Disease Association Detection Based on Word2Vec and Public Health Information Sources: Post-print

Authors: Luo Wenxin, Chen Chong, Deng Siyi

Date: 2017-11-08T00:00:00+00:00

Abstract

[Objective] To employ Word2Vec deep learning technology to identify disease associations from public-oriented health information, addressing the problem that non-medical professionals often lack understanding of associations among multiple diseases, thereby affecting the comprehensiveness and effectiveness of health information seeking. [Method] Thirty common disease topics were selected by experts, and documents corresponding to these diseases were collected from high-quality medical news websites. Word2Vec technology was utilized to construct word vectors for documents related to each disease, with vector distances calculated to determine disease associations. The accuracy of the results was evaluated through correlation analysis with expert ratings. [Results] Under optimal conditions, the correlation coefficient between Word2Vec results and expert ratings reached 0.635. By comparing the impacts of different algorithm models, optimization methods, data scales, and key parameters on the results, it was found that the Skip-Gram model combined with Negative Sampling optimization method (negative sample number of 20) yielded optimal experimental results on large-scale datasets. [Limitations] When disease topics are selected too broadly, the accuracy of Word2Vec judgments is affected. The granularity of disease topic selection in this paper requires improvement. [Conclusion] Word2Vec technology can detect disease associations in public-oriented health information sources. Its effectiveness demonstrates that this technology can be applied to improve personalized services for public health information seeking.

Full Text

Detecting Disease Associations with Word2Vec from Consumer Health Information

Luo Wenxin, Chen Chong, Deng Siyi
School of Government, Beijing Normal University, Beijing 100875, China

Abstract

[**Objective**] Average people usually do not know the complex associations among diseases, which poses negative effects to their health information seeking experience. This study tries to detect the associations among diseases using popular medical information with the help of deep learning technology (Word2Vec), aiming to improve personalized information services.

[**Methods**] First, we identified 30 common disease topics with the help of medical professionals, and then collected related reports from Medical News Today. Second, we built word vectors for each document using Word2Vec technology to calculate the semantic similarities among them. Finally, we compared the machine training results with experts' scores to evaluate the performance of the proposed method. We also investigated the impacts of different models, optimization methods, data sizes, and important parameters on the results.

[**Results**] The correlation coefficient between the Word2Vec results and the experts' scores reached 0.635 under optimal conditions. By comparing the effects of different algorithm models, optimization methods, data scales, and important parameters, we found that the Skip-Gram model combined with Negative Sampling optimization (with 20 negative samples) on large-scale datasets yielded the best experimental results.

[**Limitations**] When disease topics are selected too broadly, the accuracy of Word2Vec judgments is affected. The granularity of disease topic selection in this paper needs improvement.

[**Conclusions**] Word2Vec technology can detect disease associations from consumer-oriented health information sources. Its effectiveness demonstrates that this technology can be used to improve personalized health information seeking services for the public.

Keywords: Word2Vec, Disease Association, Non-professional Medical Text, Health Information, Personalization
Classification Codes: TP391, G350

---

1. Introduction

Traditionally, the general public obtained disease and health knowledge primarily from medical professionals. The development of the Internet now enables

people to actively search for health information they need online. In recent years, various new health service platforms have emerged, mostly focusing on disease knowledge popularization and online consultation, which greatly enriches channels for accessing medical information. However, due to their lack of specialized medical knowledge, the general public does not understand complex associations among diseases—for example, that periodontal disease may be caused by diabetes. This lack of understanding affects people' s ability to manage their health and search for comprehensive and effective medical information. If technical methods can identify associations among disease topics, this could improve personalized health information services and enhance content organization and navigation quality on information platforms. Since terminology in professional medical literature is not easily understood by the public, this paper uses non-professional medical information, such as high-quality health news, and employs Word2Vec deep learning technology to detect associations among disease topics based on disease-related documents. By comparing with expert evaluations, we find this technology can effectively be used for disease association detection.

Consumer health information services for the general public have long attracted attention. Eysenbach explicitly proposed using information technology to provide health information services for consumers, including analyzing consumers' health information needs, researching and implementing methods to provide information for consumers, and designing models to build information systems according to consumer preferences. This research area is known domestically as "consumer health informatics." Currently, consumer-oriented health services continue to emerge, providing disease knowledge popularization, customized information push, or online consultation for disease problems, promoting people to manage their health and improve their health information literacy.

