International Journal of Knowledge and Language Processing
Volume 6, Number 4, **2015**

www.ijklp.org
KLP International ⓒ2015 ISSN 2191-2734
pp.37–49

The Review of Patent Text Similarity Algorithms

Wang Qingxu
School of Software and Microelectronics
Peking University
No. 24, Jinyuan Road, Daxing District
Beijing, 102600, China
wangqingxu@pku.edu.cn

Received December 2015; Accepted December 2015

Abstract: *Patent similarity is an important part of the patent data mining, this review summarizes the currently research status of patent text similarity algorithms. Compared with the domestic and foreign research, this article classifies these algorithms, and analyses the idea, the existing problems and the development, laying the foundation of the next step research.*

Key words: patent text, similarity algorithm, patent similarity

1. **Introduction.** Nowadays, the society is an information society. As a unique information strategic resource, patent plays an important role in the development of national strategic resources. Patent is a special kind of literature, containing a large number of scientific and technical information, and patent application is one of the key ways for companies to protect intellectual property. At the time of applying for a patent, companies should search the patent databases to find similar patents, for the comparison between similar patents can help companies find new invention, the relevant prior art, the infringement of detection, competitive intelligence, and gap analysis, discovery of new opportunities for invention, patent portfolio analysis, and so on.

Traditional methods to find similar patents are generally achieved by searching for key words and then through manual filtration. However, this method is time-consuming and inefficient. Some of foreign patent analysis platforms currently integrate the function of similarity retrieval, and the analysis process is usually divided into data retrieval, cleaning, processing, analysis and application. Domestic patent analysis platform mainly realizes the

function of patent search, patent documents upload and download, and patent analysis, but the function of patent similarity measurement is still relatively rare.

At present, scholars at home and abroad proposed some patent similarity algorithms combining patent text features, including the algorithm based on patent citation, the algorithm based on patent ontology, the algorithm based on patent function trees, and the algorithm based on patent compound concepts. Based on the improvement of text similarity measurement algorithms, most of these algorithms have some advantages and limitations. From the relevant literature, the experimental data set is not large enough, and the feasibility of the application of large amounts of patent data need to be further explored.

For the research of text similarity algorithm, foreign countries started research earlier. Up to now, scholars at home and abroad have proposed many text similarity algorithms, hybrid algorithm and improved algorithm based on previous algorithms. At present, in terms of the similarity of the patent text, the algorithm involved can be divided into four categories: vector space model, semantic similarity algorithm based on ontology, implicit semantic model and the compound algorithm base on patent content. The following sections will introduce the domestic and foreign research progress of the four kinds of algorithms.

2. **Vector Space Model.**

2.1. **The Algorithms' Idea.** Proposed by G. Saltonet al. [1] at the end of 1960s, the vector space model (VSM) is the most widely used method of text similarity computation in practice. Document is viewed as a vector in n-dimensional space; a document is represented through a set of features and the corresponding rights, D{T_1, T_2, T_3,... T_i,…,T_n}; T_i is a feature item extracted from the document.

Feature weights are usually counted by TF-IDF. TF refers to the frequency of features in a text. IDF refers to the inverse document frequency, which the times of the feature items appeared in all the documents, less means more to distinguish between different documents. TFC method is which the text length normalized on the basis of the TF-IDF; On the basis of TFC, ITC uses the log value of TF instead of TF value, the weight of entropy based on information theory. The TF-IWF is based on the TF-IDF algorithm, using the IWF (the log value of the reciprocal value of the feature items frequency) instead of the IDF, and using the square of the IWF to balance the weights.

The distance between the vectors is usually calculated by the transvection between vectors, if considering the normalized, using the cosine angle between two vectors to represent the similarity coefficient. Some other similarity algorithms can also be used. For instance, Jaccard Similarity is the intersection of two sets divided by the union of two sets, in order to get the similarity between the two [2].

2.2. **Improved Algorithms in Chinese.** In foreign countries, the improvement of traditional VSM and new models are mostly for the English text, these algorithms can't be directly applied to Chinese text because of the differences between English and Chinese. At home, in view of Chinese text, many scholars have put forward a lot of improvements to traditional VSM at present.

