ISSN:0975-9646
V.D.V.Naga Lakshmi et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (4) , 2015, 4096-4101

XML Based Non Redundant Distributed Question Bank Generation

[1]V.D.V.Naga Lakshmi,[2]CH.Anupama, [3]D.Haritha

Dept of CSE, S.R.K.InstituteOf Technology,
Enikepadu, Krishna dist, AP, India

***Abstract* - The number of institutions affiliated to a university increasing enormously these days, thus demands the preparation of highly distributed, fast, reliable, secured and confidential question papers. Traditional way of question paper generation is time consuming, less efficient and less transparent. This paper aims at an automated question paper preparation in large scale distributed question bank environment. For maintaining security and confidentiality instead of a single faculty preparing the whole question paper, this system allows many faculties to prepare the question papers based on a predefined syllabus copy. From the distributed question bank the questions must be intelligently or randomly selected such that it contains no repetitive questions.Similarly from the question bank we need to prepare question paper based on the changes in the syllabus. In these applicationsquestion classification plays a major role and a new method is proposed in this paper. This method is based on XML based approach and is to augment the questions with syntactic and semantic analysis knowledge from the predefined syllabus.**

***Keywords*- Question classification, Stop words removal, Question Papers, Question bank Generation, XML tree.**

I INTRODUCTION

Examinations are important activities organized by educational institutions to evaluate student performance.With the increasing number of educational institutions and hence the assessment tests, the demand for automated and dynamic content generation systems is ever increasing. The automated content generation systems are more advantageous over conventional methods of manually generating assessment tests, as they are less error prone, secure and offer faster processing capabilities. The number of institutions affiliated to a university increasing enormously these days, thus demands the preparation of highly distributed, secured and confidential question papers. The Distributed Question Bank contains a pool of questions given by many experts related to a particular domain. If a common question paper is to be given to all the students of the institutions affiliated to a university, the administrative staff can use this question bank to automatically generate the question paper comprising of random selected questions with no repetition. Automatic question paper generation refers to extraction of the questions are extracted from the question bank automatically and a question paper that can meet the requirements of question types, difficulty levels and score distribution is constituted. As this kind of generation of the question paper takes inputs from all the experts, we can expect the equal weighted, transparent, distributed and confidential paper. Even when the syllabus of a subject changes, the question bank can be modified to a little extent and can be still used to generate the question paper according to the new syllabus. While generating the question paper, the questions must be covered from the entire syllabus and also based on the weight-age defined by the predefined rules.Two kinds of scenarios exist in these applications. First is while uploading the question paper itself the faculty categorizes the questions into different categories. The system has to identify similar questions and duplicate questions are to be eliminated from each category and based on rules defined manually, the question paper containing random questions is to be generated. In the second scenario all the questions available in the question bank can be classified into different categories based on the prior information given to the system in the form of syllabus. From each category the duplicate questions are to be identified and are to be eliminated and then based on the predefined rules question paper is to be generated.

The paper is structured as follows. Section 2 gives the overview of related work. Section 3 describes the system design methodology used to generate distinct questions in detail. It also describesquestion paper, syllabus samples that we use for our experiments and their keyword annotations. This section also demonstrates how the keyword annotations help in automatic duplicate question identification and removal. It explains in detail how we find semantically relatedquestions. In section 4, the results and discussions are given. Section 5 is the conclusion of thepaper.

*II.*RELATED WORK

[1] Dan Liu suggested adopting group intelligent searching method-ant colony algorithm in automatic test paper generation. First, they built mathematical model of constraints according to the requirements of test papers, and by using the ant colony algorithm, the optimal solution of grouping was obtained.The shuffling algorithm in an Automatic Generator Question paper System (GQS) as a randomization technique for organizing sets of exam paper is used by Jamail& Abu BakarMd Sultan[8]. The results indicate shuffling algorithm could be used to overcome randomization issue for GQS. In the field of Automatic Question Generation (AQG), most systems ([5]Heilman& Smith 2009;[9]Rus, Cai, &Graesser, 2007; [10]Wolfe, 1976) focus on the text-to-question task where a set ofcontent-related questions are generated based on a given