To help people obtain health information more efficiently and accurately, researchers have conducted extensive work, mainly divided into several aspects: (1) investigating consumers' information seeking behavior to understand what types of content they care most about when searching for medical health information on the Internet; (2) helping people understand medical terminology to solve problems of difficult comprehension or biased understanding caused by the "vocabulary gap," such as developing Consumer Health Vocabularies (CHV) and predicting users' familiarity with health terminology; (3) establishing mappings from professional medical domain concepts to general cognitive categories to handle matching problems between user health vocabulary and UMLS vocabularies.

However, due to the complex associations among diseases, ordinary people without professional medical training find it difficult to understand these relationships, which affects their ability to obtain comprehensive relevant information during information seeking. Current research in this area is relatively lacking.

Traditionally, disease association detection is a task of clinical medical research or biomedical experiments. Existing studies using text mining to detect disease associations mainly focus on professional medical literature. For example, some

scholars have used semantic expansion models and neural network clustering methods to associate disease types with disease-causing genes. These research conclusions are mostly explained at the molecular biology, gene, and chemical component levels, which are difficult for ordinary people without professional knowledge to understand.

Consumer-oriented medical health information sources include health portal websites, medical news sites, online health communities, and public health knowledge bases. Research on user posts in the online health community MedHelp discovered relationships between drugs and their adverse reactions, helping drug safety regulators effectively identify early signals of adverse drug reactions. Text clustering analysis of user posts in specific disease communities analyzed the connections and differences among three types of diseases. These studies demonstrate that consumer health information sources can reveal meaningful connections for users. However, information quality in online health communities is not optimal. To ensure research reliability, this paper selects high-quality medical news. Additionally, there are studies using social network analysis to explore relationships among health topics. For example, Liu Hongxia et al. analyzed health information topics on the WHO website, using text similarity algorithms to mine their link relationships and semantic relationships, presenting them through social networks. However, this method relies too heavily on specific website link structures, limiting the associations that can be found. The study used text similarity algorithms that did not fully reflect semantic relationships. There is considerable room for in-depth research on disease or topic relationships using semantic analysis. This paper transforms the task of discovering disease associations into detecting semantic associations from disease-related documents, using Word2Vec (Word to Vector) technology to find words closely related to specific diseases and using this bridge to discover disease associations.

In 2003, Bengio et al. proposed the Neural Network Language Model (NNLM), which obtained word vectors while modeling natural language with neural network structures. In 2013, Mikolov et al. simplified the NNLM model and proposed the CBOW (Continuous Bag-Of-Words) model and Skip-Gram model, aiming to more efficiently implement word vector representations. In the same year, Google released C language implementations of these two models, calling them Word2Vec. Currently, the gensim package in Python libraries also integrates this algorithm. Word2Vec is based on deep learning ideas. By training on text datasets, it maps different grammatical and syntactic features of words to different dimensions of vectors, representing individual words as points in high-dimensional vector space. It mainly uses CBOW and Skip-Gram models to implement word vector representation. The difference is that the CBOW model predicts the center word given the context, while the Skip-Gram model predicts the context given the current word. Related research has proven that this technology performs well in word similarity calculation, machine translation, feature extraction, sentiment classification, and other fields. Word2Vec technology is universal and relatively simple to use.

3. Disease Association Detection

Unlike previous approaches that find disease associations from biological experiments and clinical perspectives, this paper transforms the task of detecting disease associations into discovering semantic associations from disease-related documents. Specifically, we use Word2Vec technology and medical health news to find disease associations, aiming to explore a general method to find association relationships among disease topics and improve the efficiency and effectiveness of people' s health information searching. This paper focuses on two main questions: (1) How to use Word2Vec to find association relationships among diseases? (2) How to evaluate the effectiveness of Word2Vec applied to disease association detection?

On the document collection related to specific diseases, we use Word2Vec technology to find word vectors that reveal different disease topics and determine disease topic associations through their similarity. Through statistical analysis methods, we compare the results with expert scoring results and determine the optimal parameter configuration through parameter tuning experiments.

3.1 Data Collection According to our research, unlike previous studies that mine disease associations in professional medical literature, this paper selects health information that ordinary people can understand because terminology in professional medical documents is difficult for the public to comprehend. Even if associations are found, they are difficult to directly apply to information sources that ordinary people frequently browse.

