Acta Technica *62* (2017), No. 6B, 537–546

Text information semantic retrieval based on knowledge meta model and resource ontology

Wenguo Liao[1], Guangping Liao[2]

Abstract. Based on general vector space modal, this Thesis maps text, word, and concept in Google's academic sememe vector space; and based on GVSM, this Thesis express text as vector in Google's academic sememe space, and carries out calculation for test similarity through sememe similarity, thus providing a new thought for calculation of semantic similarity of text. According to comparison experiment, it shows that in semantic similarity calculation of text, the algorithm in this Thesis has been improved compared with VSM and GVSM. Test similarity has been calculated through sememe similarity; the algorithm effectiveness has been verified through experiments. A new thought for calculating text similarity is provided. During experiment process, it is found out that PVSM has certain superiority in calculating similarity of text sets without obvious characteristic items.

Key words. Knowledge element model, Ontology, Resource retrieval, Semantic space, Text clustering.

1. Introduction

Text similarity calculation is one core technology in multiple fields like information retrieval, text categorization, and machine translation. With further application and development of information technology, people have higher demand for processing capacity of text message; as basic technology for text processing, text similarity calculation has become a research hotspot for scholars.

Computing methods for text similarity are divided into two main categories: one is the method based on statistics, and the other is the method based on semantic comprehension. The most representative models in statistical methods are Vector

[1]Dean's Office, ABA Teachers University, Wenchuan County ,Sichuan Province, 623002, China
[2]Department of Humanities and Social Sciences, ABA Teachers University, Wenchuan County, Sichuan Province, 623002, China

Space Model (VSM for short) [1] and General Vector Space Model (GVSM for short) [2].

GVSM has improves the assumption that characteristic items of text in VSM are mutually orthogonal. Besides, JorgBeck[3] has proposed topic-based vector space model (TVSM for short) in which orthogonality is not needed for characteristic items and it can flexibly dispose similarity of characteristic items. Cheng Yuzhu and others[4] have proposed one kind of component frequency model (CFM for short) in which test is expressed as a space vector which takes components as characteristic items. Statistical-based methods have been widely used; however, most of these methods are based on statistical-based, thus lacking semantic information. Generally, semantic comprehension-based methods use certain knowledge base to calculate the similarity of words in text [5- 7], and then to further calculate the similarity of sentences and test. Yuan Xiaofeng[8-9] calculates the similarity of Google's academic sememe by introducing in depth and area density and then carries out the calculation for word similarity on this basis. Bai Qiuchan and others[10] use Google's academic sememe to express text as one concept vector; however, since there is no orthogonality relation among sememes, this model still has not solved the assumption of characteristic item orthogonality; besides, that use one sememe to represent words will cause the problem of losing semantic meaning; for example, in using sememe "software" to represent characteristic item "virus", the semantic information is semantic information lost.

2. Vector space modal

Vector space modal was proposed by Salton G in the 1970s; its idea is: take characteristic item of text as one vector in N-dimensional space; use characteristic item vector to represent text; and judge the similarity among texts through calculating the included angle of text vectors. In VSM, words are commonly chosen as characteristic items; each characteristic item is endowed with certain weight[11] based on its Term Frequency (TF) and Inverse Document Frequency (IDF) in the text. Text vector is expressed as follows:

$$\overrightarrow{a} = \sum_{i=1}^{m} w_{ai} \cdot \overrightarrow{t_i}\,, \tag{1}$$

$$\overrightarrow{b} = \sum_{i=1}^{m} w_{bi} \cdot \overrightarrow{t_i}\,, \tag{2}$$

$$W(t, \overrightarrow{d}) = \frac{tf(t, \overrightarrow{d}) \times \log(N/n_1 + 0.01)}{\sqrt{\sum\limits_{t \in d} [tf(t, \overrightarrow{d}) \times \log(N/n_1 + 0.01)]^2}}\,, \tag{3}$$

