Question Bank Similarity Searching System (QB3S) Using NLP and Information Retrieval Technique

Conference Paper · May 2019

DOI: 10.1109/ICASERT.2019.8934449

2 authors:

Md Raihan Mia
Bangladesh University of Engineering and Technology
4 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Abu Sayed Latiful Haque
Bangladesh University of Engineering and Technology
52 PUBLICATIONS **295** CITATIONS
SEE PROFILE

Question Bank Similarity Searching System (QB3S) Using NLP and Information Retrieval Technique

*

Md. Raihan Mia
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
Email: 1305116.mrm@ugrad.cse.buet.ac.bd

Abu Sayed Md. Latiful Hoque
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
Email: asmlatifulhoque@cse.buet.ac.bd

***Abstract*—Problem Based e-learning(PBeL) in bangla language is one of the most progressing areas of the use of ICT in education. Question Bank(QB) is the main component of any PBeL system. Searching similarity in the complex structure of QB is a challenging task in the development of PBeL system. We have been developed an efficient Question Bank Similarity Searching System(QB3S) to find similar questions, handle duplicate question and rank search result of a query input based on NLP and Information Retrieval techniques. QB3S has four modules: bangla documents processing, question structure analysis and clustered indexing by B+ tree , word-net construction and Information retrieval module. Lexical analysis, stemming by finite automata rules and stopwords removing have been used for bangla document processing. The most challenging procedures of QB3S were Analyzing the structure of data for clustered indexing in the sorted sequential file of the QB DataBase(DB) with a B+ tree data structure and improved TF-IDF algorithm with weighted functionality. A Word-net has been used for handling synonyms. Vector Space Model(VSM) has been designed from the value of TF-IDF weighted matrix. By using cosine similarity product rule, we have been Calculated the similarity value between the query input and all mcq of DB from VSM. QB3S has been evaluated in some experimental dataset to find results by imposing different test cases. The accuracy of searching performance which has found to be satisfactory.**

***Index Terms*—Tokenization, Stopwords Removing, Stemming, Clustered Indexing, B+ tree, Word-Net, TF-IDF, VSM, Cosine Similarity**

I. Introduction

Problem-Based eLearning(PBeL) systems have been proved to be effective in blended learning in classroom and also in asynchronous learning. The PBeL system [1] developed for ICT course in Higher Secondary School(HSC) level contains a rich question bank of different types and complexities. It is utmost requirement of the QB to contain only distinct questions in terms of title, contents and answers. This requires a rigorous similarity search to remove the possible duplicate questions. In this research, we have developed a similarity searching system on complex structures of QB using NLP and IR technique. Beside duplicate handling, it also can be ranked searching results of a query input based on similarity values.

Information Retrieval (IR) [2] can be defined as a set of techniques and tools dealing with access to information from a set of unstructured documents database in order to allow the user to retrieve the correct information. IR algorithms for fetching relevant information from large amount of structured data of QB database where data is represented in tabular form, need to strategically better searching techniques.

We have used NLP tools [3] like, Tokenization or lexical analysis of bangla text corpus [4], StopWord removing, Rule-based sementic analysis or stemming to modify inflected word and analyzed the structured data to build a dynamic B+ tree clustered indexing to speed up the retrieval of records. To build the generative laexicon [5] for bangla language, many paper had been already published in the field of computational linguistics morphology. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. Algorithms for stemming have been studied in computer science since the 1960s. In [6], A rule-based approach of stemming bangla word had been found an excellence solution. The addition of inflectional suffix; derivational suffix and agglutination in compound words make Morphological Parsing fairly complex for the Bangla. There are existing efforts at building a complete morphological parser for Bangla [7], where experiments have been carried out with two types of algorithms: simple suffix stripping algorithm and score based stemming cluster identification algorithm. Another rule-based intelligent Bangla parser is to ease the task of handling semantic issues in the subsequent stages in machine translation [8].

