Send Orders for Reprints to reprints@benthamscience.ae

Similarity Measure of Test Questions Based on Ontology and VSM

Jing Yu, Dongmei Li*, Jiajia Hou, Ying Liu and Zhaoying Yang

School of Information Science and Technology, Beijing Forestry University, Beijing, 100083, China

Abstract: Vector space model (VSM) is a common method for measuring test questions similarity in massive item bank system. VSM is limited in accurately representing the knowledge relationship and the potential semantic relations of different characteristic words, hence this paper proposes a method of test questions similarity measure called OVSM-TQSM which combines domain ontology and VSM. OVSM-TQSM can reveal the intrinsic relationship among words by using the constructed domain ontology which integrates with the tree structure and the graphics structure. Incorporated with eigenvectors and the weight of words in VSM, OVSM-TQSM calculates the similarity of test questions. A large number of experimental results demonstrate that the novel approach is feasible and effective. Compared with the traditional method based on VSM, OVSM-TQSM has the advantages of higher accuracy and little unnecessary laborious pre-processing.

Keywords: Domain Ontology, Eigen Word, Massive Item Bank, Similarity Measure, Test Questions, VSM.

1. INTRODUCTION

VSM is a common method for measuring test questions similarity in massive item bank system [1-3], proposed by Salton etc. in 1970s [4]. It is a relatively old algorithm which was used in measuring text similarity, and it achieves good results for documents and web pages. Though this algorithm is easy to be applied, it ignores the relations among words in documents and only uses word frequency to calculate the similarity. Thus, when word frequency is low in a shorter passage, this method is inappropriate. Therefore, Chunxia Jin introduced a new method which uses a dynamic vector calculation in short passage to measure the similarity [5]. This method constructs dynamic text vector based on HOWNET related words corpus firstly, and abstracts HOWNET to a tree structure for further calculation. There are several research works on such algorithms using tree structure to solve the similarity calculation problem [6-8].

Exam question is a kind of short passage with stronger knowledge ontology.The researchabout exam question similarity calculation was originally conducted by Junyi Zhu in the Internet-based massive question item bank [9]. With the extensive application of massive item bank, increasingly importance has been attached to the exam question similarity calculation [10-12]. At present, every item bank cannot be shared publicly because of some specific information it contained, resulting in the surplus of questions in the item bank and less effectiveness of making exam papers. Therefore, the similarity between the questions plays a very important role in eliminating the surplus questions in item bank.

Similarly, tree structure has been introduced into many research about exam question similarity calculation. In the paper by Tang and Fan [10], high-frequency words extracting algorithm based on suffix tree isused to extract content features of exam questions. Combined with metadata features of questions, a method to compute question similarity is proposed. In the calculation of word similarity in exam questions, however, the examining points are not supposed to be a tree structure merely; instead, a graph structure is supposed to be an appropriate and comprehensive structure. On the other hand, it seems that two questions based on different points are similar at first glance, but in fact, one question can be quite different from the other. In this case, these two questions cannot be defined as similar questions which cannot be identified by VSM and tree structure. Considering this special characteristic of exam questions, we introduce ontology to the calculation of similarity. An ontology is an explicit specification of a conceptualization [13] which has been applied widely into the field of problems about similarity [14, 15].

In this paper, we propose an ontology and vector space models based test questions similarity measure (OVSM-TQSM). Experiments show that this method improves the accuracy of question similarity with less pre-processing.

2. DEFINITIONS ABOUT DOMAIN ONTOLOGY

By studying the characteristic of exam questions, we construct a domain ontology consisting of domain points. In this process, we add the graph structure into the original tree structure for the relations among different points. When the domain is described as a graph, every point is regarded as a node, and Fig. (**1**) is an acknowledge network obtained by analyzing a domain ontology.

1874-4443/14 2014 Bentham Open

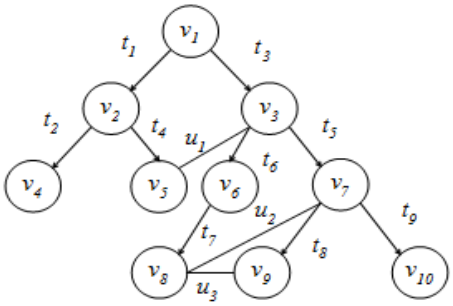

Fig. (1). Knowledge network G.

