorporating medical knowledge into encyclopedia-based semantic similarity to further optimize and improve disease semantic similarity calculation, and apply the proposed text similarity method to other research fields.

References

[1] ILAKIVA P, SUMATHI M, KARTHIK S. A survey on semantic similarity between words in semantic Web[C]//International conference on radar, communication and computing. Tiruvannamalai: IEEE, 2012: 213-216.

[2] Sha Yongzhong, Shi Zhongxian. Public crisis event case retrieval method based on semantic similarity[J]. Information and Documentation Services, 2014(6): 78-81.

[3] LIU L, YU Z. An improved knowledge push method based on semantic similarities[C]//Fourth international conference on multi-media networking and security. Nanjing: IEEE, 2012: 378-380.

[4] Wang Daoping, Zhao Yao, Liu Tao. Research on semantic similarity for knowledge service retrieval in agile supply chain[J]. Library and Information Service, 2010, 54(16): 78-81.

[5] KULMANOV M, HOEHNDORF R. Evaluating the effect of annotation size on measures of semantic similarity[J]. Journal of biomedical semantics, 2017, 8(1): 7.

[6] Li Jie, Chu Yanshuo, Cheng Liang, et al. Disease similarity calculation method based on disease ontology[J]. Progress in Biochemistry and Biophysics, 2015, 42(2): 115-122.

[7] NCBO BioPortal[EB/OL]. [2019-08-08]. https://bioportal.bioontology.org/.

[8] SCHRIML L, ARZE C, NADENDLA S, et al. Disease ontology: a backbone for disease semantic integration[J]. Nucleic acids research, 2012, 40(D1): D940-D946.

[9] Zhu Ling, Yang Feng, HE Y, et al. Analysis of important concepts in basic formal ontology and implications for Chinese medical domain ontology construction[J]. China Digital Medicine, 2018, 13(2): 27-30, 56.

[10] Chen Yunzhi. Research on hepatitis ontology construction and semantic similarity[D]. Hangzhou: Zhejiang University, 2017.

[11] JORGE M. An overview of textual semantic similarity measures[J]. Journal of library and information science, 2012, 5(4): 21-36.

[12] Qin Chunxiu, Zhao Pengwei, Liu Huailiang. Research on word similarity calculation[J]. Information Theory and Practice, 2007, 30(1): 105-108.

[13] SPAGNOLA S, LAGOZE C. Edge dependent pathway scoring for calculating semantic similarity in concept net[C]//Proceedings of the ninth interna-

tional conference on computational semantics. Tilburg: Association for Computational Linguistics, 2011: 385-389.

[14] CILIBRASI R, VITANYI M. The google similarity distance[J]. Artificial intelligence review, 2012, 42(4): 935-943.

[15] Li Feng, Li Fang. Chinese word semantic similarity calculation based on HowNet-2000[J]. Journal of Chinese Information Processing, 2007(3): 99-105.

[16] Liu Jie, Guo Yu, Tang Shiping, et al. Word similarity calculation based on HowNet-2008[J]. Small Microcomputer Systems, 2015, 36(8): 1728-1733.

[17] NGUYEN T, CONRAD S. A semantic similarity measure between nouns based on the structure of wordnet[C]//Proceedings of international conference on information integration and Web-based applications & services. Vienna: ACM, 2013: 605-619.

[18] LIU X, ZHOU Y, ZHENG R. Measuring semantic similarity in wordnet[C]//International conference on machine learning and cybernetics. Hong Kong: IEEE, 2007: 3431-3435.

[19] Zhang Junliang, Zhu Xuefang. Research on agricultural concept cluster representation based on Agricultural Dictionary[J]. Information Science, 2013, 31(7): 15-17, 22.

[20] Chen Erjing, Jiang Enbo. Survey on text similarity calculation methods[J]. Data Analysis and Knowledge Discovery, 2017, 1(6): 1-11.

[21] AMINUL I, DIANA I. Semantic text similarity using corpus-based word similarity and string similarity[J/OL]. ACM Transactions on knowledge discovery from data, 2008, 2(2): 10. [2019-08-08]. http://www.researchgate.net/publication/220345072.

[22] CHEN Q, YAO L, YANG J. Short text classification based on LDA topic model[C]//International conference on audio, language and image processing. Shanghai: IEEE, 2016: 749-753.

[23] FAROUK M. Sentence semantic similarity based on word embedding and WordNet[C]//13th international conference on computer engineering and systems. Cairo: IEEE, 2018: 33-37.

[24] Li Lin, Li Hui. A text similarity calculation method based on concept vector space[J]. Data Analysis and Knowledge Discovery, 2018, 2(5): 48-58.

[25] Zhan Zhijian, Yang Xiaoping. A short text semantic similarity calculation method based on complex networks[J]. Journal of Chinese Information Processing, 2016, 30(4): 71-80, 89.