Liu Shaohui et al. [3] proposed to add information to improve the weight of the TFIDF, classify the text, establish a class hierarchy tree, and then divided the document into sub categories and make similarity comparison. According to the test results, the accuracy and recall rate of VSM is higher than that of the traditional VSM, but the weight of the combined information is not affected by the weight of the experimental results.

Chen Zhiping et al. [4] proposed the hierarchical structure of the document and set N layers of VSM. They took 1000 computer articles from online as the test set, selected 3000 commonly used words from computer dictionary and built the feature library, where the document is divided into three parts as the title, the abstract and the text, then established three layers of vector space model, and the tests showed that the recall and precision of three layers of vector space model are 2% higher than the traditional VSM, besides that the time complexity and computation are decreased.

Cao Tian et al. [6] proposed a new algorithm based on VSM and word co-occurrence, which the relevant word sequences were added when calculating TF-IDF and the co-occurrence value of the word was added when calculating similarity. Selecting ten texts among the ten types of 2815 texts, testing the recall and correct rate when the feature of the co-occurrence words was added or not. It shows the average accuracy rate was increased from 65.7% to 75.1%, and the average recall rate increased from 79.9% to 87.6%.In addition, Zhang Zhang et al. [7], Chang Peng et al. [8] all improved the traditional VSM based on word co-occurrence, and applied it in text clustering, the accuracy and recall rate were significantly improved. But the disadvantage of this algorithm is that it is not ideal for short text processing.

2.3. **Application of Algorithm in Patent Text.** Patent text has a basic fixed structure, the literature [9] based on text mining technology to layer the patent text for patent title, abstract, claims and specifications, using the traditional VSM+TFI-DF algorithm, summed the weighted vectors of the four elements to calculate the similarity, and extracted 1286 patents in the field of carbon nanotubes in USPTO patent network database, and an empirical study was carried out to verify its feasibility.

The literature [10] is further, building the patent document structure tree which includes 6 elements (title, abstract, specification, claims, IPC classification and citation). According to the two comparable vector texts, the corresponding text of the 6 elements were expressed as vectors, the similarity between the two vectors was calculated by the cosine of the angle between the vectors, and then the similarity of each part of the patent document structure tree was weighted sum. From the three experiments that about 134 U.S. patents of Derwent patent data, compared to the weighted method that without considering the main classification number and citation, the accuracy rate, recall rate and F1 index of this method increased by 22.31%, 14.53% and 20.30% respectively; compared with the non-hierarchical VSM method, the accuracy rate, recall rate and F1 index increased by 79.26%, 51.93% and 72.63%.

Based on the description of the functional application description of the classification number and the description of the technology inheritance and evolution of citation, the

paper [11] proposed a similarity measurement method based on citation and a variety of classification numbers (IPC patent classification, Derwent manual classification, Derwent classification code), and compared and analysed with the method based on CO word. The former is better to distinguish the higher degree similarity and lower degree similarity of patents, the latter is more suitable to distinguish the normal similarity of the patent, and the combination of the two is better.

In summary, vector space model is a simple and effective similarity algorithm of text, but it assumes that it is linearly independent between words, without considering the semantic relations of the text. Aiming at this shortcoming, some improved algorithms proposed hierarchical document, combination of word co-occurrence, combination of the semantic model. Compared with the traditional VSM, the accuracy rate and recall rate of these improved algorithms have increased. Among these algorithms, the improved algorithm of hierarchical document and CO word has been applied in the patent text similarity comparison. The layer method which according to patent title, abstract and rights has certain feasibility.

3. **Ontology-based Semantic Similarity Algorithm.**

3.1. **The Algorithms' Idea.** Semantic similarity algorithm is on a lexical level, in essence, it is based on the knowledge base. Semantic similarity calculations often involve concepts' ontology. WordNet [12] or the knowledge encyclopaedia Wikipedia [13] has a major role in English semantic similarity calculation as a reference for general ontology. Domain ontology is also important. Hownet is often used in Chinese semantic similarity calculation as the ontology concept.