B+ tree data structure for indexing Object Oriented Database has been found time efficient searching, inserting and deleting algorithms in [9]. Clustered Metric Tree(CM-Tree) had been constructed for a dynamic clustered index in unstructured metric database for searching similarities [10]. Variations of the TF-IDF(Term Frequency - Inverse Document Frequency) weighting scheme are often used by IR techniques as a

central tool in scoring and ranking a documents relevance given a user query [11]. Many TF-IDF weights function has been developed for verious IR applications like, an improved TF-IDF weights function is proposed which uses the distribution information among classes and inside a class [12]. Improved version of a well functional weighting scheme of TF-IDF algorithm based on question title or option or answer terms, Vector Space Model [13] for representing documents as a vector in t-dimension (If the document contain t no. of terms) and find the similarity with query vector using cosine similarity product rules are the main area of IR techniques used in QB3S. A hash-mapping WordNet is used to handle same meaning words or synonyms. Main objective of our research showing below:

- To improve in a certain level of Bangla documents processing tools like tokenization, stopwords removing, stemming to accomplish the processing of QB data-set.
- To analyze the structure of data and indexing in sorted sequential records(Clustered Indexing) of QB database for faster access by using B+ tree data structure.
- To construct a hash mapping word-net to handle synonyms.
- To improve a weighted functional TF-IDF algorithm.

The rest of the paper has been organized as follows. Section II describes the structure of QB[14-17]. The system architecture for QB3S has been given in section III. Section IV shows the results obtained after the application of QB3S in the existing eLearning system [1]. A detailed discussions on the results has also been given in this section. Finally the conclusion is given in section V.

II. Question bank Structure and e-learning of ICT course

A sequence of papers [14-17] were published in journals or conference towards improving PBeL in ICT course of secondary and higher secondary level of Bangladesh. Problem based e-learning system (http://epbl.org) for interactive learning of ICT course, practicing and model examination from question bank and C programming, HTML, SQL, ERD learning based on higher secondary book have already been implemented successfully. Question bank data are important resource for Problem Based e-Learning(PBeL) which requires a search engine to retrieve query information from the large scale of data and this can be used by the admin of PBeL system to find the duplicate question, user can find out searching result of the relevant query input.
Let us take a closer look at the information and utilities available in a bangla question bank database. It can be Categorized the Multiple choice questions in three classes:

1) Cognitive class:
   The simplest structure where there are a question title and 4 options.Like,

   আধুনিক কম্পিউটারের জনক কে?
   ক) অ্যাডা লাভলেস
   খ) চার্লস ব্যাবেজ
   গ) বিল গেটস
   ঘ) জেমস ক্লার্ক ম্যাক্সওয়েল

2) Analytical class:
   It contains three option and the scope of answer is the four combinations of these options.As example,

   কম্পিউটারের সময় অপচয়কারী ব্যবহার –
   i) প্রোগ্রামিং শিক্ষা
   ii) ভিডিও গেম
   iii) সামাজিক যোগাযোগ
   নিচের কোনটি সঠিক ?
   ক) ii ও iii খ) i ও iii
   গ) i ও ii ঘ) i. ii ও iii

3) Higher Ability class: One of the most complex structure and scenario based questions.There are may be two or three Cognitive or Analytical mcq question based on a scenario. As example,

   নিচের অনুচ্ছেদটি পড় এবং 1-2 নং প্রশ্নগুলোর উত্তর দাও:
   বর্তমান আমাদের দেশে বেকার সমস্যা মানাতœক সামাজিক ব্যাধি। প্রায় প্রতিটি সমাজেই এ ব্যাধিতে আক্রান্ত। বেকার সমস্যা দূর করতে বর্তমানে HTML, CSS, JAVA Script সহ বিভিন্ন ধরনের CMS সফটওয়্যার এর মাধ্যমে ওয়েব ডিজাইন করে বেশ অর্থ উপার্জন করছে।
   1. CSS- এর পূর্ণরূপ কী?
   ক. Caseding Style Sheet
   খ. Compund Style Sheet
   গ. Compund Shorting System
   ঘ. Cacacing Simple Sheet
   2. কত সালে তাদের প্রতিযোগিতামূলক ব্রাউজার রিলিজ করে?
   ক. ১৯৯৫ সালে
   খ. ১৯৯৬ সালে
   গ. ১৯৯৭ সালে
   ঘ. ১৯৯৮ সালে