$\overrightarrow{t_i}$ refers to corresponding space vector of characteristic item t_i; $\overrightarrow{a}$ and $\overrightarrow{b}$ refer to text vectors; m is the number of characteristic items; w_{ai} and w_{bi} are weights

of characteristic items in the text; weights are determined by *TF-IDF* and normalization processing will be carried out for them. $W(t, \overrightarrow{d})$ represents the weight of characteristic item t in text $\overrightarrow{d}$; $tf(t, \overrightarrow{d})$ refers to the occurrence number of characteristic item t in text $\overrightarrow{d}$; N is the total number of texts; n_1 is the number of texts which contains t. Text similarity is expressed as follows:

$$sim(\overrightarrow{a}, \overrightarrow{b}) = \cos(\theta) = \frac{\overrightarrow{a} \cdot \overrightarrow{b}}{|\overrightarrow{a}| \cdot |\overrightarrow{b}|} = \frac{\sum_{j=1}^{m} w_{aj}\overrightarrow{t_j} \cdot \sum_{i=1}^{m} w_{bi}\overrightarrow{t_i}}{\sqrt{\sum_{j=1}^{m} w_{aj}\overrightarrow{t_j} \cdot \sum_{i=1}^{m} w_{ai}\overrightarrow{t_i}} \cdot \sqrt{\sum_{j=1}^{m} w_{bj}\overrightarrow{t_j} \cdot \sum_{i=1}^{m} w_{bi}\overrightarrow{t_i}}} . \tag{4}$$

In VSM, characteristic item $\overrightarrow{t_i}$ is assumed as orthometric unit vector; therefore, (4) can be simplified as:

$$sim(\overrightarrow{a}, \overrightarrow{b}) = \frac{\sum_{j=1}^{m} w_{aj} \cdot w_{bj}}{\sqrt{\sum_{j=1}^{m} w_{aj}^2}\sqrt{\sum_{i=1}^{m} w_{bi}^2}} = \sum_{j=1}^{m} w_{aj} \cdot w_{bj} . \tag{5}$$

In VSM the assumption that characteristic items are orthometric has not considered semantic similarity and correlation among words in natural language. General Vector Space Model (GVSM) has improved the assumption that characteristic items are orthometric. In GVSM, characteristic items are expressed as one set of independent non-orthometric unit vector in the space. Text space is a subspace generated by characteristic item vectors; characteristic item weight adopts the same method as VSM; the similarity of vectors in this Thesis is expressed as follows:

$$sim(\overrightarrow{a}, \overrightarrow{b}) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} w_{ai}w_{bj}\overrightarrow{t_i} \cdot \overrightarrow{t_j}}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m} w_{ai}w_{aj}\overrightarrow{t_i} \cdot \overrightarrow{t_j}}\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m} w_{bi}w_{bj}\overrightarrow{t_i} \cdot \overrightarrow{t_j}}} . \tag{6}$$

$$\overrightarrow{t_i} \cdot = |t_i| \cdot |t_j| \cdot \cos(\omega) = \cos(\omega) . \tag{7}$$

$\cos(\omega)$ represents the cosine value of included angle of two characteristic items, namely the similarity of characteristic items. The key of text similarity calculation is the calculation of characteristic item similarity. S.K.M. Wong and others[2] use the lowest term in Boolean algebra to express the co-occurrence of characteristic items with 2m minor term set and make one-to-one correspondence of the lowest term and base in a 2m-dimension space to map text vector and characteristic item vector

in 2m space represented by the lowest term; then calculate text similarity through included angle of vectors. Farahat and others[12] proposed to use covariance matrix of characteristic item of text to calculate characteristic item similarity based on GVSM model. Both these two kinds of calculation for characteristic item similarity use co-occurrence information of characteristic items in text set and they have small dependence on corpus, thus not easy to be expanded.

3. Google academics

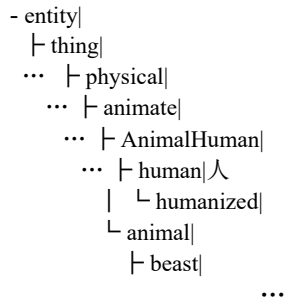

Fig. 1. Dendriform hierarchical structure of sememes

Google academics[13] is a knowledge system created by the famous machine translation expert Dong Zhengdong in our country with ten-year endeavor. It contains rich lexical semantic knowledge. There are two main concepts in Google academics: "concept" and "sememe". "Concept" refers to a description of word meaning. Each word can express several "concepts". "Concept" is described by a kind of "knowledge representation language" by which the words are used are called "sememe". "Sememe" is the smallest meaning unit for describing one "concept". Google academics tries to describe each "concept" by using a series of "sememes".

