International Journal of Computer Science and Applications,
©Technomathematics Research Foundation
Vol. 7, No. 1, pp. 79 – 93, 2010

ONTOLOGY BASED APPROACH FOR FINDING THE SIMILARITY AMONG QUESTIONS IN A DISTRIBUTED QUESTION BANK SCENARIO

REKHA RAMESH
Computer Department, Shah & Anchor Kutchhi Engineering College,
Chembur, Mumbai-400088, India,
rekha.sakec@gmail.com

M. SASIKUMAR
KBCS group, CDAC Mumbai (formerly NCST)
Kharghar, Navi Mumbai, India
the.little.sasi@gmail.com

The advance of semantic web and e-learning technologies has provided more opportunities to achieve the goal of collaborative knowledge sharing. It has also facilitated teachers to share their teaching material, tools, and experiences with others through the medium of internet and web technologies. In an earlier paper, we proposed the need for creating a Distributed Question Bank (DQB) by different experts in the related fields. We explored the possibility of using the semantic web technology and ontology in particular in addressing the issue of question similarity in a DQB. In this paper we extend it further, and propose a method of creation of subset ontologies for the questions, and comparing them to find the overlapping concepts to determine question similarity. We also tried formulating a model based on information theoretic approach using this notion of subset ontology, to measure the similarities among the questions in the data set considered.

Keywords: DQB, ontology, question similarity, similarity measure, semantic web, e-learning.

1. Introduction

The DQB contains pool of questions generated and managed by different experts for a particular domain [Vili *et al.* (2007)]. All the faculty members can use these questions to generate a question paper when needed. As the questions are set by different persons with different perspectives, the questions will be versatile. While generating a question paper from DQB, the questions taken should cover the entire topic in a fair manner. Assignment of marks to each question should be based on its expected difficulty levels, time for completion and relevance to the given syllabus. The questions need to be checked as to whether they are pertaining to overlapping concepts leading to redundancy. In this paper we are concentrating on formulating the measure of similarity among questions by analyzing the overlap among the concepts [Rekha and Sasikumar (2010)]. The extent of similarity can be as follows:

(1) Total dissimilarity among the questions meaning that there is no overlapping of concepts

(2) Partial similarity amounting to the overlapping of concepts from the domain of interest.

(3) Complete similarity with a one to one matching of concepts.

Similarity has different dimensions. The similarity between two questions is based on (i) the subject to which the question belongs (ii) the topics covered by the question (iii) the type of question and (iv) the difficulty level of question as defined by a structure like Bloom's taxonomy [Abrahams and Wei (2005)]. These attributes form the metadata associated with each question. The question metadata can be compared to find the extent of overlap among the questions.

In this paper, we explore the use of semantic web technologies and ontology in particular in a DQB scenario [Lee *et al.* (2001)] [Grigoris and Frank (2003)]. Though semantic web is concerned with web pages and documents, we can view each question (along with associated metadata) as a "web page" and borrow a lot of ideas and tools to address our problems. Ontology describes the subject domain using notions of concepts, instances, attributes, relations and axioms [Grigoris and Frank (2003)]. For example, in our question bank scenario, the syllabus of the question bank forms the domain and topics in which questions can be asked are the concepts. The relationships include typically hierarchy of concepts. The major topics can be further narrowed down to subtopics that form the subclasses in the ontology. Ontology may also include information such as properties, value restrictions, disjointness statements and specification of logical relationships between objects. We have chosen our domain as Data Structures subject in computer science field. We collected around 50 questions related to graphs from the teachers of different colleges as well as from previous years question papers and compiled them. The corpus of questions spans wide variety in terms of their difficulty, type of question, language used to frame the questions, etc, covering the entire topic of the syllabus. We studied the relationship between them in terms of their independence, overlapping concepts, difficulty levels, etc.

To find the similarity between two questions, the concepts associated with each question are mapped to the domain ontology. Comparing these mappings gives a fair idea of the overlap of concepts between two questions. We follow this intuition in the paper in computing similarity measures.