Orgatization of stroed question in database schema is showing in Fig. 1. We need to analysis the structure of question and cluster it with three decision boundaries of three different classes will be described in sec. III.

	mcqid	qtitle	op1	op2	op3	op4	op5	image	answer
	c110010030051	ফুল অ্যাডার সার্কিটের সত্যক সারণিতে কয়টি সা…	3	8	4	16			8
	c110010030052	অ্যাডার এবং রেজিস্টার ব্যবহৃত হয় কম্পিউটারে…	গাণিতিক …	যুগান্তিক	মেমোরি	মধ্যিতে			গাণিতিক ও যোক্তিক ই…
	c110010030053	শিফট রেজিস্টারে ইনপুট থাকে.	ঘড় বিটের …	1 টি	ঘড় বিটের রে…	কোনোটি নয়			1 টি
	c110010030054	একটি 4 বিট রেজিস্টারের 0000 সংরক্ষণ করার উ…	ইনপুট 00…	Reset ইন…	ইনপুট 0000 …	ইনপুট 0000 দেওয়া, …			ইনপুট 0000 দেওয়া, …
	c110010030055	D ফ্লিপ ফ্লপ এর ক্ষেত্রে আউটপুট 0 থেকে 1 করত…	0, 0 থেকে 1	1, 1 থেকে 0	1, 0 থেকে 1	1, 1 থেকে 1			1, 0 থেকে 1
	c110010030056	ক্লকের Edge কোনটি?	0 থেকে 1 …	1 থাকা অ…	0 থাকা অবস্থান…	যে কোনোটি			0 থেকে 1 রূপান্তরের স
	c110010030057	নিচের কোনটি কাউন্টারের ব্যবহার নয়-	CPU-G হ্রো…	সার্কিস সে…	ঘড়িতে	মধ্যিতে			মধ্যিতে
	c110010030058	কাউন্টারের একটি বিটের আউটপুট পরবর্তী বিটে…	অ্যাসিনক্রে…	ফাইবারি …	BCD কাউন্টার	সিনক্রোনাস কাউন্টার			অ্যাসিনক্রোনাস কাউন্টার
	c110010040001	ওয়েব সার্ভারে থাকা ওয়েব সাইটটি দেখার জন্য …	ওয়েব রাই	ওয়েব ব্রা…	ডোমেইন ব্রাউ	সার্ভার ব্রাউজার			ওয়েব ব্রাউজার
	c110010040002	WWW এর পূর্ণ রূপ হলো -	World We…	Website …	World Wide Web	Web Wide World			World Wide Web

Fig. 1. Organization of questions in database

III. QB3S system architecture

Theoretical analysis of the body of methods and principles associated with branches of procedure. QB3S has taken an input mcq or part of a mcq, processing data and performing search for finding the maximum similar mcq from entire QB DB. See Fig. 2, showing the architecture of system in terms of process flow. The whole process can be split into four modules:

1) Question structure analysis and clustering module
2) Document processing module
3) WordNet module
4) IR module

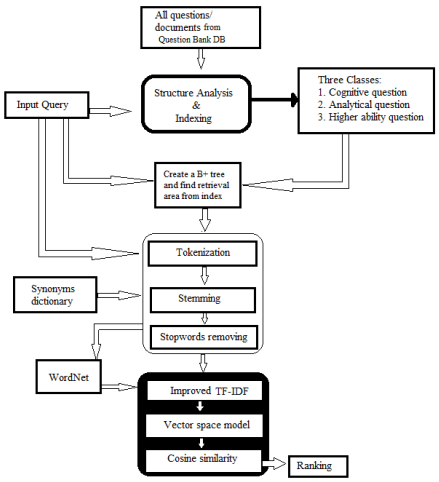


Fig. 2. System Architecture of QB3S

A. Question Structure analysis and clustering

In sec. II, we have been already discussed about the structure of question bank data and categorized them in three classes. Criteria of classification based on structure of question is showing in Fig. 3.