As basic unit for describing concept, sememes have complicated relations among each other. The 8 relations of sememes are described in Google academics; sememes form a complicated reticular structure. However, the most important relation among sememes is still the up and down position relation. According to the up and down position relation of sememes, all "basic sememes" form a dendriform hierarchical structure of sememes (see Fig.1) which is the basis for sememe similarity calculation.

4. Sememe vector space

In Google academics, sememe is the basic unit for representing concept; GVSM thought is used; sememes are regarded as one set of independent linear non-orthometric unit vectors. Sememe space is a subspace generated by sememe vectors. Sememe

vectors are expressed as follows:

$$\overrightarrow{p} = \sum_{i=1}^{n} \varepsilon_i \overrightarrow{j}_i \,. \tag{8}$$

Where, $\overrightarrow{j}$ represents base of N- dimension space; ε_i is the projection of sememe in $\overrightarrow{j}$ direction. Therefore, the inner product of vector of two sememe vectors is the cosine of their included angle, namely, similarity is:

$$\overrightarrow{p_m} \cdot \overrightarrow{p_n} = |p_m| \cdot |p_n| \cdot \cos(\theta) = \cos(\theta) \,. \tag{9}$$

4.1. Calculation of sememe similarity

The calculation of sememe similarity[8, 14–16] is based on the shortest distance and depth of sememes in hierarchical structure. The bigger the distance is, the smaller the sememe similarity will be; and the bigger the depth is, the greater the sememe information amount and similarity will be. Calculation equation for sememe similarity is as follows:

$$\mathrm{Sim}(p_1, p_2) = \frac{\alpha \times \min(depth_{p_1}, depth_{p2})}{\alpha \times \min(depth_{p_1}, depth_{p_2}) + dis(p_1, p_2)} \,. \tag{10}$$

dis refers to the shortest distance of $p1$ and $p2$ in hierarchical structure of sememes; *depth* refers to the depth of sememe in hierarchical structure; α refers to weight coefficient of depth.

4.2. Sememe vector of concept, word, and text

In Google academics, concept is described by one sememe set through one formalized language. Analyze through sememe description equation of concepts and divide them into three parts:

(1) Independent sememe description equation: use basic sememes or specific words to directly describe concepts;

(2) Relation sememe description equation: use "relation sememe = basic sememe" or "relation sememe = (specific word)" or "(relation sememe = specific word)" to describe;

(3) Symbol sememe description equation: use "relation symbol, basic sememe" or "relation symbol (specific word)" to describe concepts.

In independent sememe description equation, the first sememe description equation represents main meaning of concepts and it has rlatively high weight; therefore, independent sememe description equation is further divided into primary independent sememe and other independent sememes.

Sememes in this Thesis are space vectors; therefore, concepts are represented by linear combination of sememe vectors. According to the above division of concept component, concept vectors shall be formed by four parts of sememe vectors; and the weight of each part is determined by importance degree of sememe expression.

Concept vectors are expressed as follows:

$$\begin{aligned}\overrightarrow{S} &= \sum_{\mathrm{i}=1}^{4} \beta_i \cdot \overrightarrow{P_i} = \beta_1 \cdot \overrightarrow{p_{fi}} + \sum_{i=1}^{n} \beta_2 \cdot \overrightarrow{p_i} + \sum_{j=1}^{m} \beta_3 \cdot \overrightarrow{p_j} + \sum_{k=1}^{z} \beta_4 \cdot \overrightarrow{p_k} \\ &= \sum_{w=1}^{pn} \varepsilon_w \overrightarrow{p_w},\end{aligned} \tag{11}$$

And $\beta_4 \leq \beta_3 \leq \beta_2 \leq \beta_1$, $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$.

The definitions of $\overrightarrow{P_1}$ to $\overrightarrow{P_4}$ are as follows:

(1) $\overrightarrow{P_1}$ represents the sememe vector corresponding to the primary independent sememe description equation of the concept $\overrightarrow{S}$. As the primary independent sememe is expressed with the only sememe or word, $\overrightarrow{P_1}$ is expressed with the only sememe vector or word vector; β_1 is the weight coefficient of this part of sememe vector;

(2) $\overrightarrow{P_2}$ represents the sememe vector corresponding to the other independent sememe description equations of the concept $\overrightarrow{S}$. It is expressed with the linear addition of the sememe vectors of the other independent sememes except for the primary independent sememe or words; β_2 is the weight coefficient of this part of sememe vector;