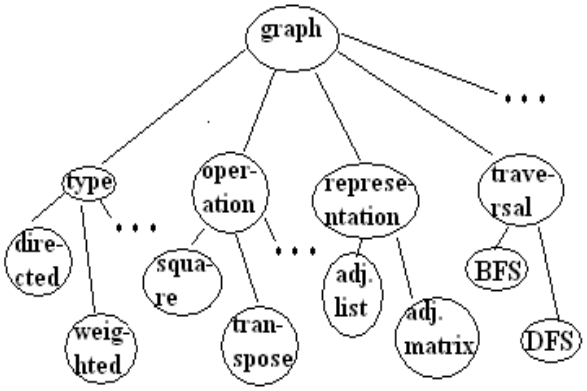
In this paper we explore the possibility of using the semantic web technology and ontology in particular to compute the similarity among questions. Section 2 gives the structure of domain ontology. Section 3 provides the details of generation of subset ontology for a question. Section 4 introduces the formula for similarity measure and different techniques of assigning weights to the nodes in the subset ontology. Section 5 gives various ways of improving the similarity measure. Section 6 provides the results obtained by applying the formula to our corpus of questions. This is followed by a conclusion of the findings in section 7.

2. Domain Ontologies

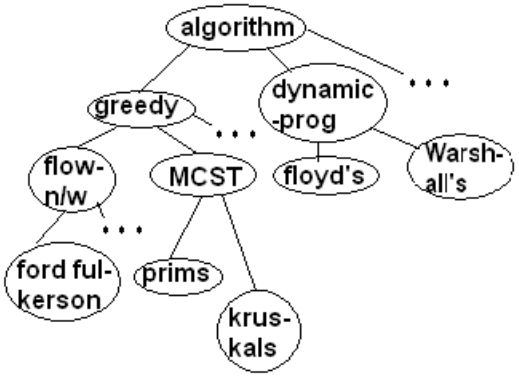
We analyzed the domain and noticed that a single large ontology may not suffice to describe the domain completely. We found that each question has three different dimensions:

(1) The basic data structure (arrays, stacks, queues, linked list, graphs, trees), types, their representations, traversals, operations on them, etc.
(2) The type of algorithm that uses specific data structures
(3) Performance analysis of algorithms

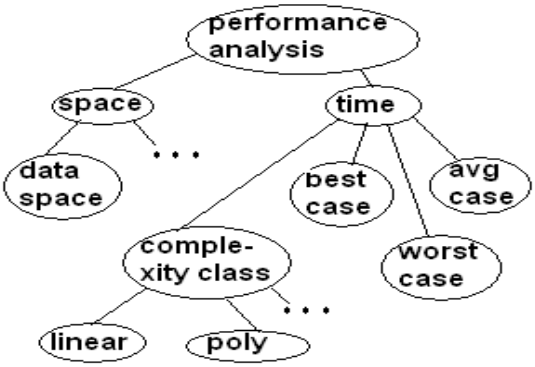
Each of these aspects has an associated vocabulary, with some linkages to other aspects. For example tree-traversal can be seen as an operation on the tree data structure or as a traversal algorithm and has a performance analysis component. We therefore defined three different ontologies for these three dimensions shown in fig 1 a, b & c [Andreas *et al*. (2008)]. These domain ontologies provide the shared conceptualization between different persons sharing the DQB. The domain ontology can be represented as a directed graph consisting of nodes and edges, where nodes denote topics in the domain chosen (e.g. Data Structure) and edges denote the binary relationship of the linked topics. The ontology at present captures only subject relationship using hierarchy.