Data comes from the Medical News Today website. News articles on this site are written by professionals with medical backgrounds and manually tagged with category labels by the website. The content is high-quality and easily understood by ordinary people. Its category tags are divided into 144 categories according to health issues that the public cares about, with each category having corresponding news documents.

This study uses 30 representative disease categories: Addiction, Allergy, Alternative Medicine, Anxiety, Arthritis, Asthma, Breast Cancer, Cardiovascular, Cholesterol, COPD (Chronic Obstructive Pulmonary Disease), Dentistry, Depression, Diabetes, Eating Disorders, Flu, Headache, Heart Disease, HIV, Hypertension, Men' s Health, Mental Health, Neurology, Nutrition, Obesity, Pregnancy, Prostate, Seniors, Sleep, Women' s Health, and Stroke.

We collected health news from each selected disease category. For medical health news webpages, we used Python' s Natural Language Toolkit (NLTK version 3.2) for text preprocessing, including removing webpage noise, tokenization, case normalization, lemmatization, and stop word removal.

To compare the impact of datasets on algorithm results, we used three dataset

sizes: 3,000, 6,000, and 9,000 webpages, denoted as 3K, 6K, and 9K respectively. The 6K dataset was obtained by continuing to crawl 100 more webpages based on the first 100 webpages already crawled for each category, and the 9K dataset was obtained similarly.

3.2 Word2Vec Model Building Word2Vec implements word vector representation using the CBOW model and Skip-Gram model. Optimization methods for algorithm efficiency include Hierarchy SoftMax (HS) and Negative Sampling (NS). Combining them pairwise yields four training frameworks, as shown in Table 1 .

(1) CBOW Model and Skip-Gram Model

The CBOW model and Skip-Gram model are essentially optimizations of the Neural Network Language Model (NNLM). NNLM is a type of statistical language model, and its working principle is shown in Figure 1 [Figure 1: see original paper].

Given a corpus C, we build a vocabulary V with total size |V|. Assuming the word predicted by the language model is wi and its context is the previous (n-1) words, the NNLM model aims to maximize Equation (1):

P(w_i | w_{i-(n-1)}, ···, w_{i-1})

NNLM has a three-layer feedforward neural network structure. The input layer x is the sequential concatenation of word vectors of the previous (n-1) words, the hidden layer is h, and the output layer y is the remaining two layers of the neural network. H is the weight matrix from input layer to hidden layer, U is the weight matrix from hidden layer to output layer, b^{(1)} and b^{(2)} are bias terms, and tanh is the hyperbolic tangent function.

x = [e(w_{i-(n-1)}); e(w_{i-(n-2)}); ···; e(w_{i-1})]
h = tanh(b^{(1)} + Hx)
y = b^{(2)} + Uh

Notably, the output layer y has |V| elements, corresponding to the probability of the next word being each word in V. The SoftMax function is used to convert it to probability values:

P(w_i | w_{i-(n-1)}, ···, w_{i-1}) = exp(y(w_i)) / Σ_j exp(y(w_j))

During training, the optimization objective is to maximize Equation (6):

Σ_i log P(w_i | w_{i-(n-1)}, ···, w_{i-1})

In actual training, parameters are continuously iterated through stochastic gradient descent, with word vectors and intermediate matrices updated in each iteration. After optimization, the corresponding word vectors are generated.

Since matrix calculation from hidden layer to output layer is most time-consuming, CBOW and Skip-Gram models remove the hidden layer based on

NNLM, greatly reducing computational complexity while ensuring accuracy through expanded training samples.

The CBOW model structure is shown in Figure 2 [Figure 2: see original paper]. The context c takes (n-1)/2 words before and after word w_i, assuming all words in the context have equal weight on the probability of the current word appearing, regardless of order. The input layer changes from concatenating word vectors e(w_i) of context c to averaging (or summing) word vectors, as shown in Equation (7). The optimization objective during iteration is to maximize Equation (8), which also achieves word vector optimization.