(3) $\overrightarrow{P_3}$ represents the sememe vector corresponding to the relation sememe description equation of the concept $\overrightarrow{S}$. It is expressed with the linear addition of the sememe vectors of the relation sememe and independent sememe or words it describes; β_3 is the weight coefficient of this part of sememe vector;

(4) $\overrightarrow{P_4}$ represents the sememe vector corresponding to the symbol sememe description equation of the concept $\overrightarrow{S}$. It is expressed with the addition of vectors of the independent sememe or words the symbol describes; β_4 is the weight coefficient of this part of sememe vector;

$\overrightarrow{S}$ represents concept vector; β_i represents the weight coefficient of each part of sememe vector; pn represents the number of sememe vectors included in the concept; ε_w represents the coefficient of the corresponding sememe vector in the concept vector.

As words generally include not only one concept, the vector of words is expressed as the weighted average of the concept vector included, as follows:

$$\overrightarrow{W} = \sum_{j=1}^{snum} \frac{\overrightarrow{S_j}}{snum} = \sum_{j=1}^{snum} \sum_{w=1}^{pn_j} \frac{\varepsilon_{wj}}{snum} \overrightarrow{p_{wj}} = \sum_{w=1}^{wn} \varepsilon_w \overrightarrow{p_w}. \tag{12}$$

$\overrightarrow{W}$ represents word vector; $snum$ is the number of concepts included by the word $\overrightarrow{W}$; $\overrightarrow{S}_j$ represents one concept of $\overrightarrow{W}$; wn represents the number of sememes in the word vector. With the TF-IDF weight of feature item, the text vector is expressed

as follows:

$$\rightarrow D = \sum_{i=1}^{wnum} w_i \overrightarrow{W_i} = \sum_{i=1}^{wnum} w_i \sum_{\mathrm{j}=1}^{wn_i} \varepsilon_{ji} \overrightarrow{p_{ji}} = \sum_{w=1}^{dn} \varepsilon_w \overrightarrow{p_w} \,. \tag{13}$$

$wnum$ is the number of feature items in $\rightarrow D$; w_i is the weight of feature item $\overrightarrow{W_i}$ which can be calculated through Equation (3); dn is the number of sememes included in the text vector. With the text sememe vector, the similarity can be expressed as follows:

$$\begin{aligned} Sim(\overrightarrow{D_1}, \overrightarrow{D_2}) = \cos(\theta) &= \frac{\overrightarrow{D_1} \cdot \overrightarrow{D_2}}{|D_1||D_2|} \\ &= \frac{\sum_{i=1}^{dn_1} \varepsilon_{1i} \overrightarrow{p_i} \cdot \sum_{j=1}^{dn_2} \varepsilon_{2j} \overrightarrow{p_j}}{\sqrt{\sum_{\mathrm{i}=1}^{dn_1} \varepsilon_{1i} \overrightarrow{p_i} \cdot \sum_{j=1}^{dn_1} \varepsilon_{1j} \overrightarrow{p_j}} \sqrt{\sum_{i=1}^{dn_2} \varepsilon_{2i} \overrightarrow{p_i} \cdot \sum_{j=1}^{dn_2} \varepsilon_{2j} \overrightarrow{p_j}}} \\ &= \frac{\sum_{i=1}^{dn_1} \sum_{j=1}^{dn_2} \varepsilon_{1i} \varepsilon_{2j} \overrightarrow{p_i} \cdot \overrightarrow{p_j}}{\sqrt{\sum_{i=1}^{dn_1} \sum_{j=1}^{dn_1} \varepsilon_{1i} \varepsilon_{1j} \overrightarrow{p_i} \cdot \overrightarrow{p_j}} \sqrt{\sum_{i=1}^{dn_2} \sum_{j=1}^{dn_2} \varepsilon_{2i} \varepsilon_{2j} \overrightarrow{p_i} \cdot \overrightarrow{p_j}}} \,. \end{aligned} \tag{14}$$