(a) Ontology for Basic data structure



(b) Ontology for Algorithm



(c) Ontology for Performance Analysis

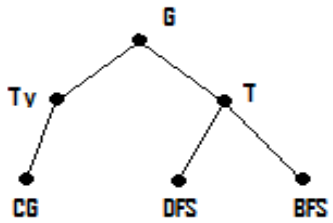
Fig. 1. Domain Ontologies

3. Subset Ontologies

We now need to map a given question to these ontologies. We extract the concepts from each question and tag the respective nodes from all the ontologies. At present, extraction is performed manually. The subgraph formed by these tagged nodes and associated edges form the subset ontology for that question.
For example, consider the following two questions

Que.1. Explain, when *DFS* and *BFS* are applied to a *connected graph* they result in a tree
Que.2. Identify the *connected components* of a graph using a *DFS*

The concepts related to Que.1 are *DFS*, *BFS* and *connected graph (CG)* in ontology 1 and those related to Que.2 are *identify connected components (FCC)* and *DFS* in ontology 2. These concepts are mapped to all three parts of domain ontologies to get the subset ontology. Figures 2 and 3 indicate the subset ontologies for Que.1 and Que.2 respectively. We will be using these two questions as running example through the paper to illustrate the measures discussed.



(a) (b) (c)

Fig. 2. Subset Ontology for Que.1

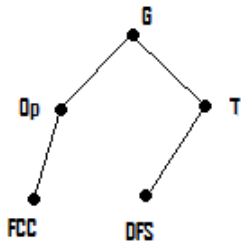


Fig. 3. Subset Ontology for Que.2

For both the questions *Algorithm* and *Performance Analysis* component is nil which is indicated by empty (b) and (c) sections. Looking at these subset ontologies gives a fair idea of the overlap of concepts between the two questions. We follow this intuition in the paper in computing similarity measures.

4. Measure of Similarity

There are several algorithms developed to compute the similarity between two sentences. They actually compute the semantic distances between them. In our approach we are finding the semantic distance between two questions by comparing their subset ontologies. We briefly surveyed existing ontology distance computation algorithms. [Jing, L., *et al.* (2006)] Based on this survey, we feel that several methods can be adopted to form a semantic distance measurement methodology. To begin with, we focused on the information theoretic approach which uses weighted directed graph [Xia *et al.* (2006)]. Each node in the subset ontology can be assigned some weight values. Nodes which are present in both can be given more weightage than nodes which are unique to one subset. Subset ontologies are compared to find commonalities and differences between them [Halang and Wang (2008)].

The definition of similarity between subset ontologies C and D (shown in figures 2 and 3) uses its commonalities and differences. The expression for calculating the similarity is as follows

$$sim(C,D) = \frac{|C \cap D|}{|C \cap D| + \alpha |C/D| + \beta |D/C|} \quad (1)$$

Where α, β Є [0, 1].

|C∩D| is the sum of the weights of common elements of C and D. The common elements may include the number of shared attributes, instances and relational classes, etc. |C/D| is the sum of the weights of the elements that are in C but not in D. |D/C| is the sum of the weights of the elements that are in D but not in C. α and β are constants. The value of similarity Sim(C, D) lies in the range of [0, 1] with 0 indicating total dissimilarity and 1 indicating total similarity. Others indicate partial similarity i.e. there is some amount of overlapping of concepts from the domain of interest.

4.1 The value of α and β

α and β are constants defining the relative importance of the noncommon nodes vs. common nodes. The value of α greater than β implies more importance is given to non matching nodes in |C/D| and vice versa. The value of α=β=1 indicates equal importance to the non matching nodes in |C/D| and |D/C| as well as giving same weightage to commonalities and differences in the subset ontologies under consideration. We have decided to keep the value of α=β=0.5 i.e. matching nodes have more importance than nonmatching nodes. Since similarity in our scenario is a symmetric relation, it is expected that α=β.