The Skip-Gram model is shown in Figure 3 [Figure 3: see original paper]. It uses a "skip certain units" approach to expand training samples, increasing combinations of context words. It randomly selects a word vector w_j from the context c of word w_i as input. The optimization objective is to maximize Equation (9):

$\Sigma_{(w,c)\ C} \Sigma_{w' \quad c} \log P(w' \mid w)$

(2) Hierarchy SoftMax and Negative Sampling

To reduce model time complexity, Hierarchy SoftMax uses classification to differentiate words by frequency, part of speech, or topic. It abstracts word groups under a certain type into a word vector, using this abstract word vector to represent such words during calculation, thereby reducing computational complexity. For example, it constructs a Huffman tree using word frequency features for hierarchical classification, using abstract intermediate node vectors to approximate all child node vectors. Negative Sampling is relatively simpler, using negative sampling to improve training speed. During model iteration, it uses random negative sampling for calculation and updates instead of calculating the probability of the next word being any word in the vocabulary. Negative sampling samples are used as substitutes for all non-current words w(i). Its implementation has various algorithms, such as weighted sampling algorithms based on word frequency.

3.3 Model Training This paper uses the Word2Vec toolkit provided by Python's gensim module. The main parameters affecting experimental accuracy and efficiency during training are shown in Table 2 .

The sg parameter corresponds to model selection: 1 represents the Skip-Gram model, 0 represents the CBOW model. The hs parameter corresponds to optimization algorithm selection: 1 represents Hierarchy SoftMax algorithm, 0 represents Negative Sampling algorithm. The negative parameter corresponds to the number of negative samples in Negative Sampling algorithm. Size is the dimension of word vectors. As the size value increases, word vector accuracy first improves, but after reaching a certain extreme value, further increase in size value actually decreases accuracy. The min_{count} parameter is used to filter low-frequency words, equivalent to preprocessing that deletes words

with frequency lower than min_{count}. The sample parameter processes high-frequency words. Updating high-frequency words during iteration occupies certain time, while word vectors corresponding to high-frequency words change little. Therefore, Subsampling technology (downsampling) is used to skip certain high-frequency words during training. As shown in Equation (10), p(w) represents the probability of word w being skipped, where f(w) is the probability of the word appearing in corpus C. When f(w) > t, the larger f(w) is, the larger p(w) is, and the higher the probability of being skipped.

p(w) = 1 - sqrt(t / f(w))

The window parameter refers to training window size, related to context construction. Each time context(w) for word w is constructed, a random integer c is generated on [1, window], and c words are taken before and after w to form context(w). Generally, larger window values are better until reaching a certain extreme value. The workers parameter is the number of parallel threads for training models; larger workers values mean faster training speed and can be increased as much as possible according to computer performance.

3.4 Disease Topic Semantic Similarity Calculation The model training result maps each disease topic word to a point in N-dimensional vector space. We calculate the distance between word vectors in vector space using the cosine distance formula as their semantic similarity. Assuming the N-dimensional word vectors of two disease topics are:

t1 = (w_{1,1}, w_{1,2}, ···, w_{1,N})
t2 = (w_{2,1}, w_{2,2}, ···, w_{2,N})

The larger the cosine value, the more semantically similar disease topics t1 and t2 are. The calculation formula is:

cos( ) = (Σ_{k=1}^N w_{1,k} × w_{2,k}) / (||t1|| × ||t2||)

We calculate the semantic distance between each pair of the 30 disease topics, obtaining 435 groups of values.

---

4. Experimental Design and Results

Word2Vec' s effectiveness is influenced by data scale, model selection, and parameter settings. Experiments will test these aspects one by one and compare them with professional doctors' scoring results on associations among the 30 diseases. We denote expert scoring values as base values and use SPSS to calculate correlation analysis between training values and base values, obtaining the impact of various factors on results and evaluating the method' s practical usability.

4.1 Data Scale The vertical axis in Figure 4 [Figure 4: see original paper] represents the Pearson correlation coefficient between training results and base values. When the dataset expands from 3K to 6K, the effect improves to some extent, and when expanded to 9K, the correlation coefficient increases significantly. Additionally, under the 3K dataset, even the best result from Skip-Gram&HS shows only a marginally significant correlation. Adjusting various parameters, the optimal result using the 3K dataset has a correlation coefficient of 0.394 at the 0.01 level, which is less than 0.4. When the dataset increases to 9K, the initial correlation coefficient of the Skip-Gram model reaches 0.454. This shows that data scale is a key factor affecting Word2Vec training quality. Larger data scales yield better model effects.

Word2Vec is based on word context relationships to establish semantic relationships. As the dataset increases, words have more complete contextual information, and the trained word vectors better reflect the semantics of vocabulary in the corpus. The experiment shows that the 3K dataset is too small to adequately measure semantic similarity between words. Word2Vec technology does not perform well with small sample sizes.