text. Usually, the answers to thegenerated questions are contained in the text. For example, Heilman and Smith presented anAQG system to generate factual questions with an 'overgenerating and ranking' strategy based onNatural Language Processing (NLP) techniques, such as Name Entity Recognizer and WhmovementRules, and a statistical ranking component for scoring questions based on features.The target applications of such systems are reading comprehension and vocabulary assessment.These are significantly different from academic writing, which is our target application.Non-statistical classification uses hand-crafted rules to identify question classes.One such effort is by Pasca&Harabagiu [2]. These rules are very efficient.However, they are not practical in identifying questions with various sentencestructures. It was also very time consuming to make each rule for every possibletype of question.Rules were also used in Silva, *et al.* [7]'s efforts through direct matching forspecific questions as well as by identifying head words that are then mappedinto the question classification by using WordNet [3]. The rule-based questionclassifier was then enhanced using a SVM resulting in an improvement inclassification compared to the stand-alone rule-based classifier.

III. SYSTEM DESIGN

This section explains the proposed system architecture and the methodology. Our proposed system shown in Figure1 is based on pipeline architecture comprising three main stages: in the first stage an input Question documentation is taken and pre-processing is done, in the second stage Syllabus Preprocessing, and Question Classification, Duplicate Questions Elimination is done and Distinct Questions identification is the final stage.The designed algorithm is implemented into three steps.

- In the first step, the prescribed syllabus copy is preprocessed and extracts the keywords from syllabus copy into an XML file. The index from the prescribed text book given in the syllabus is also preprocessed and the keywords are extracted into another XML file.
- In the second step, the Input document i.e. question paper is selected and preprocessed. This involves segmentation into questions and by eliminating the stop words and extracts the keywords. Step two is repeated for all question papers.
- In the third step, compare each question keywords based on the keyword annotations present in the syllabus keyword XML file/ text content XML file. Based on keywords categorize each question to the corresponding Unit. Check for the redundant questions in each Unit and generate the distinct questions.

A.Question Paper Pre-processing

Question paperpreprocessing is done as follows.
First step is to segment the question paper into questions.
Second step is to tokenize the questions with a string tokenizer.We normalize tokens by removing noise terms and stop-words. We usedEnglish language-dependent stop-word lists for this purpose.
Third step is to identify question context keywords (when, how, what, explain, discuss,) that are subset of stop-words that help in identifying the context of the question.
Next step is to identify the keywords that help in classifying the question to a particular unit/class.
Question paper ID, Subject ID and Question ID is a composite key that uniquely identifies the question along withFeature vector comprising of annotations of Keywords and question context keywords and is stored in the database.
Annotation of a keyword is taken from the hash table derived from the syllabus preprocessing. While storing the annotations they are stored in ascending order that helps in easy identification of duplicate questions.