4.2 Assigning Weights

Subset ontology is a weighted graph. Each node in the graph represents a topic related to the question. As we go deeper to the lower level in the graph, the topics get narrower. Questions will be less similar if they share common nodes at the higher level in the graph and they will be more similar if they have more number of common nodes at the lower level. A challenging problem is to assign the weights to reflect such variation. The weights assigned to the nodes affect the similarity measure of the questions as the weights gets summed up while finding the commonalities and differences between the subset ontologies under consideration. There are many considerations in assigning weights to the nodes in the subset ontology.

The simplest scenario would be to assign a constant weight of value one to all the nodes in the subset ontology. Here every node in the subset ontology has equal importance. Considering again the subset ontologies of Figures 2 and 3 with the weights attached to it (shown in figures 3 and 4). Nodes such as Graph (G), Traversal (T) and DFS are common in both the subset ontologies (indicated by nodes with square shape). Node such as Type (Ty), Connected Graph (CG) and BFS are unique to ontology for Que. 1. Similarly nodes such as Operations (Op) and Find Connected Components (FCC) are unique to ontology for Que.2. Based on the above information we can find the values of |C∩D|, |C/D|, D/C|.

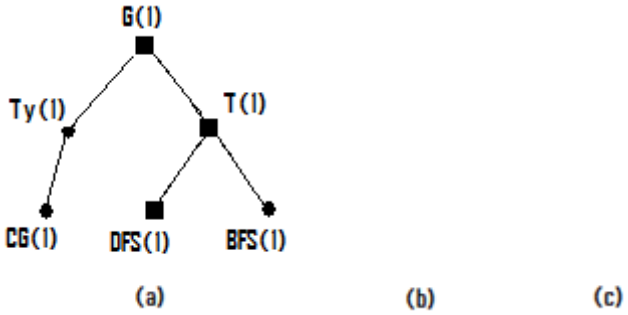


Fig. 3: Subset Ontology for Que. 1 with associated constant weights (concept C)

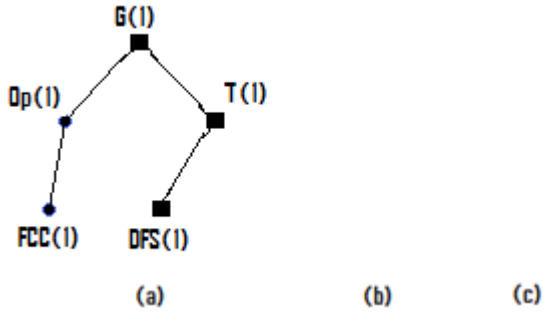


Fig. 4: Subset Ontology for Que. 2 with associated constant weights (concept D)

|C∩D| = sum of weights of the common elements of concept C and D
= 1+1+1 = 3
|C/D| = 1+1+1 = 3
|D/C| = 1+1 = 2

$$Sim(C,D) = \frac{3}{3 + 0.5(3) + 0.5(2)} = 0.545$$

The similarity value of 0.545 indicates that the questions are about 54 percent similar. This may be because both of them belong to same graph topic and in that traversal subtopic. But when we look at it with the human perspective the value seems to be on the higher side.

Here we have considered all the nodes to be of same importance by giving equal weightage to them which may not be true in all cases. Even though the questions belong to same topic their intentions may be entirely different which make them less similar. This aspect of similarity needs to be further explored and should be accounted somewhere in the formula.

Some nodes can be identified as "important" nodes. These nodes represent the lead part of the question. Such nodes can be given higher weight values than others. Generally the words (or concepts) following the constructs such as *solve, list, perform, design, find, identify, give, construct, explain, implement*, etc are the important nodes in a question. They essentially capture the essence of the question. Suitable algorithm need to be explored for finding the important nodes when given a specific question. For example, considering the above questions again we can say that for Que.1 *DFS, BFS, CG* nodes are important. For Que. 2, *FCC and DFS* are important nodes. What values should be assigned to these nodes needs further consideration. Suppose we assign double the value of other nodes (value of 2) to these nodes. Then the similarity value:

$$Sim(C,D) = \frac{4}{4 + 0.5(5) + 0.5(3)} = 0.5$$

In this case the similarity value has reduced to 50% from 54% in the previous case. For this particular example important nodes are present in set of common nodes (C∩D) and the set of noncommon nodes (C/D and D/C). If the important nodes are present only in the set of noncommon nodes then the similarity value will reduce drastically even if there

are more number of common nodes. This means that the questions may belong to same topic but their intentions may be totally different. These intentions are captured by the important nodes. There are still some issues which need to be considered such as:

- What is the significance of doubling the value of important nodes?
- Whether to assign same values to all the important nodes or the values should depend on the level at which these nodes are present?
- Can we consider higher value for the important nodes if they are present in the set of common nodes?
- What value should be assigned to the node in (C∩D) if it is an important node in one subset ontology but is not important in the other?

We are currently working further on these issues

Another concern in weight assignment is the importance of level at which a node occurs. The weights associated with each node in the subset ontology can be made dependent on the level at which the node is present. Level 1 nodes will have the weight value of 1, level 2 will have weight value of 2, and so on. Considering the Que.1 and Que. 2, the following figures (5 and 6) show the subset ontologies with the weight values based on level.

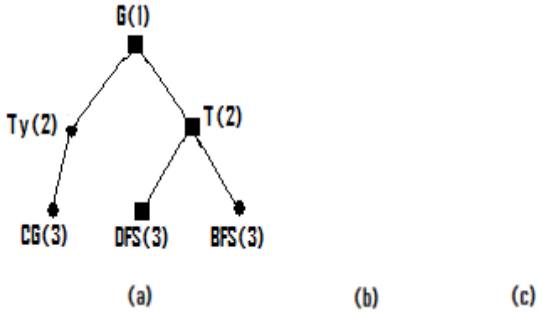


Fig. 5. Subset Ontology for Que. 1 with the weight values based on level (concept C)

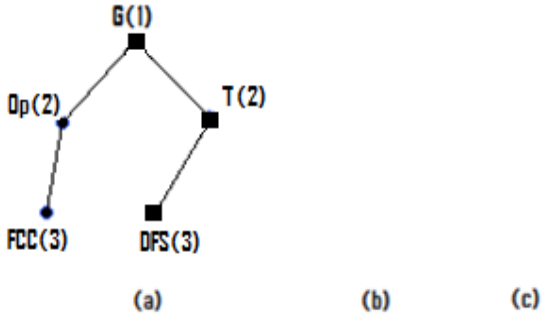


Fig. 6. Subset Ontology for Que. 2 with the weight values based on level (concept D)

The similarity value now is:

$$Sim(C,D) = \frac{6}{6 + 0.5(8) + 0.5(5)} = 0.48$$

In this case the similarity measure may vary with the depth of the ontology rather than the concepts involved in the question. As the level of the ontology increases the weight values also increases. Introducing more number of nodes in the middle level of the ontology affects the similarity measure considerably. As this is not desirable this technique should be used with care.

5. Enhancing the similarity measure

The matching process can be further improved by enhancing the formula for the similarity measures. Various factors affecting the similarity can be taken into consideration while calculating the values of |C∩D|, |C/D| and |D/C|. We studied the relationship among various parameters such as number of matching nodes, number of non matching nodes, total number of nodes in each ontology, etc. We tried to use this knowledge to improve the accuracy of similarity measure.

(1) Number of matches against total number of nodes

For finding the similarity the most important task is to find the number of matches |C∩D| in both the subset ontologies of the respective questions. But the matching nodes cannot be considered as absolute value as both the ontologies may not be of the same size. Suppose size of the first ontology is m, size of the second ontology is n and p is the sum of the weights of matching nodes. Assuming weights of all nodes is equal to 1:
The number of matches and mismatches can be made proportional to total number of nodes in the respective ontologies.

$$\left|C \cap D\right| = \frac{p}{m} + \frac{p}{n}\ , \qquad \left|C / D\right| = \frac{m-p}{m}\ , \qquad \left|D / C\right| = \frac{n-p}{n} \tag{2}$$