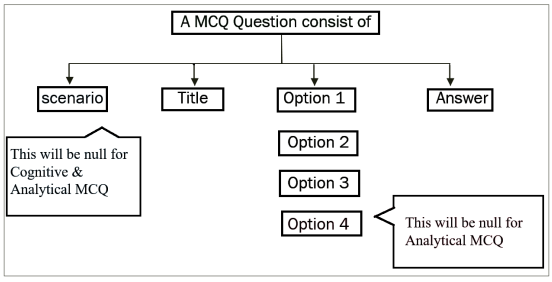
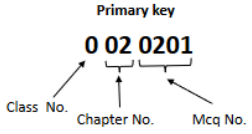


Fig. 3. Question structure analysis

PBeL system of ICT course database schema has been designed such a way that a table allocated for each class. Each table for a specific class contains identical questions of all chapters.

Clustered indexing by B+ tree:

B+ tree data structure has been used to track the clustered index(primary key on sorted order) of database table. B+ tree efficiently reduce the access time of search space in database of QB. In case of ICT course QB, it has 6 chapters and 3 classes of question. Primary key of table has been maintained as,



We had been compelled to track 18 indexes in the leaves node, each one pointed to the starting key of sorted data and degree of the B+ tree 6 in this case(see Fig. 4). So the height of tree ceil(log(18)/log(6/2)) = 2.

During the search operation, h nodes are read from the disk to the main memory where h is the height of the B+ tree and $h = log_t(n)$, where n is the number of the keys stored in the tree and t is size of a block or node . In addition of the disk reads, B+ tree searching algorithm performs a linear search in every node read from the disk. The time complexity of each linear search is O(t) . Thus, the total time complexity of the B + tree search operation is $O(tlog_t n)$ and same time complexity for inserting and deleting [18].

In case of higher number of chapter, it needs higher order and for adding other courses in QB, extra level will be added dynamically by imposing precondition [19].

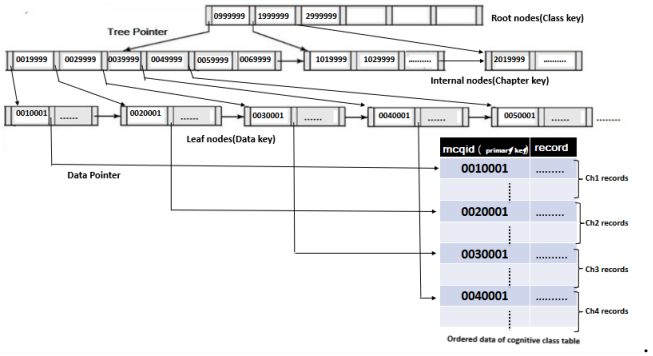


Fig. 4. B+ tree structure for clustered indexing

B. Bangla Document Processing

In the field of Bangla language processing, we have been already discussed the existing work of tokenization, Stemming and stop words removing in Sec. I. Now we are going to explained the tools of language processing which have been used in QB3S.

1) Tokenizer: Tokenization is the process of breaking up the given text into units called tokens.Tokenization refers lexical analysis in natural language processing.

We use a standard tokenizer, which splits text into terms on word boundaries using white space, as defined by the Unicode Text Segmentation algorithm [20]. It removes most punctuation symbols. Array-list data structure has been used to store the terms after tokenization.

2) Stemmer: A very simple-to-use rule based stemmer [6] for Bangla word parsing is used for semantic analysis. Which take a inflected word as input and output the corresponding root stem. Some common rules which is used for semantic analysis showing below and See examples of common rules text file contains in Fig. III-B2.