[26] Li Hui. Survey on word similarity algorithms[J]. Modern Intelligence, 2015, 35(4): 172-177.

[27] BOLLEGALA D, ISHIZUKA M, MATSUO Y. Measuring semantic similarity between words using web search engines[C]//International conference on World Wide Web. Banff: ACM, 2007: 757-766.

[28] ZHU G, IGLESIAS C. Computing semantic similarity of concepts in knowledge graphs[J]. IEEE transactions on knowledge and data engineering, 2017, 29(1): 72-85.

[29] RADA R, MILI H, BICHNELL E, et al. Development and application of a metric on semantic nets[J]. IEEE transaction on systems, man, and cybernetics, 1989, 19(1): 17-30.

[30] BANU A, FATIMA S S, KHAN K U. A new ontology-based semantic similarity measure for concepts subsumed by multiple super concepts[J]. International journal of Web applications, 2014, 6(1): 14-22.

[31] ZHU X, LI F, CHEN H, et al. An efficient path computing model for measuring semantic similarity using edge and density[J]. IEEE transactions on knowledge and data engineering, 2007, 19(3): 370-383.

[32] Li Wenqing, Sun Xin, Zhang Changyou, et al. An ontology concept semantic similarity calculation method[J]. Acta Automatica Sinica, 2012, 38(2): 229-235.

[33] SAHNI L, SEHGAL A, KOCHAR A, et al. A novel approach to find semantic similarity measure between words[C]//2nd international symposium on computational and business intelligence. New Delhi: IEEE, 2014: 89-92.

[34] YANG Y, PING Y. An Ontology-based semantic similarity computation model[C]//IEEE international conference on big data and smart computing. Shanghai: IEEE, 2018: 561-564.

[35] PESQUITA C, FARIA D, FALCAO A O, et al. Semantic similarity in biomedical ontologies[J]. PLoS computational biology, 2009, 5(7): e1000443.

[36] DUTTA P, BASU S, KUNDU M. A new hybrid semantic similarity measure using information content and topological features of the Gene Ontology graph[C]//International conference on computer communication and informatics. Coimbatore: IEEE, 2017: 1-5.

[37] JEONG J, CHEN X. A new semantic functional similarity over gene ontology[J]. IEEE/ACM transactions on computational biology and bioinformatics, 2015, 12(2): 322-334.

[38] DUTTA P, BASU S, KUNDU M. Assessment of semantic similarity between proteins using information content and topological properties of the gene ontology graph[J]. IEEE/ACM transactions on computational biology & bioinformatics, 2018, 15(3): 839-849.

[39] AL-MUBAID H, NGUYEN H. Using MEDLINE as standard corpus for measuring semantic similarity in the biomedical domain[C]//Sixth IEEE international symposium on bioInformatics and bioEngineering. Arlington: IEEE, 2006: 315-318.

[40] Li Wenqing. Research on medical domain ontology-based semantic similarity algorithm[D]. Taiyuan: Taiyuan University of Technology, 2013.

[41] ZHANG J, ZHU X, ZHU G. Designing an automated FAQ answering system for farmers based on hybrid strategies[J]. Chinese journal of library and information science, 2012, 5(4): 21-36.

[42] BLEI D, NG A, JORDAN M I, et al. Latent dirichlet allocation[J]. Journal of machine learning research, 2003, 3(3): 993-1022.

[43] He Weilin, Xie Hongling, Feng Guohe. Survey on latent Dirichlet allocation model[J]. Journal of Information Resources Management, 2018, 8(1): 55-64.

[44] Liu Ming, Wang Xiaolong, Liu Yuanchao. Semantic-based high-dimensional data clustering technology[J]. Acta Electronica Sinica, 2009, 37(5): 925-929.

[45] Disease ontology[EB/OL]. [2019-08-08]. http://www.disease-ontology.org/.

[46] Baike Mingyi[EB/OL]. [2019-08-08]. http://www.bakemy.com/.

[47] Python[EB/OL]. [2019-08-08]. http://www.python.org/.

[48] HanLP[EB/OL]. [2019-08-08]. http://hanlp.linrunsoft.com/.

[49] NumPy[EB/OL]. [2019-08-08]. http://www.numpy.org/.

[50] gensim: Topic modeling for humans[EB/OL]. [2019-08-08]. http://radimrehurek.com/gensim/.

[51] Zhou Aiming. Practical multivariate statistics in library and information science[M]. Zhengzhou: Zhengzhou University Press, 2017.

[52] Guan Peng, Wang Yuefen. Research on determining optimal topic number for LDA topic model in scientific and technological intelligence analysis[J]. New Technology of Library and Information Service, 2016(9): 42-50.

Note: Figure translations are in progress. See original paper for figures.