This algorithm is relatively mature in English text processing field. Chinese researchers have made a survey on it. This article [14] has made a comprehensive and clear summery. Semantic similarity algorithm based on the ontology is divided into two categories: methods based mainly on ontology tree; methods based on directed graph.

1. Algorithms based mainly or wholly on tree-structure

The article [21] put these algorithms into algorithms based on simple structures and algorithms based on complex structures, semantic similarity and relevance calculation based on the content, semantic similarity and relevance calculation based on properties and hybrid algorithms.

The algorithms based on a simple structure mainly refer to the semantic similarity calculation based on the distance calculation. The basic idea is to calculate the distance of two concept words in the ontology tree, namely the semantic distance. The farther the semantic distance between two concepts of words are, the lower their similarity is; vice versa. Representative algorithms are Shortest Path method [15], Weighted Links method [16], Wu and Palmer method [17], Leacock and Chodorow method [18] and so on.

Algorithms based on complex ontology structure take ontology structure into consideration. J.W.Kim [19], who proposed CP / CV concept propagation methods depending on the semantic relationships between concepts in the ontology hierarchy; on this basis, the article [20] proposed a concept vector model based on the local density of

related concept node. First it defines the concept nodes in ontology structure tree and assigns different weights to them in order to form concept vectors. Experiments show that the improving effect of the method is relatively good.

Semantic similarity algorithm based on the information content is to compare the information that contained by parent nodes shared by concept word pairs. Lord et al [21] proposed to calculate the similarity between word pairs by information of the nearest public parent nodes words. Resnik [22] proposed to use the most informative parent node. Lin [23] thinks that in addition to shared information, the information owned by each concept should also be taken into consideration. When concept words belong to the same set of ontology, Lin method is better. Jiang and Conrath method [24] on the basis of Lin, directly calculate the semantic distance to show the similarity between the concept pairs.

Algorithms based on property thinks that the similarity will be higher if the number of public properties is larger between two concepts. A representative algorithm is Tversky algorithm [25]. It mainly uses a set of properties information of related ontology. Banerjee and Pedersen [26] and Patwardhan et al. [27, 28] proposed a method based on concept gloss. The overlap of two concepts' gloss can show their similarity.

Hybrid algorithm is actually a combination of the above algorithms, that is to say the hybrid methods taking the location of the concept, edge type and properties information into account, such as algorithms proposed by Li et al. [29], which compares the shortest path between words, the depth of nearest common parent nodes as well as the local density information of the words position. Marco [30], who proposed algorithm based on graph model, combines the ontology structure and this method. Literature[s] [31-38] are mostly based on improved one of the above calculation methods, hybrid methods and specific applications.

Mihalcea, R, et al in the literature[39] made experiment to evaluate the algorithm that based on corpus (PMR-IR, LSA) and algorithm that based on the ontology (J&C, L & C, Lesk, Lin, W & P, Resnik), and evaluate some of these combined algorithms . The results show that the accuracy rate of the six algorithms which based on semantic reached 70.3%, the F value reached 81.3%, the effect was the best.

2. Semantic Similarity Algorithm Based On Directed Graph Structure

This kind of algorithm is no longer based on WordNet, but is based on Wikipedia. Wikipedia has good structured information; it can be seen as two huge networks: a composed of webpages and another composed of categories. These two networks both can be abstracted into directed graphs (DAG), so the processing of the category network and webpage network of Wikipedia can be regarded as the processing of the directed graph. Representative algorithms are WikiRelate! [40], Semantic Analysis Explicit (ESA) [41] and Link Vector Model Wikipedia (WLVM) [42]. Compared with the algorithm based on tree, this kind of algorithm has a low degree of correlation. Although Wikipedia contains rich semantic information, but its data noise is larger, compared with WordNet, the structure of its data is not strong, so the calculation effect is generally poor.