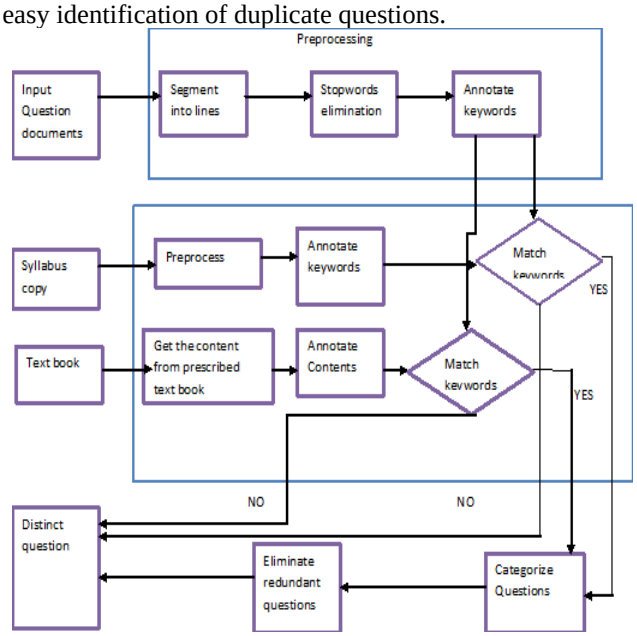

Fig1.Distributed Question Bank Generation System

```
Algorithm Qp_processing(Questionpaper QP,
SubcodeSub_Code, Dictionary Sub_Code_value, Stopword SW)
{
segment QP into Questions QS[]
        for each qsin QS do
i←0
        segmentqs into wordsetQPWord
if (QPWord[i] is digit)
then
i←i+1
else if (QPWord[i] is alphabet and QPWord[i+1] is')')
then i←i+1
else
for each QPWord  until the QPWord is '.'or '
if(QPWord[i] Є SW)
addQPWord[i] to set QS(stopwords
else
addQPWord[i] to set QS(keywords)
end if
end if
end for
```

Fig2.**Algorithm for Question Paper generation**

B.Syllabus Pre-processing:

Authorized administrative staff input the default syllabi for all the subjects received from the heads of the academic units, which our systemstores in an XML file. The advantage of using the XML format instead of a relational

database is the reduction in system development cost. Briefly, XML provides a way to describe structured data. That is, an XML file is a plain-text file that uses XML tags to define the logical structure of the document in a hierarchical fashion. XML uses a set of tags to delineate data elements, each of which comprises subelements and attributes. These subelements, in turn, contain other subelements, and so on forming the hierarchy.

The given syllabus comprises of different units.We use node index based prefix scheme XML format to represent the each keyword in the syllabus uniquely. Node index is a value that represents the location of a node in the XML tree. Node index is a number helps to find the node's parent, ancestors, children and siblings. In Prefix based node indexing, index value contains two parts i.e prefix value and node value and is separated by dot. The prefix value is the parent code and the node value represents the location of the node. To allow structural query each node i.e the keyword is given unique index value. Each entry in the dictionary is represented with the keyword and its index value.

Thus structural query about keyword can be accessed from dictionary without referring to the actual document i.e. syllabus copy.Each keyword is uniquely annotated based on the XML format.

```
AlgorthmSyllabus_Pre_Process(SubcodeSub_Code, Syllabus S,
Dictionary Sub_Code_value)
{
segment  S into wordset W
unit_count←0
for all words w[i] € W do
if (w[i] begins with "Unit")
then
subcount←0,  count←1,  increment unit_count.
else if (w[i] is '--')
subcount←0,
for each w[i] do
incrementsubcount,
sub_code_value(w[i])←unit_count.count.subcount;
else
sub_code_value(w[i])←unit_count.count;
end if
end for
}
```

Fig3.**Algorithm for syllabus pre-processing**

In the similar manner prescribed text book chapter index is used to extract the keywords and another XML file is generated.

UNIT-I :OSI, TCP\IP, other network models - Network Topologies, WAN, LAN, MAN ;

UNIT-II :Transmission media, twisted pair, wireless switching, encoding, asynchronous communications, Narrow band ISDN, broad band ISDN, ATM

Fig4: Sample syllabus content

```
<?xmlversion="1.0"encoding="UTF-8"
standalone="no" ?> - <syllabus>
-<A>/*A is used for representing UNIT-I*/
<A.a>OSI</A.a>
<A.b>TCP\IP</A.b>
<A.c>other network models</A.c>
<A.c.a>Network Topologies /A.c.a>
<A.c.b> WAN/A.c.b>
<A.c.c> LAN </A.c.c>
<A.c.d>MAN</A.c.d>
</A>
- <B>
<B.a>Transmission</B.a>
<B.b>media</B.b>
<B.c>twisted pair</B.c>
<B.d>wireless switching </B.d>
<B.e>encoding</B.e>
<B.f>asynchronous communications </B.f>
<B.g>Narrow band ISDN </B.g>
<B.h>broad band < ISDN /B.h>
<B.i>ATM</B.i>
</B>
```

Fig 5: XML File generated for the syllabus

C.Question Classification:

Many supervised learning approaches have been proposed for question classification ([11]Liand Roth, 2002; [6]Blunsom et al., 2006; [12]Huang et al., 2008). These approaches mainly differin the classifier they use and the features they extract.Most of the studies assume that a question is unambiguous, i.e., it has only one classand therefore assign the question to the most likely class. Some other studies ([4]Li andRoth, 2002, 2004) on the other hands, have more flexible strategy and can assign multiplelabels to a given question.If the set of possible classes represented by $C = \{c_1, c_2, ..., c_m\}$ then the task of aquestion classifier is to assign the most likely class $\mathbf{c_i}$to a question $\mathbf{q_j}$if the question canonly belong to one class. If a question can belong to more than one class then the decisionmodel will be different.

Question classification in this application is different from the common text categorization task in the sense that questions are relatively short and contains less word-based information compared with classification of the entire text. This work presents a machine learning approach to this task. Our approach is to augment the questions with syntactic and semantic analysis as input to the text classifier. Machine Learning based classifiers typically take as input a feature-based representation of the domain element (e.g., a question). For the current task, a question sentence is represented as a vector of features and treated as a training or test example for learning. The mapping from a question to a class label is a linear function defined over this feature vector.

Question Classification is important in handling Community based Distributed Question Banks or Question Answering systems. Given a question, Question Classification module is responsible for matching the question to the semantic class/unit in the syllabus of the respectivesubject. This information will be used later to reduce the searching space.The idea is to focus the question classification only on those keywords related to the syllabus prescribed.Each question is represented by a feature vector. Each element of the vector is the annotation

of the keyword and is obtained from the corresponding keyword present in the XML notation of the syllabus.Based on the root tag of the majority of the annotations present in the feature vector of the question, the question is classified into the respective class/unit.

```
Algorithm Distinct_questions(Questionpaper QP,
SubcodeSub_Code, Dictionary
Sub_Code_value,KeywordsKW,Annotations  A){

Select questions from database D

Keywords Є Words[Qp]+words[SP]

for all words w[i] Є W do

if (w[Qp] compares with W[sp]) then do
                W[Qp].indexof(W[SP]>0

Annotations Є(A.a,A.b,----------H.Z)
else if(Annotations A selected from database)
then do
Annotations Є(Empty)
for all Annotations [i] Є A  do
add Annotations[i] to Array
Annotations_similar  Count← 0
Questions_Distinct add to set Question paper[QP]
else
Annotations_similar  Count← 1
end if
end if
end for
}
```

Fig6.**Algorithm for generating distinct questions**

Duplicate Questions Elimination: Once the classification of the questions of all the questionnaires completes, each class/unit is considered for the duplicate questions removal in that class. For duplicate removal feature vectors of questions is considered. If a feature vector of one question is a subset of feature vector of the other question, question having less feature vector size is considered to be duplicate.For deleting the duplicate questions , the keyword set Ks is used to evaluate the similarity measure. Ks Keyword set is the collection of three possible keyword sets such as common keyword set {Kc} (common in both questions), keywords in Q1 question alone but not in Q2 question, {Kq1} and keywords in Q2 question which are not in the Q1 question, {Kq2}, represented in Eq (1).

$K_s=\{\{K_c\}+\{K_{q1}\}+\{K_{q2}\}\}$

If the size of $K_c$is equal to or more than the size of K_s, then the question having larger size is considered as the distinct question and the question having smaller keyword size is identified as duplicate question and is removed.

IV. EXPERIMENTS AND RESILTS

For experimentation we took Computer Networks question papers and classified each question. The following tables specify the questions and their respective annotations using syllabus and text book index taken from two question papers. The last column indicates the unit to which that question was categorized. It was observed that few questions were not categorized.