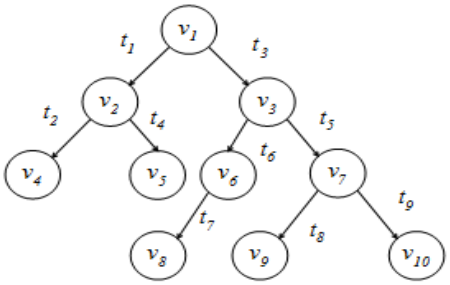

Fig. (2). Knowledge tree T.

Definition 1.Knowledge network. Knowledge network is characterized by a three-tuple set, denoted as G = (V, TE, UE), where V is the finite nonempty set of points, consisting all the nodes; TE is a finite set of parent-child pairs, consisting of directed arcs; UE is a finite set of non-parent-child pairs, consisting of undirected arcs.

In Fig. (**1**), V = {v1,…,v10}, TE={t1,….,t9}, UE = {u1, u2,u3}.

Obviously, the knowledge network has the same characteristics to the ontology introduced in [16]; namely, the knowledge network can reflect the upper and lower relations between each pair of nodes.

Definition 2. Knowledge tree. Knowledge tree is a two-tuple set from two sets in the knowledge network V and TE, denoted as T = (V, TE).

In Fig. (**2**), V={v1,…,v10}, TE={t1,….,t9}

Definition 3. Ancestor knowledge. Ancestor knowledge PVi is the set of ancestors of the node vi in the knowledge network.

In Fig. (**1**), the ancestor knowledge of the node v9 is PV9(v1, v3, v7, v9).

Definition 4. Related knowledge. Related knowledge RVi is the set of nodes connected with vi *via* undirected arc uen.

In Fig. (**1**), the related knowledge of v9 is PV9(v8).

In [7], a similarity calculation method based on tree structure is proposed. It uses the level of sememe, obatins the similarity of sememe by calculating the distance of paths, and takes node depth into consideration. We improve this method for the domain characteristic of exam questions. Therefore, the definition of concept analyzing is given as follows.

Definition 5. Concept analyzing. Concept analyzing is a process where the ancestor knowledge PVi and the related knowledge RVi of a certain node vi are united as a union set, namely, PVi∪RVi, denoted as CVi. The node vi is regarded as a word, and all the related nodes will be defined as sememe, then concept is the sememe of word.

In Fig. (**1**), the concept analyzing of v9 is v9 (v1,v3,v7,v8,v9), and that of v10is v10(v1,v3,v7,v10). According to the method in [7], the similarity between the two nodes is high because of the same ancestor. However, since they belong to different knowledge, they cannot be compared together as to the huge similarity. But if the method described in definition 5 is applied, it is more proper to set v9 (v1,v3,v7,v8,v9) and v10 (v1,v3,v7,v10).

3. THE PROPOSED METHOD OF SIMILARITY

3.1. Model of the Exam Questions

If we denote a word *i* as a vector *xi* or *yi*, then an exam question can be denoted as:

$$S\left[x_1,x_2,x_3,\ldots,x_n\right] \tag{1}$$

The similarity of two questions $S_1\left[x_1,x_2,x_3,\ldots,x_n\right]$ and $S_2\left[y_1,y_2,y_3,\ldots,y_m\right]$ is denoted as *sim(S1, S2)*. In this paper, we use ontology to measure the word similarity in VSM, namely, to compare the similarity of *xi*and *yj*, denoted as *sim(xi, yj)*. According to the definition of concept analyzing, every word in the ontology word corpus can derive more concepts from domain ontology by analysis, obtaining word (sememe1, sememe2, ..., sememe n).

If the vector g can be denoted as a concept, the vector *x*derived from the concept can be denoted as:

$$x\left[g_1,g_2,\ldots,g_n\right] \tag{2}$$

and the matrix model constructed can be denoted as

$$S=\begin{bmatrix} g_{11} & \cdots & g_{m1} \\ \vdots & \ddots & \vdots \\ g_{1n} & \cdots & g_{mn} \end{bmatrix} \tag{3}$$

3.2. Procedures of Calculating

OVSM-TQSM gets some certain exam questions from the item bank, segments, and then calculates the similarity.The method mainly includes the following two steps.