3.2. **Improved Algorithms in Chinese.** Chinese words semantic similarity computation is usually based on HowNet, a knowledge system which is created by the famous Chinese machine translation experts Mr. Zhen Dong Dong. [43] The meaning of HowNet is to describe something through the sememe. Sememe is the most basic minimum meaningful unit in Chinese. There is a network of words. Liu Qun [44] and some Chinese scholars analysed the knowledge description structure of HowNet, using the hypernym-hyponym relation of sememe to calculate the similarity, and got the similarity of words.

$$\mathrm{Sim}(S_2,\ S_2) = \frac{a}{a+distance(S_2,\ S_2)} \tag{1}$$

On the basis of Liu Qun [44], Yu Gang et al. [45] proposed the method to calculate text similarity based on lexical semantic, each document is represented as one of the feature vectors, not being normalized and disambiguated, computing the similarity of any word entry of the two documents to find the maximum weight, adding the weight of the word to get the similarity of the two documents. This algorithm is similar to solve the maximum weight matching of a complete bipartite graph.

According to the characteristics of the different structure of the ontology, considering the factors that affect the similarity are semantic coincidence degree, semantic distance, the width, depth and density of concept, Li Wenjie[46] and some scholars calculated the similarity of these factors, and multiplied them to obtain the final similarity, establishing a semantic similarity algorithm based on ontology structure.

Yang Fangying [47] and others took advantage of the structure of the ontology, according to the method based on edge and the method based on the vertex, combined with Tversky's point of view about attributes, weighted the distance, attributes, the common parent node level and the information feature to calculate the similarity. The experimental part used the amino acids ontology public released by Wikipedia as an example, by comparing with Liu Qun[44], Resnik[22], Lin[23], calculating the Euclidean distance between the results and results of the field experts. It is proved that the effect of this algorithm is best.

On the basis of Li Wenjie[46], Yang Nana et al. [48] proposed that the semantic similarity is firstly affected by the attributes of the concept, and secondly related to the structure factor of concept tree. They defined the the relationship between the concepts as synonyms, inheritance, part and the whole, the other, and gave different weight respectively, summed the weight to calculate the similarity. The feasibility and accuracy of the experiment are explained by using the semantic similarity of the entities in the land use classification.

3.3. **Application of Algorithm in Patent Text.** Zhou Qunfang et al. [49] proposed a similar patent detection method based on ontology. This method actually is a mixed method that combined with ontology, traditional VSM, and LCS algorithm. Firstly, the paper represented the sibling class in ontology model as GUID. Secondly, it used the GUID replaced the cut words, calculated the TFI-DF value of each word as the weight, using cosine similar degrees to extract the similar documents. Finally, it matched the similar sentences using the last longest common subsequence matching algorithm.

In conclusion, the ontology-based semantic similarity algorithm is an algorithm based on

knowledge base, taking into account the semantic relevance of texts; the calculation is comprehensive and accurate through the network knowledge of the knowledge. But this kind of similarity calculation is limited range of words or sentence, the calculation efficiency of the full text is not high. In terms of patents similarity, there are scholars combine it with traditional VSM or LCS algorithm to calculate patent text similarity. The experiments demonstrated their feasibility, but due to the limited experimental set, the feasibility under large scale corpus needs to be further studied.

4. **Latent Semantic Model.**

4.1. **The Algorithms' Idea.** Latent semantic model was proposed by Deerwester [50] et al. in 1990. It is assumed that the relevant words will appear in similar texts, in fact, is an extension of the generalized vector space model. [51] LSA uses terms as line and documents as column to build a matrix, the elements of the matrix are the TF-IFD value of the terms, using Singular Value Decomposition (SVD) to achieve the effect of dimensionality reduction of the matrix.

$$A=U \Sigma V^T$$

$$L= U \Sigma^{-1}$$

$$sim(q,d)=\cos(LTq, LTd)$$

Matrix A can be decomposed into the product of three matrices U , $\sum$and V^T, the column vectors of the matrix U and V matrix is orthonormal, the matrix $\sum$ is a diagonal matrix. U is m * t matrix, $\sum$ is t * t matrix, V is n * t matrix. It can reduce the number of columns, and maintain the similar structure of each line. By comparing the vector cosine angle between any two lines, the word similarity can be compared.