X #When X appears at the end of a word, remove it

$Y \rightarrow Z$ #When Y appears at the end of a word, replace it with Z

$Y.Z \rightarrow A.B$ #When Y, followed by some character a, followed by Z appears at the end of a word,replace it with AaB

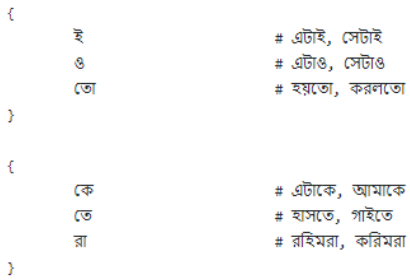


Fig. 5. Examples of some common rules for stemming

3) Stop words Remover: 'Stop Words' usually refers to the most common words in a language.There is no single universal list of bangla stop words. We build stop words list in followed way:

1) From some pre defined words
2) From the most frequently occurring word in Question Bank database

All stop words are removed from the array-list by using UTF-8 string matching library in java [21].

C. Word-Net module

As you know, synonyms are words that have similar meanings. A synonym set, or synset, is a group of synonyms. A synset, therefore, corresponds to an abstract concept.We have been constructed a runtime hash mapping from text file which contains bangla synonym dictionary(see Fig. 6).

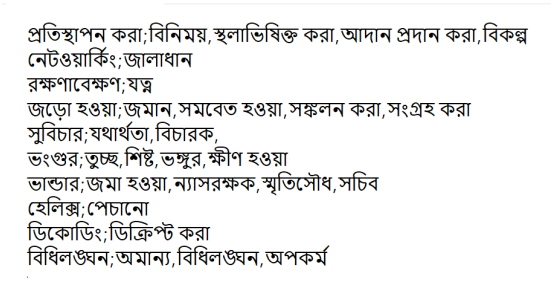


Fig. 6. Bangla synonym dictionary txt file

Step of hash-mapping word-net construction:

1) Take input from bangla dictionary text file
2) Tokeniztion
3) Stop-word removing
4) Stemming
5) Use hash function, $f(n) = charAt(0) \mod 128$ for finding hash value (total 128 bangla character,unicode value 0980H-09FFH).

Hash-map of the first line of Fig. 6 showing in Fig. 7.

Searching time complexity : O(1)

Fig. 7. Representation of dictionary hash-map

D. Information Retrieval module

1) Improved version of Weighted TF-IDF: Term Frequency(TF) measures how frequently a term occurs in a document.Suppose we have a query or input mcq consisting terms, $t_1, t_2, t_3, t_4.....t_n$ and there are many documents $mcq_1, mcq_2, mcq_3.....mcq_m$ where the tf-idf will be performed.Then,

$$TF(t, mcq_j) = \frac{f(t, mcq_j)}{\sum_{T \xi mcq_j} f(T, mcq_j)} \quad (1)$$

$Where, f(t, mcq_j) =$ Numbers of time Term t appear in mcq_j

Inverse Document Frequency(IDF) Measures how important a term is. Log base e of ratio between total number of documents and Number of documents with that term t in it.

$$IDF(t) = ln\frac{N}{n} \quad (2)$$

Where, N = Total number of documents And n = Number of documents with term t

We have been developed an improved version of this algorithm where we used some pre-defined weighting factor based on the classification of question structure.
Weighting Factor,

$$W(t) = \begin{cases} w1 & if\ t\ is\ a\ Answer \\ w2 & if\ t\ is\ an\ option \\ w3 & otherwise \end{cases}$$

Where, w1> w2>w3
So tf-idf equation with pre-define weighting factor as like,
$\text{TF-IDF}_{improved}(t, mcq_j) = \frac{f(t,mcq_j)*W(t)}{\sum_{T \xi mcq_j} f(T,mcq_j)*W(T)} * ln\frac{N}{n}$ (3)

2) Vector space Model Representation from TF-IDF: Vector space model or term vector model [22] is an algebraic model for representing text documents as vectors of identifiers like index terms. Formally, a vector space is defined by a set of linearly independent basis vectors.The basis vectors correspond to the dimensions or directions of the vector space. The basis vectors are linearly independent because knowing a vectors value on one dimension does not say anything about

its value along another dimension. Documents and queries are represented as vectors.

$$d_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$$

$$q = (w_{1,q}, w_{2,q}, \ldots, w_{n,q})$$

Each dimension corresponds to a separate term. If a term occurs in the document, its value in the vector is non-zero. Each dimension corresponds to a separate term. If a term occurs in the document, its value in the vector is non-zero. Graphical representation of vector space in Fig. 8.