4.2 Model Selection Figure 4 also compares results from the four training architectures. The Skip-Gram effect is significantly better than CBOW, but the latter has shorter actual running time. For the same corpus, Skip-Gram uses the "skip certain units" approach to expand training samples, which can also be seen as an increase in "data scale," thereby bringing increased model performance and longer training time.

From the perspective of optimization algorithms, taking the Skip-Gram model as a prerequisite, although Skip-Gram&HS is slightly more accurate than Skip-Gram&NS when the dataset is 9K, there is little difference between the two under 3K and 6K datasets, as shown in Figure 4. The Negative Sampling compared here has a negative sample value (negative) of 5. In fact, the negative sample value also affects the effectiveness of the Negative Sampling method.

Table 3 shows further comparison of negative sample values. Since training on the 9K dataset takes too long, we first reduce the word vector dimension size to 50 to shorten training time, then explore the impact of negative values on results. In the Skip-Gram&NS method, the larger the negative value, the higher the correlation coefficient. Although when negative is 5, the training result is not as good as the Hierarchy SoftMax algorithm; when negative is 20, the correlation coefficient reaches 0.539, which is much higher than the Hierarchy SoftMax algorithm.

In summary, Skip-Gram and Negative Sampling, when the negative sample value is 20, yield better training results. The following parameter selections are all based on the premise of the Skip-Gram&NS method.

4.3 Parameter Comparison To understand the impact of word vector dimension size on algorithm results, we first use the smaller and faster dataset to find the effect of size parameter changes on results, then select the parameter value range that yields optimal results for further observation. On the 3K dataset, controlling size values within [50, 500], we find that word vector dimension values are not larger-the-better. Values in [50, 100] yield better results, as shown in Figure 5 [Figure 5: see original paper]. Further, on the 9K dataset, we narrow size values to the [50, 100] range and find that when the word vector dimension is 50, Skip-Gram&NS has the highest correlation with expert scores, as shown in Figure 6 [Figure 6: see original paper].

We then investigate the impact of high-frequency word sampling threshold sample on results, fixing the already-tested parameters at their optimal values (negative=20, size=50). The Google Word2Vec toolkit recommends changing sample values within [1e-5, 1e-3], so we set this parameter to several values as shown in Figure 7 [Figure 7: see original paper]. The results show that the smaller the sample value, the higher the correlation between training results and base values, and the significantly shorter training time. High-frequency words appear many times in the corpus and provide less useful information, with corresponding word vectors changing little during training. From Equation (10), the smaller the sample value, the more words in the corpus have appearance probabilities higher than sample, and the more high-frequency words are skipped in subsampling, leading to higher accuracy. On the 9K dataset, sample value of 1e-5 yields better results, with a correlation coefficient of 0.614, showing significant improvement.

Setting sample=1e-5, we further examine the low-frequency word threshold min_{count}, changing its value from 40 with a step size of 20. As shown in Table 4 , within the (40, 100) range, changes in min_{count} have little impact on results. Removing words with frequency lower than 40 before training when creating the vocabulary yields better results.

Regarding context window window, theoretically larger values are better, but expanding window increases training time. Changing the window parameter in the [50, 200] range and comparing with expert scores yields results shown in Figure 8 [Figure 8: see original paper]: when window is around 50, the correlation coefficient no longer increases, reaching 0.635. At this point, when Skip-Gram takes context samples, it generates a random integer c in the [1, 50] range, then takes c words before and after word w to form context(w).

In summary, through exploration of parameter factors, we find that using the Skip-Gram model combined with Negative Sampling algorithm (negative sample value of 20), with word vector dimension of 50, high-frequency word sampling threshold of 1e-5, low-frequency word threshold of 40, and context window of 50, the similarity measurement results from the trained model are optimal, with a correlation coefficient with base values reaching 0.635. We denote this result as W2V.

5. Analysis of Word2Vec' s Disease Association Detection Effect

We conduct detailed comparative analysis between the optimal experimental result W2V and base values. After normalizing W2V values, we obtain a scatter plot as shown in Figure 9 [Figure 9: see original paper]. We sort the 435 groups of values by base values from largest to smallest, number them from 1 to 435, and use the numbers as the horizontal axis with base values and W2V values as the vertical axis.

As base values decrease from large to small, W2V values also show an overall decreasing trend. In regions with higher semantic similarity, W2V values are more dispersed; in regions with lower similarity, W2V values are relatively more concentrated. Word2Vec calculation results have good overall performance but need further improvement in local performance.