4.2. **The Development of Algorithm.** In order to overcome the shortcomings of LSA, Hofmann et al.[52] proposed a probabilistic latent semantic analysis (PLSA) model in 2001. PLSA avoids the matrix decomposition and complex calculation of SVD by introducing the idea of probability and statistics. In PLSA, it firstly calculated the conditional probability distribution of the documents and the potential topics, the words and the potential topics, then used EM algorithm to estimate parameters so that it can get the probability distribution matrix of the documents and the potential topics in the end. PLSA also has some disadvantages: the probability matrix will become very large with the number of documents and words increasing, and the EM algorithm needs to be iterated repeatedly, the amount of computation is also great.

In view of the shortcomings of PLSA, Blei et al. [53] proposed a new topic model LDA (Latent Dirichlet Allocation) in 2003, which is a hierarchical Bayesian model. The parameters of the model are also considered as random variables, there are two main methods to estimate the parameters of LDA: Gibbs Sampling method (the computational quantity is large, but relatively simple and accurate) and Variational Bayesian Inference method (small amount of calculation, low precision).

In China, on the basis of LDA model, Wang Zhenzhen et al. [54] and Sun Changnian et al. [55] also improved the similarity measurement method for Chinese text. First, the text was

pre-processed and expressed as a document feature words matrix, and then built the LDA model, using the JS (Jessen-Shannon) distance between the texts to calculate similarity to get the similarity matrix, and finally used the K-means algorithm to cluster to evaluate the accuracy of the similarity calculation. In the experiment of the paper, the parameters were estimated by Gibbs sampling method, extracting 8 subset of Fudan Chinese corpus to compare the F value (the balance index of accuracy and recall rate) of the result with the F value of the traditional TFI-DF. The result showed that F value fluctuated markedly in different themes, but the F value of LDA model was significantly higher than that of the traditional TF-IDF.

4.3. **Application of Algorithm in Patent Text.** Literature [56] divided the patent claims into hierarchical levels, used the dependency tree to analyse the relationship between two words in a claim sentence and made POS tagging for the words, then it established templates for each cluster and used the templates to extract the key components of patents (attributes and functions). In order to solve the problem of synonyms in patents, the literature used PLSA (Latent Semantic Analysis Probabilistic) theme model on the extracted patent attributes and functions to get the theme-document probability matrix, and then used the calculated theme distance to express the similarity.

Then the literature designed an experiment to compare the efficiency of extracting the similarity of the key text and the whole text. It chose the patents in the field of oil exploration in 2015 in USPTO on-line patent database as the experimental text, and the results showed an 87.5% increase in the text analysis speed and a 41.1% increase in the precision. However, this algorithm is a model for the English patent text, and the design of the template for extracting the key components is very important, an over-complicated template will also cause problems, so whether this algorithm is effective for Chinese patents should be further discussed.

In summary, the advantage of latent semantic model is that it transformed the text feature space to concept space by the decomposition of the singular value,, and the calculation of the cosine between the inner product and the included angle of the concepts is more reliable; while the disadvantage of the model is that its effect for the sparse corpora is not good as it depends on the context information. The improved algorithms include PLSA, LDA and so on, these algorithms introduce the probability model, but they also face the problem of too much computational work. Among them, PLSA has been applied in the similarity comparison of patent text. The experimental results proved its high accuracy, but as the application in this aspect is still relatively rare, its effect needs to be further verified.

5. **Compound Algorithm Base on Patent Content.** Patent text is a kind of special text which has its special structure and characteristics. When some scholars are studying the patent text similarity comparison algorithm, starting from the text characteristics, they put forward some composite algorithms based on the patent text content, mainly are the algorithm based on the invention of the function tree and the algorithm based on the extraction of composite concept.

5.1. **Based on the Invention of Functional Tree.** At present, the traditional text similarity calculation method is dependent on the language style of the author. Although a same inventor or patents from the same company was different, they were clustered together. However, the similar invention concepts are separated because of the use of the wording. In the literature [57], the invention of function tree was used to measure the similarity among patents. Compared with the components of the invention and their level, functional connectivity, meanwhile, a sample of current circuit breaker field experiment was carried out. It is assumed that the two are used in the same area, the same parts and organizational structure, and has the same functional interaction technology system, from theoretically speaking is exactly the same.