Step 1. Measuring the word similarity by using domain ontology.

Step 2. Measuring the test questions similarity based on VSM by using the two words which have the weighted maximal similarity as the eigenvector.

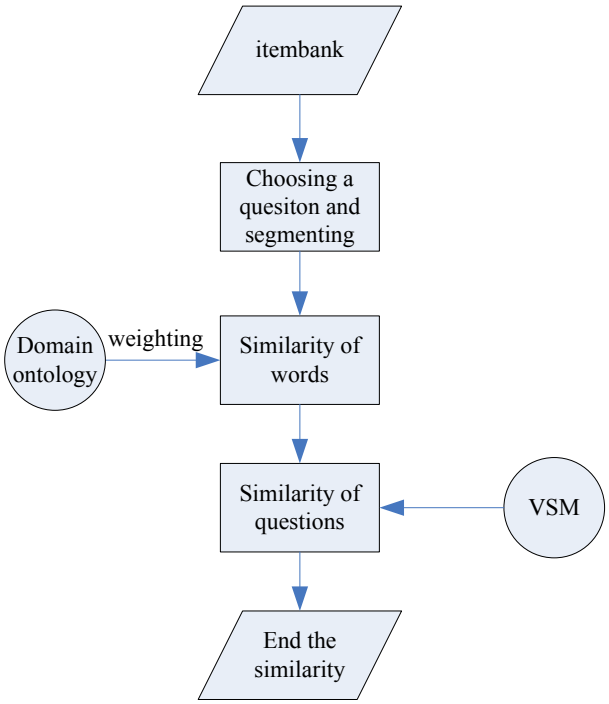

Fig. (3). The flowchart.

The flowchart is in Fig. (**3**). Comparedwith the traditional method, OVSM-TQSM has two advantages:

① It combines tree structure and graph structure, which strengthens the relations or difference between the words and improves the accuracy of the similarity. It is appropriate to construct a graph structure, because a vector of a question belongs to a certain subject with relations to others, and this is a many-to-many relation. But tree structure is useful, therefore we combine tree and graph structure to achieve more accurate results.

② It weights the word in the domain and eliminates stopwords, requiring less pre-processing. In order to compare effectively the domain words with common words, it is useful to enlarge the weights because of the strong domain character of exam questions. OVSM-TQSM modifies the traditional method of word matching and integrates the intrinsic relations based on frequency.

3.3. Ontology-based Similarity

As to two questions $S_1[x_1,x_2,x_3,\ldots,x_n]$ and $S_2[y_1,y_2,y_3,\ldots,y_m]$, we can get the eigenvector $x_i[g_1,g_2,\ldots,g_n]$and $y_j[h_1,h_2,\ldots,h_m]$ (i∈1,…, n, j∈{1,…,m}) after segmenting, where sememe*g* and *h* are obtained from concept analyzing. For convenience, we use $sim(x_i,y_j)$ to represent the similarity between *xi* and *yj*. The calculation is described as follows:

$$\begin{cases} sim(x_i,y_j)=1 & when\ x_i=y_j \\ sim(x_i,y_j)=\dfrac{f}{m+n-f} & when\ x_i\neq y_j \end{cases} \quad (4)$$

where*f*is the number of the same sememe, *m* and *n* are the numbers of sememe of the word *x* and the word *y* respectively.

3.4. VSM-based Similarity

According to the traditional VSM method, the distribution of the word k in a quesiton $IDK_k = lg(N/n_k)$ should be calculated firstly, where *N* is the number of the eigen words in the question and *nk* is the number of the eigen word *k*. Usually, the frequency of eigen word and that of non-eigen word in a question has little difference, so this traditional method is inappropriate to be applied into the exam questions. Therefore, when the domain ontology is being constructed, we weight the eigen word higher, and use the word similarity in 3.3 for further calculation.