Dividing data into 7 intervals according to base value ranges, corresponding to base values from high to low, we can easily find from Table 5 that in intervals with higher base values, the corresponding minimum W2V values are higher than minimum values in lower intervals. The distribution of means is consistent with the trend of base value intervals—higher intervals have larger means, which also matches conclusions from Figure 9. However, relatively speaking, the range of W2V mean variation is smaller. The median change trend is basically consistent with the mean. The interval with base value of 1 has the maximum standard deviation of 0.157; the interval with base value of 0 has the minimum standard deviation of 0.127. This indicates that in intervals with highest similarity, W2V points have greater dispersion, while in intervals with lowest similarity, W2V points have the most concentrated distribution.

From disease topic semantic associations obtained through Word2Vec training, sorted by similarity from high to low, the top 10 most similar disease topic pairs are shown in Table 6 . Among the corresponding base values, 6 groups of diseases are highly related, and 3 groups are also highly related. Only the relationship between Men' s Health and Women' s Health shows high correlation in Word2Vec calculation but only 0.5 in expert scoring, with a large difference. This may be because Word2Vec uses vector addition to represent phrases—Men' s Health and Women' s Health vectors are respectively the sum of vectors for words "health" and "men" / "women." During calculation, the similarity between the two increases accordingly. Especially since words "men" and "women" have high frequency in the corpus, are not specific, and are semantically similar, thus overestimating the association between Men' s Health and Women' s Health. Moreover, these two disease categories themselves have broad scopes, affecting Word2Vec' s judgment accuracy.

6. Conclusion

This paper selected 30 disease topics, collected news text from Medical News Today, used Word2Vec technology to calculate associations among diseases, and compared results with expert scoring. The study found that larger data scales yield better model effects but require longer training time. The Skip-Gram model combined with Negative Sampling optimization (with 20 negative samples) yields optimal experimental results on large-scale datasets. Smaller high-frequency word subsampling thresholds produce better training effects and shorter training times. Under optimal conditions, the correlation coefficient between training results and expert scores reaches 0.635. In regions with higher semantic similarity, Word2Vec training values are more dispersed; in regions with lower similarity, Word2Vec training values are relatively more concentrated. Sorting Word2Vec training results by similarity from high to low, 9 of the top 10 disease relationship groups also show high correlation in expert scoring.

Word2Vec technology can detect disease associations from consumer-oriented health information sources. Its effectiveness demonstrates that this technology can be used to improve personalized health information seeking services for the public.

Future research will expand in the following directions: expanding the dataset —Word2Vec shows significant improvement with increased data scale, and using more data in practice can yield more ideal results; adjusting disease types and conducting association research at finer granularity.

References

[1] Kempson E. Review Article: Consumer Health Information Services [J]. Health Libraries Review, 1984, 1(3): 127-144.
[2] Eysenbach G. Recent Advances: Consumer Health Informatics [J]. BMJ Clinical Research, 2000, 320(7251): 1713-1716.
[3] Hou Xiaoni, Sun Jing. Research on Internet Health Information Searching Behaviors of Outpatients from Tertiary Referral Hospital in Beijing [J]. Library and Information Service, 2015, 59(20): 126-131, 11.
[4] Klavans J L, Muresan S. Evaluation of the DEFINDER System for Fully Automatic Glossary Construction [C]. In: Proceedings AMIA Annual Symposium. 2001: 324-328.
[5] Zeng-Treitler Q, Tse T. Exploring and Developing Consumer Health Vocabularies [J]. Journal of the American Medical Informatics Association, 2006, 13(1): 24-29.
[6] Zeng-Treitler Q, Goryachev S, Tse T, et al. Estimating Consumer Familiarity with Health Terminology: A Context-based Approach [J]. Journal of the American Medical Informatics Association, 2008, 15(3): 349-356.
[7] Burgun A, Bodenreider O. Mapping the UMLS Semantic Network into General Ontologies [C]. In: Proceedings of Annual Symposium. 2001: 81-85.