This method took into account the distance of the concept of patent, so the patent can be classified into a more appropriate concept group, and can produce more similar clusters in nature, which avoid the results cause by simply determine the similarity among patents by co-occurrence of patents. But the set of the weight coefficient in the formula, and the threshold setting of choosing components, will greatly affect the similarity results, and needed to be reasonably adjusted.

5.2. **Based on the Extraction of Composite Concept.** The paper [58] proposed the definition of window to extract the compound concept of the patent; firstly, it extracted a single concept, and then extracted the compound concept based on the single concept. After extracting the compound concept, it established a model based on the set theory to calculate the similarity. According to the extracted compound concepts, defined four subsets and six variables and calculated the variables and their relationship. Then establish a model to link the subsets to variables. According to the relevant variables, the text similarity coefficient is divided into two groups, the 1 group is based on the analysis of the patent double set, and the 2 group is based on the analysis of the patent of one side / double side. The coefficients of the 1 group were calculated by Jaccard, Inclusion and Cosine, and the coefficients of the 2 group were calculated by DSS-Jaccard, DSS-Inclusion and DSS-Gamma-Inclusion.

From the experimental results, in the analysis of patent priority and the convergence of patent in different areas, this model works well. However, the step of extracting compound concept from the single concept should be designed according to the comparable patent text. The design of the file window size will affect the results of the similarity, which needs to be reasonably adjusted.

In summary, these compound algorithms based on patent content mainly focuses on the pre-process of patent text, and transform the text similarity comparison into the similarity comparison of patent concept. The appropriate extraction method will greatly improve the accuracy of the calculation of the concept distances, but the size of the extraction window and other factors need to be decided by the corpus, and there are some uncertain factors. This kind of algorithm needs to be verified by large-scale corpus.

6. **Summary.** In this paper, the domestic and foreign research status of patent text similarity algorithm is summarized. In order to facilitate the research, this article divides

the algorithms applied in patent text into four categories: the vector space model, the semantic similarity algorithm based on ontology, the latent semantic model and the compound algorithm base on patent content. Compared with the domestic and foreign research, this article analyse these four kinds of algorithms and their existing problems respectively, laying the foundation of the next step of the patent similarity algorithm research.

Acknowledgement. This work is partially supported by National Key Project of Scientific and Technical Supporting Programs No. 2013BAH21B02; the author also gratefully acknowledges the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] G. Saltan, A. Wong, and C.S. Yang, A Vector Space Model for Information Retrieval, *Journal of the ASIS*, 18:11,613-620, November 1975

[2] Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. B*ulletin de la Société Vaudoise des Sciences Naturelles 3*7, 547-579

[3] Liu Shaohui, Dong Mingkai, Zhang Haijun, Li Rong, Professor zhongzhi Shi. An Approach of Multi-hierarchy Text Classification Based on Vector Space Model. *Journal of Chinese information Processing*, 2001,16 (3): 8-14

[4] Chen Zhiping, Lin Yaping, Tong Diaosheng. An Information—Retrieval Method Based on N-level Vector Model. *Journal of Computer Research and Development,* 2002 (10)

[5] Li fan, Lin Aiwu, Chen Guoshe. A Chinese text categorization system based on the improved VSM. *Journal of Huazhong University of science and technology (natural science Edition),* 2005,33 (3): 53-55

[6] Cao Tian, Zhou Li, Zhang Guo-xuan. Text similarity computing based on Word co-occurrence. *computer engineering and science*, 2007,29 (3)

[7] Zhang Zhang, and Fan Xiaozhong. Improved VSM Based on Chinese Text Categorization. *computer engineering and design*, 2006,27 (21): 4078-4080

[8] Chang Peng, Feng Nan, Ma Hui. Document Clustering algorithm based on co-occurrence. *computer engineering*, 2012,38 (2): 213-214

[9] Peng Jidong and Tan Zongying. Patent similarity measurement method based on text mining and its application . *Theory and practice of intelligence*, 2010,12:114-118.