S.No	Question Description	Annotations from the syllabus	Annotations from the textbook index	Unit catogerized
1	Differentiate between OSI and TCP\IP reference models.	A.a,A.b		Unit 1
	Explain about the design of ARPANET .	Nil	Nil	
2	What is guided media ? Explain about various types of guided media available .	Nil	a.a	Unit 1
3	What is framing ? Explain its purpose in data link layer .	C.b		Unit 3
	What is sliding window protocol ?Explain .	C.h		Unit 3
4	What are the different collision-free protocols ? Explain	C.a		Unit 3
	How is Manchester encoding implemented ? Explain		g.b	Unit 7
5	How do you implement hierarchical routing ?	E.b.c		Unit 5
	What are the advantages and disadvantages of flooding	E.b.b		Unit 5
6	How do you implement broadcast routing ? Explain.	E.b.d, F.b		Unit 5,6
	How is congestion prevented in different layers ? Explain		a.c	Unit1
7	What are the different transport primitives ? Explain		c.a	Unit 3
	Explain about AAL layer protocol .		e.d	Unit 5
8	Define WWW . Explain about dynamic web documents		f.d,h.b	Unit 6.8
	How do you provide Network Security in application layer	H.a		Unit 8

Table1. Question paper 1

S.No	Question Description	Annotations from the syllabus	Annotations the textbook index	Unit categorized
1	What is OSI reference model ?Explain .	A.a,A.b		Unit 1
	What is 802.11 ? Explain	Nil	Nil	
2	What is guided media ? Explain about various types of guided media available.	Nil	a.a	Unit 1
3	What are the different collision-free protocols ?Explain .	C.b		Unit 3
	How is Manchester encoding implemented ? Explain	B.d		Unit 3
4	Explain about various design issues of DLL .	C.a		Unit 3
	What is HDLC ?Explain .		g.b	Unit 7
5	How do you implement hierarchical routing ?	E.b.c		Unit 5
	What are the advantages and disadvantages of flooding ?	E.b.b		Unit 5
6	How do you implement broadcast routing ?Explain .	E.b.d, F.b		Unit 5,6
	How is congestion prevented in different layers ? Explain.		a.c	Unit1
7	What are the different transport primitives ?Explain .		c.a	Unit 3
	Explain about AAL layer protocol .		e.d	Unit 5
8	Define WWW . Explain about dynamic web documents		f.d,h.b	Unit 6.8
	How do you provide NetworkSecurity in applicationlayer?	H.a		Unit 8

Table2. Question paper 2

Unit	Qp_ID	Q-ID	Description of the question
1	CN_2	CN1-1	Differentiate between OSI and TCP\IP reference models
	CN_1	CN2	What is guided media? Explain about various types of guided media available.
	CN_1	CN6_2	How is congestion prevented in different layers? Explain.
2	CN_1	CN2-2	How is Manchester encoding implemented? Explain
3	CN_1	CN4-1	Explain about various design issues of DLL
	CN_2	CN3-2	What is sliding window protocol? Explain
	CN_1	CN7-1	What are the different transport primitives? Explain.
5	CN_1	CN6-1	How do you implement broadcast routing? Explain.
	CN_2	CN5-2	What are the advantages and disadvantages of flooding?
	CN_1	CN7-2	Explain about AAL layer protocol.
6	CN_1	CN6-1	How do you implement broadcast routing? Explain.
7	CN_1	CN4-2	What is HDLC? Explain
8	CN_1	CN8-2	How do you provide Network Security in application layer
	CN_1	CN8-1	Define WWW. Explain about dynamic web documents

Table3. Distinct Questions from two Question papers

a). Distinct Unit wise Questions

Unit wise distinct questions are selected from the syllabus paper by comparing different question papers.

These are the questions which are categorized from syllabus copy which matched questions need to be deleted.

The questions which are not matched either from syllabus copy or textbook index are given below.

Table4. Unclassified Questions from two Question papers

QP_ID	Q_ID	QUESTIONS
CN_1	CN1- 2	Explain about the design of ARPANET .
CN_2	CN1-2	What is 802.11 ?Explain .