[8] Keselman A, Smith C A, Divita G, et al. Consumer Health Concepts that do not Map to the UMLS: Where do They Fit? [J]. Journal of the American Medical Informatics Association, 2008, 15(4): 496-505.
[9] Yang Z H, Lin H F, Li Y P, et al. TREC 2005 Genomics Track Experiments at DUTAI [C]. In: Proceedings of the 14th Text REtrieval Conference. 2005: 1-9.
[10] Yang Z H, Lin H F, Li Y P, et al. DUTIR at TREC 2006 Genomics and Enterprise Tracks [C]. In: Proceedings of the 15th Text REtrieval Conference. 2006: 1-10.
[11] Jiang Q, Wang Y, Hao Y, et al. miR2Disease: A Manually Curated Database for microRNA Deregulation in Human Disease [J]. Nucleic Acids Research, 2009, 37(Database issue): D98-104.
[12] Yang H, Yang C C. Using Health Consumer Contributed Data to Detect Adverse Drug Reactions by Association Mining with Temporal Analysis [J]. ACM Transactions on Intelligent Systems & Technology, 2015, 6(4): Article No.55.
[13] Chen A T. Exploring Online Support Spaces: Using Cluster Analysis to Examine Breast Cancer, Diabetes and Fibromyalgia Support Groups [J]. Patient Education and Counseling, 2012, 87(2): 250-257.
[14] Liu Hongxia, Zhang Jin, Chen Jinghao. Social Network Analysis of Semantic Links Relationships Among Health Topics in WHO English Website [J]. Library and Information Service, 2014, 58(13): 75-82.
[15] Bengio Y, Schwenk H, Senécal J-S, et al. A Neural Probabilistic Language Model [J]. Journal of Machine Learning Research, 2003, 3(6): 1137-1155.
[16] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space [OL]. [2016-05-13]. http://arxiv.org/pdf/1301.3781v3.pdf.
[17] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and Their Compositionality [A]. //Advances in Neural Information Processing Systems [M]. 2013: 3111-3119.
[18] Handler A. An Empirical Study of Semantic Similarity in WordNet and Word2Vec [D]. Columbia University, 2014.
[19] Amunategui M, Markwell T, Rozenfeld Y. Prediction Using Note Text: Synthetic Feature Creation with Word2Vec [J]. Computer Science, 2015(3): 1-6.
[20] Ju R, Zhou P, Li C H, et al. An Efficient Method for Document Categorization Based on Word2Vec and Latent Semantic Analysis [C]. In: Proceedings of the 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM). IEEE, 2015: 2276-2283.
[21] Su Z, Xu H, Zhang D, et al. Chinese Sentiment Classification Using a Neural Network Tool —Word2Vec [C]. In: Proceedings of the 2014 International Conference on Multisensor Fusion and Information Integration for Intelligent Systems (MFI). IEEE, 2014: 1-6.

Author Contributions

Chen Chong: Proposed research ideas, designed research plan, revised paper.
Luo Wenxin: Conducted experiments, collected, cleaned and analyzed data, drafted paper.
Deng Siyi: Participated in literature research.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data is self-archived by authors, E-mail: luowenxin1994@foxmail.com, chenchong@bnu.edu.cn.

[1] Luo Wenxin. Word2Vec Disease Similarity Results.xlsx. The disease associations obtained by Word2Vec results (i.e., W2V) with the highest correlation with expert scoring results (Pearson 0.635) in this paper.
[2] Luo Wenxin. Word2Vec Word Vector Training Results.txt. The word vectors corresponding to W2V results, used to calculate semantic similarity of diseases.
[3] Chen Chong. 30 Disease News Webpages.html. 9,000 news webpages crawled from the Medical News Today website.

Received: 2016-05-16
Revised: 2016-05-22

EBSCO Further Funds Koha

EBSCO announced continued advocacy for open source and open access, providing further funding for Koha. Koha is the world' s first feature-rich, free, open-source integrated library system, used by more than 15,000 libraries of various types worldwide as their integrated library system.

EBSCO began providing financial support for Koha in February 2015. This latest funding for Koha will help Koha with next-stage functional improvements, such as additional system interoperability, as well as acquisition and electronic resource management functions. Specifically, this includes: (1) developing a procurement API; (2) fully implementing ordering and invoicing systems; (3) improving interoperability between Koha and CORAL, providing an open-source solution that combines traditional integrated library system workflows with ERM functions.

Koha will adhere to open-source traditions, and these enhanced functions for Koha will also be open-source, available for others to use, modify, and redeploy.

These enhanced functions are expected to be completed in the first quarter of 2017.

(Compiled from: https://www.ebsco.com/news-center/press-releases/ebsco-information-services-continues-to-support-open-source-technology)
(Journal News)

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