TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol. 12, No. 9, September 2014, pp. 6932 ~ 6939
DOI: 10.11591/telkomnika.v12i9.4554

Applying Ontology and VSM for Similarity Measure of Test Questions

Jing Yu, Dongmei Li*, Shudong Hao, Jiajia Hou, Jianxin Wang
School of Information Science and Technology, Beijing Forestry University,
Beijing, 100083 China
*Corresponding author, e-mail: lidongmei@bjfu.edu.cn

Abstract

Vector space model (VSM) is a common method for measuring test questions similarity in large-scale 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. Comparing with the traditional method based on VSM, OVSM-TQSM has the advantages of higher accuracy and little unnecessary laborious pre-processing.

Keywords*: domain ontology, VSM, test questions similarity, large-scale item bank*

1. Introduction

VSM is a common method for measuring test questions similarity in large-scale 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 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 research about exam question similarity calculation was originally conducted by Junyi Zhu in the Internet-based smart question item bank [9], which has been the important field in smart item bank [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 is used 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

Received September 29, 2013; Revised June 27, 2014; Accepted July 25, 2014

specification of a conceptualization [13] which has been applied widely into the field of problems about similarity [14-16].

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 Figure 1 is acknowledge network obtained by analyzing a domain ontology.

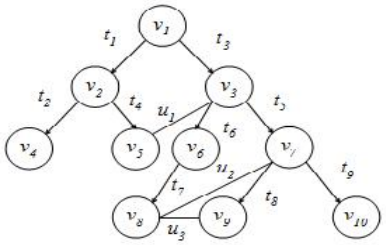

Figure 1. Knowledge Network G

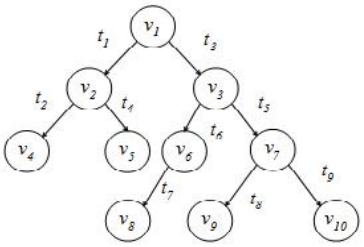

Figure 2. Knowledge Tree T

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 the Figure 1, V = {v1,…,v10}, TE={t1,….,t9}, UE={u1,u2,u3}.

Obviously, the knowledge network has the same characteristics to the ontology introduced in [17]; 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 the Figure 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 the Figure 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 the Figure 1, the related knowledge of v9 is PV9(v8).

In the paper [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 the Figure 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:

$$S[x_1, x_2, x_3, \ldots, x_n] \quad (1)$$

The similarity of two questions $S_1[x_1, x_2, x_3, \ldots, x_n]$ and $S_2[y_1, y_2, y_3, \ldots, y_m]$ 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:

$$x[g_1, g_2, \ldots, g_n] \quad (2)$$

And the matrix model constructed can be denoted as:

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

3.2. Procedures of Calculating

OVSM-TQSM gets some certain exam questions from the item bank, segments, and then calculates the similarity. Firstly, it measures the word similarity using domain ontology; then it weights words with the highest similarity, and uses them as the eigenvector; finally it calculates the exam questions similarity based on VSM. The flowchart is in the Figure 3.

Comparing to 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 combines 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.

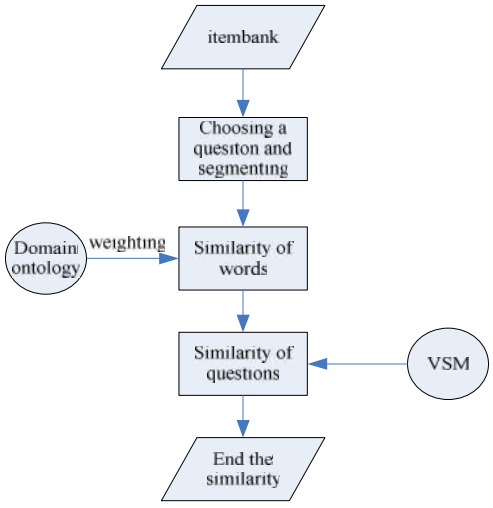

Figure 3. The Flowchart

3.3. Similarity Based on Ontology

As to two questions $S_1[x_1, x_2, x_3, \dots, x_n]$ and $S_2[y_1, y_2, y_3, \dots, y_m]$, we can get the eigenvector $x_i[g_1, g_2, \dots, g_n]$ and $y_j[h_1, h_2, \dots, h_m](i \in \{1, \dots, n\}, j \in \{1, \dots, m\})$ 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:

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

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. Similarity Based on VSM

According to the traditional VSM method, the distribution of the word k in a quesiton $IDK_k = lg\left(\frac{N}{n_k}\right)$ 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:

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

Where *q* is the number of the word *k*, *m* and n are the number of the words whose frequencies are $\gamma 1$ and γ_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:

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

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*.

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 the Figure 4 and Figure 5.

The dataset we use is from our own item bank of the course. 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, 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 with different points from the same ancestor, as v4 and v6 from the same ancestor v2 in the Figure 6.

(2) The questions with the same points and different descriptions. In the Figure 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.