For example, consider the following two questions

Que.3 Find the *path matrix* of G using *Warshall's algorithm.*
Que.4. Apply *Warshall's algorithm* to find the *transitive closure* of the *diagraph* defined by the *adjacency matrix***.** Analyze the *running times* of your algorithm.
.
Figures 7 and 8 shows the subset ontologies for Que.3 and Que. 4

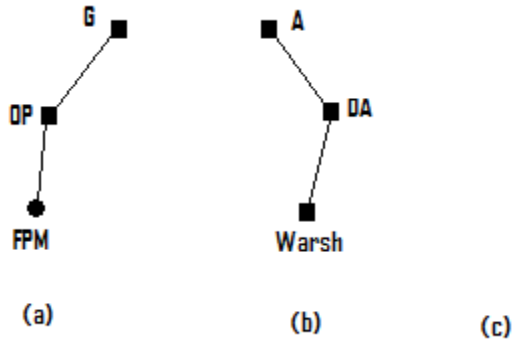


Fig. 7. Subset Ontology for Que. 3

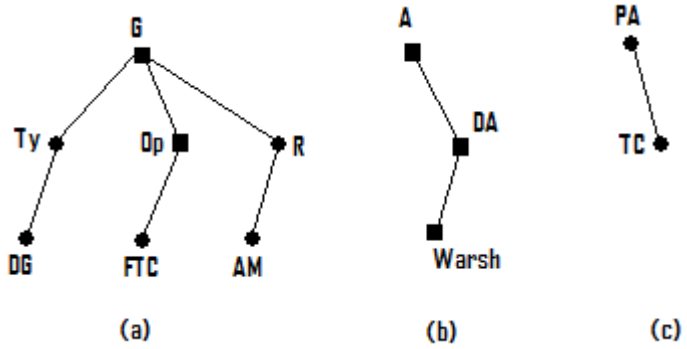


Fig. 8. Subset Ontology for Que. 4

From the subset ontology p = 5, m = 6, n =12. So n>>m

$$\left|C \cap D\right| = \frac{5}{6} + \frac{5}{12} = 1.25$$

$$\left|C / D\right| = \frac{1}{6}\ , \qquad \left|D / C\right| = \frac{7}{12}$$

$$Sim(C, D) = \frac{1.25}{1.25 + 0.08 + 0.29} = 0.77$$

(2) Number of matches against size of union of two ontologies

The denominator for all the terms can be considered as the union of nodes in both the subset ontology for the questions under consideration ($U_{c,d}$).

$$\left|C \cap D\right| = \frac{p}{U_{c,\ d}}, \qquad \left|C / D\right| = \frac{m-p}{U_{c,\ d}}, \qquad \left|D / C\right| = \frac{n-p}{U_{c,\ d}} \tag{3}$$

For the above questions $U_{c,d} = 13$

$$\left|C \cap D\right| = \frac{5}{13} = 0.3846 \ , \ \left|C / D\right| = \frac{1}{13} \ , \ \left|D / C\right| = \frac{7}{13}$$

$$Sim(C, D) = \frac{0.3846}{0.3846 + 0.5(0.076) + 0.5(0.538)} = 0.556$$

(3) Number of matches against number of mismatches

From the first ontology, the number of non matching nodes is m-p. From the second subset ontology, the number of non matching nodes is n-p. From this

$$\left|C \cap D\right| = \frac{p}{m-p} + \frac{p}{n-p} \tag{4}$$

From the subset ontology of Que.3 and Que.4

$$\left|C \cap D\right| = \frac{5}{1} + \frac{5}{12} = 5.41$$

$$Sim(C, D) = \frac{5.41}{5.41 + 0.5(1) + 0.5(7)} = 0.57$$

(4) Number of matches against maximum of two Ontologies

The denominator for each term should be equal to maximum of the size of the two ontologies which is max (m, n).