So these above questions need to be treated as unmatched questions from syllabus and textbook and therefore we need to avoid these questions for further comparison.
In the same way we are performing for 5 different subjects and finally categorizing distinct questions from individual subject as unitwise.

b). Performance Metrics in Question Classification
Typically, the performance of a question classifier is measured by calculating the accuracyof that classifier on a particular test set. The accuracy in question classification is definedas given below.

$$\text{Accuracy} = \frac{\text{no. of Correctly Classified Samples}}{\text{Total no. of Tested Samples}}$$

We have taken 5 subjects JAVA,WEB TECHNOLOGIES, DATA STRUCTURES, DBMS and COMPUTER NETWORKS each with 10 question papers. In each subject 6 question papers are taken into the Training set and 4question papers into Test set. The following table gives Accuracy, Precision and Recall measures of Training and Test sets.

Set	Accuracy
Training set	0.8904
Test Set	0.8672

Table5: performance of the proposed system

c). Discussions:
This work mainly aims to improve the efficiency involved in duplicate question detection as well as categorization of questions to different units. It facilitates the collection of different questions pertaining to each unit and hence simplifies the task of question paper preparation even if the syllabus changes.
Algorithm performance i.e duplicate question elimination is quite good. If the questions are not classified unitwise every question is to be compared with every other question in the question papers. If there are ***m*** question papers each with ***n*** questions each question must be compared with ***(m-1)*n*** questions. But in this proposed method as a first step all questions are categorized into different units and then each question is to be compared with the remaining questions of the same unit. Hence the time complexity reduces by a factor of ***l***, where ***l*** is the number of units.

V. CONCLUSIONS

In this paper we proposed new approach for XML based Non-Redundant Distributed Question Bank Generation to meet the requirements of the educational institutions.An automated question paper preparation in large scale distributed question bank environment is presented. This algorithm is more efficient in terms of time complexity and supports flexibility. As and when there is change in the syllabus the question bank need not be thrown away.

REFERENCES

[1] Dan Liu Jianmin Wang LijuanZheng AutomaticTest Paper Generation Based on AntColony Algorithm JOURNAL OF SOFTWARE,VOL. 8, NO. 10, OCTOBER 2013 p2600-2605

[2] Dan Moldovan, SandaHarabagiu, Marius PascaRoxana Girgu, " The Structure and Performance ofan Open-domain Question Answering System",Proceedings of the 38th Annual Meeting on Association for Computational Linguistics Hon Kong, 2000, PP. 563-570.

[3] Fellbaum, C. (Ed.). (1998). WordNet: An electroniclexical database. MIT Press

[4] Li, X., K. Small, and D. Roth. 2004. The role of semantic information in learning question classifiers. In Proceedings of the First Joint International Conference on Natural Language Processing

[5] Michael Heilman and Noah A. Smith (2009) Good question! Statistical ranking for questiongeneration.*In Proceedings of The 2010 Annual Conference of the North American Chapter ofthe Association for Computational Linguistics*, pages: 609-617, Stroudsburg, PA.

[6] Phil Blunsom, KrystleKocik, and James R. Curran.2006. Question classification with log-linear models.In SIGIR '06: Proceedings of the 29th annualinternational ACM SIGIR conference on Research and development in information retrieval, pages 615–616, New York, NY, USA. ACM.

[7] silva,J.a., Coheur L., Mendes, A. &Wichert, A. From Symbolic to Sub- Symbolic Information inQuestion Classification, Artificial Intelligence Review,35, pp. 137-154, 2011.

[8] Shuffling Algorithms for Automatic Generator Question Paper System Nor ShahidabtMohdJamail& Abu BakarMd Sultan :Computer and Information Science vol.3 No. 2; May 2010

[9] VasileRus, ZhiqiangCai, Arthur C. Graesser Experiments on Generating Questions About Facts

[10] Wolfe, J. 1976. Automatic question generation fromtext-an aid to independent study.In Proceedings of ACM SIGCSE-SIGCUE.

[11] Xin Li and Dan Roth. Learning question classifiers.In Proceedings of the 19th international conference on Computational linguistics, pages 1–7,Morristown, NJ, USA, 2002. Association for Computational Linguistics.doi.

[12] Z.Huang, M. Thint and Z. Qin. 2008.Question Classification Using Head Words andtheir Hypernyms. Proceedings of the conference on Empirical Methods in Natural Language Processing, pp. 927-936.