Step 1. Calculating the weight of every word. The adjustment factor is γ_1 if the word k belongs to the ontology O, and γ_2 otherwise. The weight wk of the word *k* in question S1 and S2 is:

$$w_k=\begin{cases} \dfrac{q\gamma_1}{m\gamma_1+n\gamma_2}, & k\in O \\ \dfrac{q\gamma_2}{m\gamma_1+n\gamma_2}, & k\notin O \end{cases} \quad (5)$$

where *q* is the number of the word *k*, *m* and n are the number of the words whose frequencies are γ_1 and $\gamma_2$respectively. Here, $\gamma_1+\gamma_2=1$ and $\gamma_1>\gamma_2$.

Step 2. Calculating the weighted similarity β_k. Obtaining the maximum similarity $sim(x_1,y_1)$from 3.3, we get:

$$\beta_1=sim(x_1,y_1)w_{x1}w_{y1} \quad (6)$$

and then we eliminate the word x_1 and y_1; From the remaining similarities, we get the maximum $sim(x_2,y_2)$, get:

$$\beta_2=sim(x_2,y_2)w_{x2}w_{y2} \quad (7)$$

and eliminate the word x_2 and y_2. Repeat such steps until all the eigen words in a question have been completely extracted.

Step 3. The similarity of question S1 and S2 is:

$$sim(S_1,S_2)=\frac{\sum_{k=1}^{l}\beta_k}{\sqrt{\sum_{k=1}^{m}w_{S_1k}^2\sum_{k=1}^{m}w_{S_2k}^2}} \quad (8)$$

where*m* and *n* are the number of eigen words in S1 and S2 respectively, and l is the less one between *m* and *n*.

3.5. ApplicationInstance

We give an example in Fig. (**4**) to illustrate the execution process of the algorithm.

1）Giving a graph with 5 vertices and 12 edges, if adjacency matrix is used as storage structure, how much is space complexity?
2）Giving a graph with 3 vertices and 5 edges, if adjacency listis used as storage structure, how much is space complexity?
2）3 个顶点 5 条边的图，若使用邻接表存储，则空间复杂度是多少?

Fig. (4). An application instance.

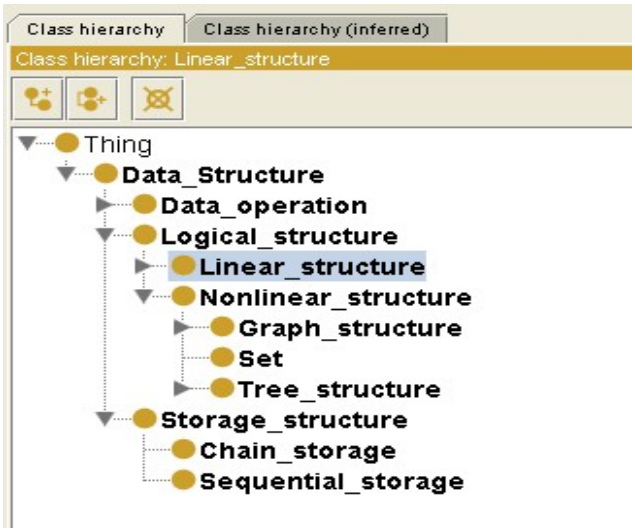

Fig. (5). The class of ontology.

Step 1. Word segmentation.

Step 2. Calculating the similarity between each word by usingformula 4, such as the similarity of "vertices " and " vertices " is 1,and the similarity of "adjacency matrix" and "adjacency list" is 0.67.

Step 3. Calculating the weight of the wordby using formula 5.

Step 4. Calculating the word similarity with weights by using formula 6.

Step 5.Obtaining similarity measure of test questions by using formula 9.

4. EXPERIMENTS

4.1. Experiment Settings

Taking the course Data Structure for an example to measure the effectiveness of our new method, we construct the domain ontology of Data Structure, illustrated in Fig. (**5**).

The dataset we use is from our own massiveitem bank of the course. After similarity measure of test questions is calculated by using OVSM-TQSM, we delete the redundant questions.Now the item bankis serving our teaching with high-quality test qusestions.

In this item bank, we import almost all of the questions in Analyzing Algorithm and Data Structure Graduate Test (The second edition) published by China Machinery Press, written by Shoukong Chen etc, consisting of 318 multi-choice questions, 335 fill-blank questions, 232 judgement questions,

1）There are ( ) nodes in a k-depth complete binary tree at least.
2）Here are ( ) nodes in a k-depth complete binary tree at least.
a) 2^k b) 2^k-1-1 c) 2^k-1 d) 2^k+1

Fig. (6). An example of question.