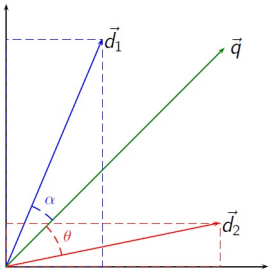


Fig. 8. Vector space model

Vector Space Model (VSM) is a typical method to describe the text feature in text classification at present. It adopts TF-IDF weights to compute the term weighting in each dimension of the text feature. However, it only considers the relationship between the term and the whole text but neglects the relationship between different terms. VSM has been used for representing tf-idf weight of a mcq as a vector in this space.Let's assume we have a query mcq '**Q**' like,

নিচের কোনটি উচ্চস্তরের ভাষা?
A. পাইথন
B. ভিজুয়াল বেসিক
C. জাভা
D. সব ক'টি
Answer: সব ক'টি

After processing it using the method described in Subsec. III-B '**Q**' will be transformed into:

নিচ উচ্চস্তর ভাষা পাইথন ভিজুয়াল বেসিক জাভা সব

Query '**Q**' can be represented in the 7-dimensional vector space. Calculated the value of TF-IDF in a wighting matrix where each row represents a vector of a document in the space of query document.After performing improvedTF-IDF algorithm the weighting matrix is showing in Table I.

TABLE I
VECTOR REPRESENTATION FROM TF-IDF VALUE IN VSM

Doc	$term_1$	$term_2$	$term_3$	$term_4$	$term_5$	$term_6$	$term_7$
mcq_1	0.0	0.05	0.05	0.05	0.05	0.05	0.05
mcq_2	0.05	0.08	0.0	0.01	0.0	0.03	0.0
mcq_3	0.0	0.12	0.22	0.33	0.44	0.55	0.83
mcq_4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
mcq_5	0.0	0.0	0.03	0.03	0.03	0.03	0.03
mcq_6	0.09	0.0	0.11	0.1	0.1	0.1	0.1

3) Similarities measurement using Cosine Similarity: Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. This metric is a measurement of orientation and not magnitude, it can be seen as a comparison between documents on a normalized space because were not taking into the consideration only the magnitude of each word count (TF-IDF) of each document, but the angle between the documents. To assign a numeric score to a document for a query, the model measures the similarity between the query vector and the document vector. The similarity between two vectors is once again not inherent in the model. Typically, the angle between two vectors is used as a measure of divergence between the vectors, and cosine of the angle is used as the numeric similarity [23]. Cosine has the nice property that it is 1.0 for identical vectors and 0.0 for orthogonal vector. If $\vec{D}$ is the document vector and $\vec{Q}$ is the query vector, then the similarity of document D to query Q (or score of D for Q) can be represented as:

$$sim(\vec{D}, \vec{Q}) = \cos\theta = \frac{\vec{D} \cdot \vec{Q}}{\|\vec{D}\|\|\vec{Q}\|} = \frac{\sum_{i=1}^{n} D_i Q_i}{\sqrt{\sum_{i=1}^{n} D_i^2}\sqrt{\sum_{i=1}^{n} Q_i^2}} \quad (4)$$

Cosine Similarity will generate a metric that says how related are two documents by looking at the angle instead of magnitude, like: Using the eq. 4, we can find out the similarity

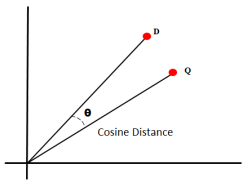


Fig. 9. Cosine distance between $\vec{D}$ and $\vec{Q}$, where $0 \leq cos\theta \leq 1$

between Query vector and other mcqs in VSM showed in table. I .The **Cos**θ values for different mcqs or docs in range of 1 (Most Similar) and 0 (less similar) and ranking(see table. II) from calculated cosine value.