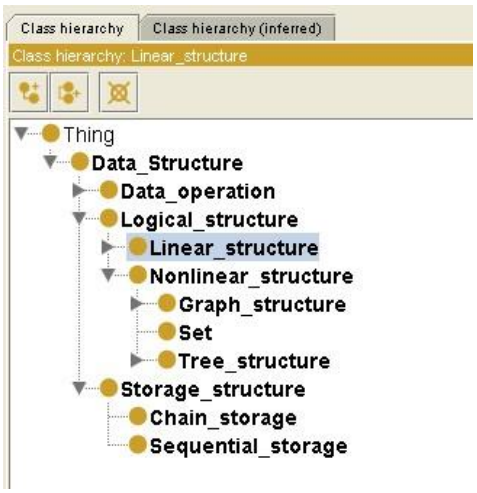

Figure 4. The Class of Ontology

Figure 5. The Adjacent List Related Content and Class

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

Figure 6. An Example of Question

4.2. Experimental Results

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

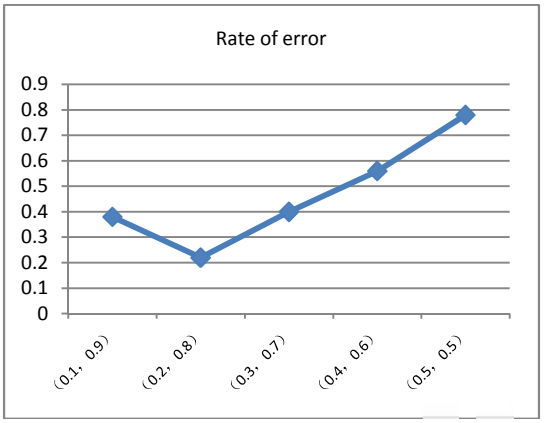

Figure 7. The Rate of Error of (γ_1, γ_2)

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 the Figure 8.

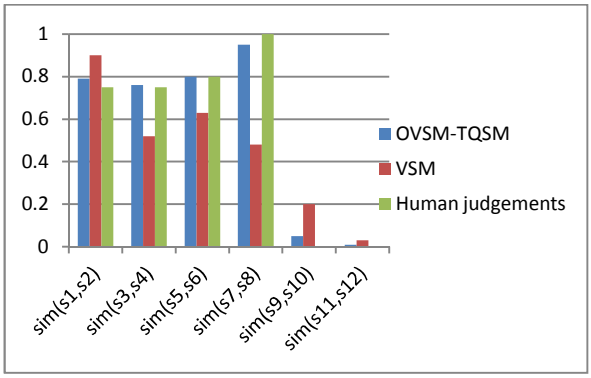

Figure 8. A Comparison

We discuss the reuslts as follows. In the Figure 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 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 the Table 1.

Table 1. Experimental Results

Method	Number of questions	Number of tolerable accuracy	Accuracy
OVSM-TQSM	1561	1426	91.4%
VSM	1561	1164	74.6%

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.

5. 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.

Acknowledgements

This work is supported by the National College Students' Training Programs of Innovation and Entrepreneurship (No. 201310022051) and the National Nature Science Foundation of China (No. 61170628).

References

[1] Hage H, Aïmeur E. ICE: A System for Identification of Conflicts in Exams. AICCSA. 2006: 980-987.
[2] Tsinakos A, Kazanidis I. Identification of conflicting questions in the PARES system. *The International Review of Research in Open and Distance Learning*. 2012; 13(3): 297-313.
[3] Qin B, Liu T, Wang Y, Zheng SF, Li S. Question answering system based on frequently asked questions. *Journal of Harbin Institute of Technology.* 2003; 35(10): 1179-1182.
[4] S Galton, CBuckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*. 1988; 24(5): 513–523.
[5] Jin CX, Zhou HY. Chinese short text clustering based on dynamic vector. *Computer Engineering and Applications*. 2011; 47(33): 156-158.
[6] Wang G, Zhong GX. Study on text clustering algorithm based on similarity measurement of ontology. *Computer Science.* 2010; 37(9): 222-224.
[7] Li F, Li F. An new approach measuring semantic similarity in Hownet 2000. *Journal of Chinese Information Processing*. 2007; 03(3): 99-105.
[8] Ning YH, Fang XH, Wu Y. Short text classification based on domain word ontology. *Computer Science*. 2009; 36(3): 142-145.
[9] Zhu JY. Redundancy、consistency and integrity analysis for an intelligent item bank system on computer networks. National ChiNan University. 1998.
[10] Tang SP, Fan XZ. Itembank redundancy checking based on multi-instance learning. *Transactions of Beijing Institute of Technology*. 2005; 25(12): 1071-1074.
[11] Xiao JW. Semantic analysis of redundancy and consistency for an intelligent network-based testing bank system. National ChiNan University. 2000.
[12] Wang YY, Chen Z, Su XH. Question similarity identification in automatic generation of test papers. *Journal of Harbin Institute of Technology*. 2009; 41(1): 1179-1182.
[13] Gruber TR. A Translation Approach to Portable Ontology. *Specifications. Knowledge Acquisition.* 1993; 5(2): 199-220.

[14] Batet M, Sánchez D, Valls A. An ontology-based measure to compute semantic similarity in biomedicine. *Journal of biomedical informatics*. 2011; 44(1): 118-125.
[15] Varelas G, Voutsakis E, Raftopoulou P, et al. *Semantic similarity methods in wordNet and their application to information retrieval on the web*. Proceedings of the 7th annual ACM international workshop on Web information and data management. ACM. 2005; 10-16.
[16] Chen J, Cai Y, Liu Y, et al. Multi-prototype based semantic similarity of concepts in ontology. ICMLC. 2012; 107-111.
[17] Hao WN, Feng B, Chen G, Jing DW, Zhao SN. Document vector space model construction based on domain ontology. *Application Research of Computers*. 2013; 3(30): 764-767.