Fig. (7). The rate of error of(γ_1, γ_2).

450 application questions and 226 algorithm designing questions, 1561 totally. We measure the similarity in the interval of [0, 1], and the closer to 1, the higher similarity.

There are three different situations in our experiment:

① The questions withdifferent points from the same ancestor, as v4 and v6 from the same ancestor v2 in Fig. (**6**).

② The questions with the same points and different descriptions. In Fig. (**6**), questions 1 is a fill-blank question while the question 2 is a multi-choice question. But in fact they have the identical descriptions.

③ The questions with irrelevant points.When measuring the effectiveness, we use VSM-based method to compare.

4.2. Experimental Results

We get different results when using different values of γ_1 and γ_2 in formula 5. Fig. (**7**) is the result of experiments on these values. From Fig. (**7**), we choose (0.2, 0.8) as the best of (γ_1, γ_2)and conduct the subsequent experiments.

In order to prove the advantages of OVSM-TQSM, we separate the 1561 questions into three situations. We choose six groups to compare with other methods like traditional VSM and human judgements, shown in Fig. (**8**).

We discuss the reuslts as follows. In Fig. (**8**), the comparisons in the first two groups are in the situation ① described in the section 4.1, where the first group includes the questions with less characters and lower similarities, whereas the second group includes the questions with more character and higher similarities. In accordance to the first

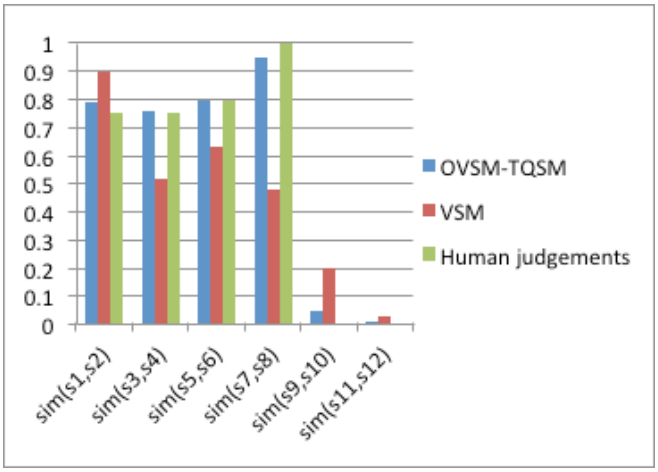

Fig. (8). A comparison.

Table 1. Experimental results.

Method	Number of Questions	Number of Tolerable Accuracy	Accuracy
OVSM-TQSM	1561	1426	91.4%
VSM	1561	1164	74.6%

group, it is obvious that OVSM-TQSM can enlarge the difference between two questions with less characters; the second group shows that OVSM-TQSM can easily find out the similarities with more characters.

The middle two groups are in the situation ② where the questions have higher similarities. In the group 4 especially, the questions are expressed in the same way but the numbers in the questions are different. OVSM-TQSM can make the similarity closer to similarity 1.

The last two groups are in the situation ③ where the questions are irrelevant to each other, namely, the similarities by human judgements are 0. The similarities gained by OVSM-TQSM are obviously less than that gained by VSM.

Making the advantages of OVSM-TQSM clearer, we compare the OVSM-TQSM and traditional VSM to human judgements separately. If the bias between the similarity calculated and human judgements is less than 5%, we regard it tolerable, and define the accuracy as:

$$accuracy = \frac{number\ of\ tolerable\ result}{number\ of\ the\ total\ questions} \quad (9)$$

Then the result is shown in Table **1**.

Analysing the result, we conclude that traditional VSM is not a proper method because the frequencies have little difference when there are fewer words in the questions. OVSM-TQSM compares every word, and weights the words in ontology higher, which can be effectively applied and achieve higher accuracy than VSM.