TABLE II
RANKING FROM COSINE VALUE

Doc	$CosineValue$
mcq_1	0.93
mcq_3	0.81
mcq_2	0.64
mcq_5	0.32
mcq_6	0.24
mcq_4	0.00

IV. RESULT AND DISCUSSION

QB3S has been evaluated its first experiment in a multiple choice question bank(ICT) database. It contains total 1289 mcq. Information of experimental data-set is showing in

table. III.

TABLE III
EXPERIMENTAL DATA-SET

Chapter ID	Chapter Name	Total Question
c11001001	Information and Communication Technology	304
c11001002	Communication Systems and Networking	245
c11001003	Number system and digital device	273
c11001004	Web Design Contacts and HTML	166
c11001005	Programming language	152
c11001006	Database management system	149

We implied two test case with present or absent of condition and analysis the result graphically. Accuracy of searching performance, Sensitivity and specificity have been calculated from the value of True Positive(detects the condition when the condition is present), False Negative (does not detect the condition when the condition is present), True Negative(does not detect the condition when the condition is absent) and False Positive(detects the condition when the condition is absent) [24].

Test 1: Select 50 existing full or part of a mcq as query inputs from 'Communication Systems and Networking' chapter and found the maximum cosine valued mcq. Graphical result showing in Fig. 10.

Test result can be True Positive(TP) or False Negative(FN) depends on the value of cosine similarity.In case of TP, cosine value greater than 0.5, it detects the condition where query mcq belongs to this chapter is satisfied. otherwise FN.

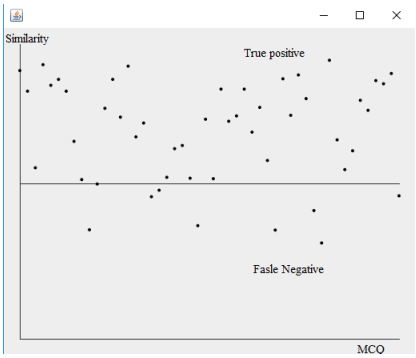


Fig. 10. Result of test 1 where 42 True Positive(TP) and 8 False Negative(FN) value was detected

Test 2: Select 50 full or part of a mcq as query inputs from other chapter except "Communication Systems and Networking" chapter(Not exist in this chapter) and find the mahapterximum cosine value holder mcq.Graphical result showing in Fig. 11.

Test result can be True Negative(TN) or False Positive(FP) depends on Cosine value. In case of TP, cosine value less than 0.5, it detects the condition where query mcq doesn't belongs to this chapter, otherwise FP.

In this experiment of QB3S, we found TP=42, FN=8, TN=43, FP=7

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP} = \frac{42+43}{42+8+43+7} = 85\%$$

Fig. 11. Result of test 2 where 7 False Positive(FP) and 43 True Negative(TN) value was detected.

$$Sensitivity = \frac{TP}{TP+FN} = \frac{42}{42+8} = 84\%$$
$$Specificity = \frac{TN}{TN+FP} = \frac{43}{43+7} = 86\%$$

V. CONCLUSION

Information retrieval techniques are effective in finding similar documents from a text database. These techniques are not efficient in finding the similarity of questions in a Question Bank developed for PBeL systems for many reasons: i) the document structure is different in nature that the conventional text documents, and ii) the contents of the different parts of the documents have different weights.

In this paper, we have analyzed the structure and weight of the different types of questions in the QB and developed an improved TF-IDF algorithm based on the weight. At the initial stage we applied NLP tools and techniques like tokenization, stemming, stopword removal in the bangla text. Handling synonym is critical in any similarity searching system. We have developed a WordNet based on hash technique technique count of of synonym in TF-IDF. We have created a the Vector Space Model using the improved TF-IDF weighted matrix. For faster and efficient access to the QB DB, we have used a B+ tree index structure.

Using the above techniques, we have developed a Question Bank Similarity Searching System(QB3S). We have applied QB3S in a real life dataset emerging from the PBeL systems for ICT course of HSC level in Bangla. We have achieved an accuracy level of 85% for similarity search in the QB.

REFERENCES

[1] G. M. M. Bashir, A. Latiful Haque, and B. Chandra Dev Nath, E-learning of php based on the solutions of real-life problems, vol. 3, 12 2015.