$$\left|C \cap D\right| = \frac{p}{\max(m, n)} \ , \ \left|C / D\right| = \frac{m-p}{\max(m, n)} \ , \ \left|D / C\right| = \frac{n-p}{\max(m, n)} \tag{5}$$

From the subset ontology of Que.3 and Que.4

$$\left|C \cap D\right| = \frac{5}{12} = 0.416 \ , \ \left|C / D\right| = \frac{1}{12} \ , \ \left|D / C\right| = \frac{7}{12}$$

$$Sim(C, D) = \frac{0.41}{0.41 + 0.5(1) + 0.5(7)} = 0.55$$

Using the basic similarity measure without any enhancement we get |C∩D|=5, |C/D|=1, |D/C|=7 and *Sim (C, D)* = 0.55. But Looking at Que.3 and Que.4 we can say that they are almost (70%) similar. To improve the calculated similarity measure the different similarity enhancement measures discussed in this section can be combined with the different weight assignment methods illustrated in the section 4. One such combination is being explored in the next section.

6. Current Results

In section 5, item (2) we considered the denominator for all the terms in the formula as ($U_{c, d}$). But weights of all the nodes were taken as same and equal to 1. To further improve the result we combined node importance discussed in section 4 with the formula. The weight values for each node is determined based on its importance in the subset ontologies for the questions under consideration. So the terms |C∩D|, |C/D| and |D/C| included the weighted sum of nodes in them. So revisiting Eq. (1), the formula for calculating the similarity between two ontologies C and D is

$$sim(C, D) = \frac{|C \cap D|}{|C \cap D| + \alpha|C / D| + \beta|D / C|} \tag{6}$$

Where α, β ∈ [0, 1],

$$|C \cap D| = \frac{Weighted\ \ sum\ of\ \ the\ common\ nodes\ in\ C\ \ and\ \ D}{U_{c, d}}$$

$$|C / D| = \frac{Weighted\ \ sum\ of\ \ the\ common\ nodes\ in\ C\ but\ not\ in\ D}{U_{c,\ d}}$$

$$|D / C| = \frac{Weighted\ \ sum\ of\ \ the\ \ nodes\ in\ D\ but\ not\ C}{U_{c,\ d}}$$

To evaluate usefulness of our formula, we tested it with a corpus of questions. Each of the questions was compared with every other question in the set and similarity measures were obtained. The similarity values thus obtained was put into the similarity matrix as shown in the table 1. The same sets of questions were then given to teachers teaching the same subject for comparing. They were told to quantify some value as similarity measure in the range of (0, 1) depending upon how they feel about the similarity of the questions. The similarity matrix thus constructed manually is shown in the table 2. Currently this is done only for a single person and in future we will extend it to multiple persons.

Table 1. Similarity matrix calculated using formula

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Q1	1	0.46	0.13	0.25	0.22	0.37	0.12	0.1	0.09	0.33
Q2		1	0.11	0.096	0.19	0.33	0.33	0.28	0.26	0.8
Q3			1	0.93	0.14	0.102	0.16	0.09	0.12	0.13
Q4				1	0.26	0.19	0.15	0.12	0.11	0.12
Q5					1	0.6	0.28	0.23	0.22	0.23
Q6						1	0.22	0.28	0.27	0.37
Q7							1	0.8	0.37	0.37
Q8								1	0.2	0.33
Q9									1	0.7
Q10										1

Table 2. Similarity matrix found manually

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Q1	1	0.2	0.1	0.2	0.2	0.3	0.1	0.1	0.1	0.2
Q2		1	0.1	0.1	0.1	0.2	0.2	0.2	0.3	0.8
Q3			1	0.9	0.2	0.1	0.1	0.1	0.1	0.1
Q4				1	0.2	0.1	0.1	0.1	0.1	0.1
Q5					1	0.7	0.1	0.2	0.1	0.1
Q6						1	0.1	0.1	0.1	0.2
Q7							1	0.9	0.2	0.2
Q8								1	0.2	0.2
Q9									1	0.6
Q10										1