[10] Wang Xiuhong, Yuan Yan, Zhao Zhicheng, Li Jieyu, Liu Haijun, Yang Guoli. Patent document structure tree model and its application in similarity computation.*ITA*, 2015,03:107-111.

[11] Wang Xin, Zhao Yunhua, Gao Fang. Research on measuring method of patent similarity based on classification number and citation. *2015 Digital Library Forum (1)*

[12] Fellbaum C. WordNet: An Electronic Lexical Database [M]. *MIT Press*, 1998.

[13] http://www. wikipedia. org/.

[14] Liu Hongzhe, Xu De. Ontology Based Semantic Similarity and Relatedness Measures Reviewed.

computer science, 2012,39 (2): 8-13.

[15] Rada R, Mili H, Bicknell E, et al. Development and application of a metric on semantic nets[J]. IEEE Transactions on Systems, Man and Cybernetics, 1989,19(1) : 17-30.

[16] Richardson R, Smeaton A F. Using WordNet in a Knowledge- Based Approach to Information Retrieval [Z]. Working Paper, CA-0395. School of Computer Applications, Dublin City University, Ireland, 1995.

[17] Wu Z, Palmer M. Verbs semantics and lexical selection[C]// Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics. Morristown, NJ, USA, 1994 : 133-138.

[18] Leacock C, Chodorow M. Combining local context and WordNet similarity for word sense identification[C]//Fellbaum C, ed. WordNet: An Electronic Lexical Database. MIT Press, 1998: 265-283.

[19] Kim J W, Candan K S. Cp/ev: concept similarity mining without frequency information from domain describing taxonomies[C]// Proceedings of the 15th ACM International Conference on Infor- mation and Knowledge Management. New York, NY, USA, ACM Press, 2006 : 483-492.

[20] Liu Hong-zhe, Bao Hong, Xu de. Concept Vector for Similarity Measurement based on Hierarchical Domain Structure. *Computing and informatics*,2011 (30) : 1001-1021.

[21] Lord P W, Stevens R D, Brass A, et al. Investigating Semantic Similarity Measures across the Gene Ontology: The Relationship Between Sequence and Annotation. *Bioinformatics*, 2003, 9 (10):1275-1283.

[22] Resnik Po Semantic similarity in a taxonomy: An information based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research,* 1999,11:95-130.

[23] Lin D. An information-theoretic definition of similarity [C]//Proceedings of the 15th International Conference on Machine Learning. Wisconsin, USA, July 1998: 296-304.

[24] Jiang J J, Conrath D W. Semantic similarity based on corpus statistics and lexical taxonomy[C]//Proceedings of the 10th International Conference of Research on Computational Linguistics. Taiwan, August 1997.

[25] Tversky. A. Features of Similarity. *Psychological Review*, 1977, 84(4): 327-352.

[26] Banerjee S, Pedersen T. Extended gloss overlaps as a measure of semantic relatedness[C]//Proceedings of IJCAI. Mexico 2003: 805-810.

[27] Patwardhan S, Pedersen T. UsingWordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts [C]// Proceedings of the EACL Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together. Trento, Italy, April 2006: 1-8.

[28] Wan S, Angryk R A. Measuring semantic similarity using word-net-based context vectors[C]//Systems, Man and Cybernetics. 2007:908-913.

[29] Li Y, Bandar Z, MeLean D. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering,* 2003, 15 (4):871-882.

[30] Marco A A, SeungJin L. A Graph Modeling of Semantic Similarity between Words[C]//International Conference on Semantic Computing (ICSC 2007). 2007:355-362.

[31] Jike G,Yuhui Q. Concept Similarity Matching Based on Semantic Distance[C]//SKG: 380-383.

[32] Anna F. Concept similarity by evaluating information contents and feature vectors: a combined approach. *Communications of the ACM,* 2009, 52(3): 145-149.

[33] Gerasimos S, Georgios S, Andreas $. A hybrid Web-based measure for computing semantic relatedness

between words [C]// 2009 21st IEEE International Conference on Tools with Artificial Intelligence, ICTAI. 2009:441-448.