[2] Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto. Modern information retrieval. Vol. 463. New York: ACM press, 1999.

[3] Chowdhury, Gobinda G. "Natural language processing." Annual review of information science and technology 37.1 (2003): 51-89.

[4] F. Alam, S. Habib, and M. Khan, Text normalization system for bangla, tech. rep., BRAC University, 2008.

[5] J. Pustejovsky, The generative lexicon, Computational linguistics , vol. 17, no. 4, pp. 409441, 1991.

[6] S. Das and P. Mitra, A rule-based approach of stemming for inflectional and derivational words in bengali, in Students Technology Symposium (TechSym), 2011 IEEE , pp. 134 136, IEEE, 2011.

[7] A. Das and S. Bandyopadhyay, Morphological stemming cluster identification for bangla, Knowledge Sharing Event-1: Task , vol. 3, 2010.

[8] G. K. Saha, Parsing bengali text: An intelligent approach, Ubiquity , vol. 2006, no. April, p. 1, 2006.

[9] Ramaswamy, Sridhar, and Paris C. Kanellakis. ”OODB indexing by class-division.” ACM SIGMOD Record. Vol. 24. No. 2. ACM, 1995.
[10] Aronovich, Lior, and Israel Spiegler. ”CM-tree: A dynamic clustered index for similarity search in metric databases.” Data Knowledge Engineering 63.3 (2007): 919-946.
[11] N. Wang, P. Wang, and B. Zhang, An improved tf-idf weights function based on information theory, in Computer and Communication Technologies in Agriculture Engineering (CCTAE), 2010 International Conference On , vol. 3, pp. 439441, IEEE, 2010.
[12] Ramos, Juan. ”Using TF-IDF to determine word relevance in document queries.” Proceedings of the first instructional conference on machine learning. Vol. 242. 2003.
[13] Turney, Peter D., and Patrick Pantel. ”From frequency to meaning: Vector space models of semantics.” Journal of artificial intelligence research 37 (2010): 141-188.
[14] A. Habib and A. L. Hoque, Towards mobile based e-learning in bangladesh: A frame- work, in Computer and Information Technology (ICCIT), 2010 13th International Con- ference on , pp. 300305, IEEE, 2010.
[15] A. Hoque, M. M. Islam, M. I. Hossain, and M. F. Ahmed, Problem-based e-learning and eval-uation system for database design and programming in sql, International Journal of E-Education, EBusiness, E-Management and E-Learning-IC4E , pp. 537542, 2013.
[16] A. S. M. L. Hoque, G. M. M. Bashir, and M. R. Uddin, Equivalence of problems in problem based e-learning of database, in Technology for Education (T4E), 2014 IEEE Sixth International Conference on , pp. 106109, IEEE, 2014.
[17] G. M. M. Bashir and A. S. M. L. Hoque, An effective learning and teaching model for programming languages, Journal of Computers in Education , vol. 3, no. 4, pp. 413437, 2016.
[18] Pollari-Malmi, Kerttu, and Eljas Soisalon-Soininen. ”Concurrency control and i/o-optimality in bulk insertion.” International Symposium on String Processing and Information Retrieval. Springer, Berlin, Heidelberg, 2004.
[19] Bruso, Kelsey L., and James M. Plasek. ”Dynamic preconditioning of A B+ tree.” U.S. Patent No. 7,809,759. 5 Oct. 2010.
[20] Davis, Mark, and L. Iancu. ”Unicode text segmentation.” Unicode Standard Annex 29 (2012).
[21] Duerst, Martin. ”The properties and promises of UTF-8.” Proc. 11th International Unicode Conference, San Jose. 1997.
[22] Lee, Dik L., Huei Chuang, and Kent Seamons. ”Document ranking and the vector-space model.” IEEE software 14.2 (1997): 67-75.
[23] A. Singhal et al. , Modern information retrieval: A brief overview, IEEE Data Eng. Bull. , vol. 24, no. 4, pp. 3543, 2001.
[24] D. M. Powers, Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation, 2011.