5. Experiment

Proposed methods in this thesis such as PVSM, VSM and GVSM are compared via text clustering experiment. Chinese text classification corpus offered by Sougou Lab is selected for classification text, and one hundred and fifty texts of different lengths are selected from each kind of six kinds totally as experimental data. Fifty texts in each kind are selected as training texts for VSM and GVSM, and the other one hundred texts are considered as clustering texts. Distinction of three kinds of models is compared via using F-metric, which is a kind of balancing index combining precision ratio and recall ratio. If n_i is text number of type i, and n_j is text number of clustering j, and n_{ij} is text number in clustering j subordinating to type i, then precision ratio $p(i,j)$, recall ratio $r(i,j)$ and F-metric $F(i,j)$ can be respectively defined as:

$$p(i,j) = \frac{n_{ij}}{n_j} \,. \tag{15}$$

$$r(i,j) = \frac{n_{ij}}{n_i} \,. \tag{16}$$

$$F(i,j) = \frac{2 \times p(i,j) \times r(i,j)}{p(i,j) + r(i,j)} \,. \tag{17}$$

Calculation experiment on conception similarity in literature[15–17] can be re-

ferred to to determine experimental parameters α, β_i, and they can be adjusted in the experiment according to the effect. Specific setting is as follows $\alpha = 1.6$, $\beta 1 = 0.5$, $\beta 2 = 0.2$, $\beta 3 = 0.17$, $\beta 4 = 0.13$, and weight calculation formula of characteristic item is introduced as follows:

$$w_t = \frac{t_t}{t_i t_d}. \tag{18}$$

Then set of characteristic item can be selected. t_t presents total number of times of characteristic item in training text, and greater value of it indicates stronger text denoting capability; t_i presents number of characteristic item in training set type, and greater value of it indicates stronger text capability of characteristic item to denote this type; t_d presents number of text in characteristic item, and greater value of it indicates weaker capability of characteristic item to denote text type. Preceding N words are selected as characteristic item according to size of weight. Experimental result is shown in following figure via using clustering algorithm of K-means:

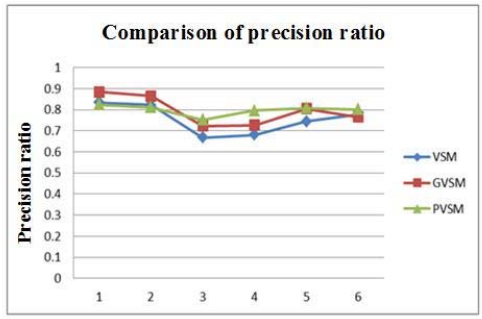

Fig. 2. Comparison Of Precision Ratio

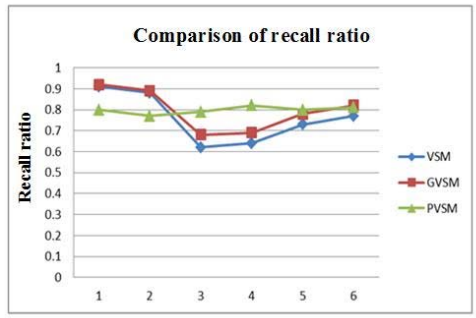

Fig. 3. Comparison Of Recall Ratio

Via comparison of recall ratio, precision ratio and F-metric for three algorithms, algorithms proposed in this thesis are all superior to VSM and GVSM in text set 3, 4, 5, 6, but they are inferior to VSM and GVSM in text sets 1 and 2. It is found via analysis that themes of text sets 1 and 2 are respectively military and tourism. In these two texts, there are obvious high-frequency words of characteristic item, such as "weapon" and "tourism" etc. For this kind of text set of obvious

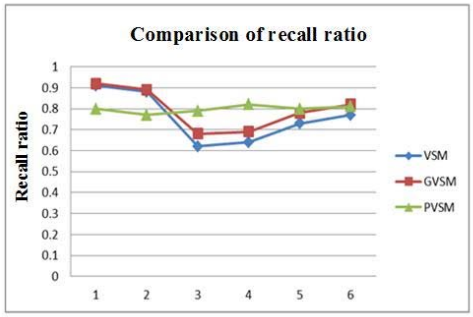

Fig. 4. F-Comparison chart of metric

characteristic item, VSM and GVSM are relatively suitable. Text sets 3 and 4 are respectively correspondent to topic, culture of novel and film, so content scope of texts is very extensive without obvious characteristic item, so PVSM algorithm is relatively effective this moment. While in text sets 5 and 6, effect of PVSM and GVSM are close on health care and technology, and they are superior to VSM, and there are some commonly used characteristic words in these two kinds of thesis, but they are not so obvious as text sets 1 and 2. To sum up, PVSM model based on Google's academic sememe vector space is superior to GVSM and VSM in the aspect of semantic similarity calculation, but it is inferior to GVSM and VSM in text set of obvious characteristic item. Words, conception, sememe are all known in Google's academic, so sememe similarity, sememe vector denotation of conception and words can be calculated in advance, and they can be directly used at the time of calculating text similarity without training corpus, so its application scope is more extensive compared with VSM and GVSM.