[34] Zhao Zhong-cheng, Yan Jian-zhuo, Fang Li-ying, et al. Measuring Semantic Similarity Based On WordNet[C]//Web information system and application conference. 2009:89-92.

[35] Cai Song-mei, Lu Zhao. An Improved Semantic Similarity Measure for Word Pairs[C]//International Conference on e-Education, e-Business, e-Management and e-Learning. 2010:212-216.

[36] Qin Peng,Lu Zhao,Yan Yu,et al. A New Measure of Word Semantic Similarity based on WordNet Hierarchy and DAG Theory[C]// International Conference on Web Information Systems and Mining. 2009:181-185.

[37] Sheng Yan, Li Yun, Luan Luaru A Concept Similarity Method in Structural and Semantic Lev¥els[C]//Second International Symposium on Information Science and Engineering: 620-623.

[38] Shi Bin, Fang Li-ying, Yan Jian-zhuo, et al. Ontology-Based Measure of Semantic Similarity between Concepts [C]// World Congress on Software Engineering. 2009, 2: 109-112.

[39] Mihalcea, R., Corley, C. & Strapparava, C. (2006). Corpus based and knowledge-based measures of text semantic similarity. In Proceedings of the American Association for Artificial Intelligence. (Boston, MA).

[40] Strube M, Ponzetto S P. WikiRelate! Computing Semantic Relatedness Using Wikipedia[C]//Proc. of AAAI. 2006.

[41] Gabrilovich E, Markovitch S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis [C]//IJCAI. 2007:1606-1611.

[42] Milne D. Computing semantic relatedness using Wikipedia link structure[C]//NZCSRSC'07.

[43] HowNet [R]. HowNet's Home Page. http://www. Keenage.com.

[44] Liu Qun, Li Sujian. Based on the "HowNet" lexical semantic similarity calculation 2002, 7 (2): 59-76.

[45] Yu Gang, Pei Yangjun, Zhu Zhengyu. Research of text similarity based on word similarity computing. *computer engineering and design* .2006, 27 (2): 241-244

[46] Li Wenjie, Zhao Yan. Semantic Similarity between Concepts Algorithm Based on Ontology Structure. *Computer engineering*, 2010, 23:4-6.

[47] Yang Fangying, Jiang Zhengxiang, Zhang Shanshan. Semantic similarity Measurement based on ontology. *Computer technology and development,* 2013, 23 (7): 52-56.

[48] Yang Nana, Zhang Qingnian, Niu Jiqiang. Spatial entity semantic similarity computation model based on ontology structure. *Science of Surveying and mapping*, 2015, 03:107-111+84.

[49] Zhou Qunfang, Gu Jun. Study on similarity patents detection on ontology [A].2012

[50] S. Deerwester, S.T. Dumains, G.W. Furmas, Indexing by Latent Semantic Analysis, Journal of the ASIS, 1986-1998, September 1990.

[51] Landauer, T.K. & Dumais, S.T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge", Psychological Review, 104.

[52] Thomas Hofmann. Unsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning. 2001 (1-2).

[53] Blei,D M Ng A Y, Jordan M I. Latent Dirichlet allocation. Journal of Machine Learning Research. 2003.

[54] Wang Zhenzhen, He Ming, Du Yongping. Text similarity computing based on Model LDA. Computer science, 2013, 12:229-232.

[55] Sun Changnian, Zheng Cheng, Xia Qingsong. Chinese text similarity computing based on LDA.

Computer technology and development, 2013, 01:217-220.

[56] Po HU, Minlie HUANG*, Xiaoyan ZHU. Patent Key Component Extraction with the Application of Patent Similarity Analysis. *Journal of Computational Information Systems*.10: 13 (2014) 5813-5820

[57] G Cascini,M Zini. Measuring patent similarity by comparing inventions functional trees. Computer-aided Innovation, 2008

[58] Moehrle M G, Gerken J M. Measuring textual patent similarity on the basis of combined concepts: design decisions and their consequences. Scientometrics, 2012, 91(3): 805-826.