CONCLUSION

To mitigate the deficiency of traditional VSM, we propose a new method called OVSM-TQSM based on ontology and VSM to calculate the similarities between the exam questions. This method firstly constructs an ontology of a certain course, considering the ancestors in a knowledge tree and nodes with special relations in the knowledge network. Then it combines the thoughts of eigenvector and weighted words in VSM to calculate the similarity. The experiments show that OVSM-TQSM uncovers the intrinsic relations between words, reduces much pre-processing and achieves higher accuracy.

In the future, two researches can be conducted further. One is to find out more precise adjustment factors; the other is to expand this work to other fields where the sentence similarity is applied.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work is supported by the Fundamental Research Funds for the Central Universities (No.TD2014-02), Beijing Higher Education Reform (Computer Application Professional Core Courses Reform and Teaching Resources Construction Aiming at Improving Programming Design and Software Development Ability), Beijing Forestry University Resources Sharing Course (Data Structure), Beijing Forestry University Special Research of Campus Informatization (Construction of the Teaching Resources Sharing Platform for Program Design and Algorithms Courses).

REFERENCES

[1] H. Hage, and E. Aïmeur, "ICE:A System for Identification of conflicts in exams", *AICCSA*, pp. 980-987,2006.

[2] A. Tsinakos, and I. Kazanidis, "Identification of conflicting questions in the pares system", *The International Review of Research in Open and Distance Learning*, vol. 13, no. 3, pp. 297-313,2012.

[3] B. Qin, T. Liu, Y. Wang, S. F. Zheng, and S. Li, "Question answering system based on frequently asked Questions", *Journal of Harbin Institute of Technology*, vol. 35, no. 10, pp. 1179-1182, 2003.

[4] S. Galton, and C. Buckley, "Term-weighting approaches in automatic text retrieval", *Information Processing & Management*, vol. 24, no. 5, pp. 513-523, 1988.

[5] C. X. Jin, and H. Y. Zhou, "Chinese short text clustering based on dynamic vector", *Computer Engineering and Applications*, vol. 47, no. 33, pp.156-158, 2011.

[6] G. Wang, and G. X. Zhong, "Study on text clustering algorithm based on similarity measurement of ontology", *Computer Science*, vol. 37, no. 9, pp. 222-224, 2010.

[7] F. Li, and F. Li, "An new approach measuring semantic similarity in hownet 2000", *Journal of Chinese Information Processing*, vol. 03, no. 3, pp. 99-105, 2007.

[8] M. Batet, D. Sánchez, and A. Valls, "An ontology-based measure to compute semantic similarity in biomedicine", *Journal of Biomedical Informatics,* vol. 44, no. 1, pp. 118-125, 2011.

[9] J. Y. Zhu, "Consistency and integrity analysis for an intelligent item bank system on computer networks", *National Chinan University*, 1998.

[10] S. P. Tang, and X. Z. Fan, "Itembank redundancy checking based on multi-instance learning", *Transactions of Beijing Institute of Technology*, vol. 25, no. 12, pp. 1071-1074, 2005.

[11] J. W. Xiao, "Semantic analysis of redundancy and consistency for an intelligent network-based testing bank system", *National Chinan University*, 2000.

[12] Y. Y. Wang, Z. Chen, and X. H. Su, "Question similarity identification in automatic generation of test papers", *Journal of Harbin Institute Of Technology*, vol. 41, no. 1, pp. 1179-1182, 2009.

[13] T. R. Gruber, "A translation approach to portable ontology specifications", *Knowledge Acquisition*, vol. 5, no. 2, pp. 199-220, 1993.

[14] M. Batet, D. Sánchez, and A.Valls, "An ontology-based measure to compute semantic similarity in biomedicine", *Journal of biomedical informatics*, vol. 44, no. 1, pp. 118-125, 2011.

[15] K. Saruladha, G. Aghila, and S. Raj. "A survey of semantic similarity methods for ontology based information retrieval", Proceedings *of the 2nd International Conference on Machine Learning and Computing(ICMLC)*, pp. 297-301, 2010.

[16] W. N. Hao, B. Feng, G. Chen, D. W. Jing, and S. N. Zhao, "Document vector space model construction based on domain ontology", *Application Research of Computers*, vol. 3, no. 30, pp. 764-767, 2013.

Received: September 22, 2014 Revised: November 04, 2014 Accepted: November 06, 2014