6. Conclusion

Problem of "polysemy" and "many words for a single meaning" in natural language cannot be handled by traditional spatial vector model, so knowledge base system of Google's academic is utilized in this thesis to improve general space vector model, and vector space model based on semantics is proposed. Text is denoted as a vector in vector space via TF-IDF weight of characteristic item in this model, and text similarity calculation is realized via sememe in knowledge base of Google's academic. To verify effectiveness of proposed algorithm, K-means clustering algorithm is adopted in this thesis, and similarity as clustering distance among texts is calculated using respectively VSM model, GVSM model and method proposed in this thesis, and experimental result indicates that the aspect of semantic similarity calculation can be improved to some degree by method proposed in this thesis compared with GVSM and VSM models.

Acknowledgement

The key scientific research project of ABA Teachers University Foundation of China under Grant No. ASA16-07; the research project of Sichuan Provincial Department of Education Foundation of China under Grant No. 17ZB0001.

References

[1] an S, Li M, Al-Sultany G, et al.: (2011) *Semantic based file retrieval on resource limited devices with ontology alignment support*[C]// Eighth International Conference on Fuzzy Systems and Knowledge Discovery. IEEE, 2011:2699-2702.

[2] Wu J, Wang H: (2008) *Study on Semantic Knowledge Retrieval based Context*[C]// IEEE International Symposium on Knowledge Acquisition and Modeling Workshop, 2008. Kam Workshop. IEEE, 2008:1006-1009.

[3] Zhai J, Li M, Sun P: (2011) *Knowledge Modeling and Semantic Retrieval for Sports Information Based on Ontology*[J]. Advanced Materials Research, 187:45-50.

[4] Zghal H B, Aufaure M A, Mustapha N B: (2007) *A model-driven approach of ontological components for on- line semantic web information retrieval*[J]. Journal of Web Engineering, 6(4):309-336.

[5] Xia C, Cheng X, Zhang L, et al.: (2010) *Ontology-based Semantic Information Retrieval*[C]// World Automation Congress. 2010:183-187.

[6] Fadzli S A, Setchi R: (2010) *Semantic Approach to Image Retrieval Using Statistical Models Based on a Lexical Ontology.*[C]// Knowledge-Based and Intelligent Information and Engineering Systems -, International Conference, Kes 2010, Cardiff, Uk, September 8-10, 2010, Proceedings. DBLP, 2010:240-250.

[7] Wang S L, Zhang G J: (2013) *Ontology Based Domain Resource Semantic Retrieval Model*[J]. Applied Mechanics & Materials, 347-350:2804-2808.

[8] Tang L, Chen X: (2015) *Ontology-Based Semantic Retrieval for Education Management Systems*[J]. Journal of Computing & Information Technology, , 23(3):255.

[9] Wang W L, Huang M, Wang Y: (2014) *Construction of XBRL Semantic Metamodel and Knowledge Base Based on Ontology*[J]. Applied Mechanics & Materials, 571-572:1119-1128.

[10] Zhai J, Liang Y, Jiang J, et al.: (2008) *Fuzzy Ontology Models Based on Fuzzy Linguistic Variable for Knowledge Management and Information Retrieval*[C]// Intelligent Information Processing Iv, Ifip International Conference on Intelligent Information Processing, October 19-22, 2008, Beijing, China. DBLP, 2008:58-67.

[11] Zhai J, Yu Y, Liang Y, et al.: (2008) *Traffic Information Retrieval Based on Fuzzy Ontology and RDF on the Semantic Web*[C]// International Symposium on Intelligent Information Technology Application. IEEE Xplore, 2008:779-784.

[12] Teodorescu R, Racoceanu D, Leow W K, et al.: (2008) *Prospective Study for Semantic Inter-Media Fusion in Content-Based Medical Image Retrieval*[J]. Medical Imaging Technology, 26(1):48-58.

[13] Zhou Y, Tian B, Zeng Z Y, et al.: (2014) *Semantic Retrieval for Ontology-Based Aircraft Fault Knowledge*[J]. Advanced Materials Research, 945-949:3410-3417.

Received May 7, 2017