Comparing the two tables (table 1 and table 2) we can see that there is good agreement between manual and calculated values. The absolute difference between the similarity values of two tables is shown in table 3. Looking at the shaded regions it is seen that most of the values are in the range is between 0 and 0.2 which are within 6% of human judgment. The average difference and the standard deviation is found to be 0.061927 and 0.065447. The accuracy can be further enhanced by the proper selection of important nodes, refining the domain ontology, having a relationship between multiple domain ontology, etc.

Table 3. Match between manual and calculated values

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Q1	0	0.3	0.03	0.05	0.02	0.07	0.02	0	0.01	0.13
Q2		0	0.01	0.004	0.09	0.13	0.13	0.08	0.04	0
Q3			0	0.03	0.06	0.002	0.06	0.01	0.02	0.03
Q4				0	0.06	0.09	0.05	0.02	0.01	0.02
Q5					0	0.1	0.18	0.03	0.12	0.13
Q6						0	0.12	0.18	0.17	0.17
Q7							0	0.1	0.17	0.17
Q8								0	0	0.13
Q9									0	0.1
Q10										0

Legend

	0 -0.09
	0.1-0.19
	0.2-0.3

Average= 0.061927
Standard deviation = 0.065447

7. Conclusion

In this paper we described the problem of identifying similarities between questions in DQB using ontology as the base. We have introduced the notion of subset ontology, identified various issues in defining a distance measure for this problem and proposed a initial model for similarity detection. Experimental results for the domain of graphs in Data structure have been reported. The result is encouraging and is able to come within 6% of human judgment in most cases. We are currently investigating extension of the model to larger domain and addressing the various issues mentioned in that context.

References

Abrahams, B., Dai, W. 2005. Architecture for Automated Annotation and Ontology Based Querying of Semantic Web Resources. Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence. WI 2005. Sep. 19-22. France. IEEE Computer Society Press. Pp 413-417.

Andreas, P., Konstantinos, K., Konstantinos, K. 2008. Automatic Generation of Multiple Choice Questions from Domain Ontologies. e-Learning (2008), 427-434.

Berners-Lee, J. Hendler, and O. Lassila. 2001. The Semantic Web**.** Scientific American. 284(5) (2001), 35-43.

Grigoris, A. and Frank, van, H. A Semantic Web Primer**.** The MIT Press, Cambridge, Massachusetts, London.

Jing, L., *et al.* .(2006). Ontology-based Distance Measure for Text Clustering. Proceedings of the Fourth Workshop on Text Mining Sixth SIAM International Conference on Data Mining Hyatt Regency Bethesda Bethesda, Maryland April 22, 2006

Rekha, R., Sasikumar, M. (2010). Use of Ontology in Addressing the Issues of Question Similarity in Distributed Question Bank**.** (ICWET'10) , February 26–27, (2010)

Vili, P., Luka, P., Marjan, H. 2007. Semantic web based integration of knowledge resources for supporting collaboration. Informatica. March, (2007)

What is an Ontology? http://www-ksl.stanford.edu/kst/what-is-an-ontology.html.

Xia, W., Yihong, D., Yi, Z. 2006. Similarity Measurement about Ontology-based Semantic Web Services. Workshop on Semantics for Web Services (SemWS'06) in conjunction with 4th European Conference on Web Services (ECOWS'06), (2006).

Yi Zhao Halang, W. Xia Wang. 2008. A Rough Similarity Measure for Ontology

Mapping. Internet and Web Applications and Services, 2008. ICIW '08, 136-141.

Jing, L., *et al.* (2006). Ontology-based Distance Measure for Text Clustering. Proceedings of the Fourth Workshop on Text Mining Sixth SIAM International Conference on Data Mining Hyatt Regency Bethesda Bethesda, Maryland April 22, 2006