WordNet and Cosine Similarity based Classifier of Exam Questions using Bloom's Taxonomy

http://dx.doi.org/10.3991/ijet.v11i04.5654

K. Jayakodi[1], M. Bandara[2], I. Perera[2], and D. Meedeniya[2]
[1] Wayamba University, Kuliyapitiya, Sri Lanka
[2] University of Moratuwa, Moratuwa, Sri Lanka

***Abstract*—Assessment usually plays an indispensable role in the education and it is the prime indicator of student learning achievement. Exam questions are the main form of assessment used in learning. Setting appropriate exam questions to achieve the desired outcome of the course is a challenging work for the examiner. Therefore this research is mainly focused to categorize the exam questions automatically into its learning levels using Bloom's taxonomy. Natural Language Processing (NLP) techniques such as tokenization, stop word removal, lemmatization and tagging were used prior to generating the rule set to be used for this classification. WordNet similarity algorithms with NLTK and cosine similarity algorithm were developed to generate a unique set of rules to identify the question category and the weight for each exam question according to Bloom's taxonomy. These derived rules make it easy to analyze the exam questions. Evaluators can redesign their exam papers based on the outcome of this classification process. A sample of examination questions of the Department of Computing and Information Systems, Wayamba University, Sri Lanka was used for the evaluation; weight assignment was done based on the total value generated from both WordNet algorithm and the cosine algorithm. Identified question categories were confirmed by a domain expert. The generated rule set indicated over 70% accuracy.**

***Index Terms*—Question classification, Teaching and Supporting Learning, Bloom's taxonomy, Learning Analytics, Natural Language Processing, Cosine similarity**

I. INTRODUCTION

Assessments are the systematic collection, review and use of information in educational programs, undertaken for the purpose of improving learning outcomes and student development. Effective style of questions plays an important role in learner assessment. Through the art of thoughtful questioning teachers can extract not only factual information, but also help learners in connecting concepts, making inferences, increasing awareness, encouraging creative and constructive thoughts. There are different taxonomies that have been developed to identify the level of the assessment being practiced such as Bloom's [1] and SOLO [2]; they are useful to identify the levels of the questions. While questions can be given throughout the course, mid semester and the end semester exam questions often carry a considerable weight for the overall assessment. When questions are prepared, there should be an effective balance between questions that assess the high level of learning and questions that assess the basic level of learning [3]. Often the exam questions used to assess the level of the university students are at low cognitive levels [4].This may happen due to the lack of tools available to evaluate the exam papers and limited knowledge of the examiners about existing learning taxonomies and how their exam questions fit into the taxonomies. Poorly designed assessments usually fail to examine the achievement of course outcomes, which can lead to low quality graduates who do not fit with the employer expectations. Ultimately this can fail the goals of examination and result in degradation of the standards of degree program.

An exam question often falls into more than one level of assessment categories of a given taxonomy. It is difficult to categorize exam questions and even more difficult to identify the portion each taxonomy level of assessment belongs to. Therefore this research was carried out to generate an appropriate rule set using NLP and WordNet similarity algorithm, which was then combined with cosine similarity algorithm to assign the weight for each category of the question, according to Bloom's taxonomy.

The paper is arranged as follows: Section II presents the related literature on educational taxonomies and natural language processing techniques used for exam question evaluation. Section III elaborates the research methodology and Section IV presents the results and analysis. Sections V, VI and VIII discuss research contributions and conclude.

II. LITERATURE REVIEW

A. Educational Taxonomy

Educational taxonomies can be used to measure the achievement of course objectives. Taxonomy not only does explain about the topics to be covered in a course but also help to understand the depth of each subject topic [5]. Once we identify the relationship of the chosen level of the taxonomy and the course outcome we can assess students at the chosen level through a suitable choice of questions [6]. Educational researchers have developed several taxonomies useful for the development of assessments, learning outcomes and educational resources. Out of those Bloom's taxonomy [1] is in the foreground. In his study Bloom identified six main categories within cognitive domain. It starts from the lowest level (Fig. 1) and increasingly moves to complex and abstract higher levels. Bloom's categories were considered as the degree of difficulty to achieve learning outcomes. The highest order is classified as Evaluation and the lowest is classified as Knowledge level. It is expected that lowest level should be mastered before moving into higher levels. Anderson *et al* [7] have already improved the noun list of Bloom's taxonomy into a verb list (Table 1). Apart from that Anderson identified the level of knowledge which makes Bloom's levels into a Matrix. For example, factual

knowledge, Conceptual knowledge, Procedural knowledge and Metacognitive knowledge were identified as the knowledge level dimensions [8].

Structure of observed learning outcome [SOLO] taxonomy [2] is another model, which concerns about student understanding of the subject. SOLO provides a simple, reliable and robust model for three levels of understanding such as surface, deep and conceptual. It is up to the examiner to define the type of content in the answer that is expected. There are five main stages to be followed sequentially: Pre-structural, Uni-structural, Multi-structural, Relational, and Extended Abstract. The lower level of SOLO taxonomy is important to focus on individual items of what is being assessed. The higher level is more concerned with the broader range of elements or attributes to be examined.

B. Educational taxonomy and NLP for exam evaluation

Bloom's taxonomy has been widely researched for student assessments efficiency. Jerzy *et al* [9] analyzed the contents of laboratory exercises and lab tests to identify the knowledge level. Bloom's taxonomy of learning outcomes has been applied to classify the exam questions. In general, disregarding the Bloom's pyramid structure was found as the leading source of laboratory failure [9]. Turkish high-school physics examination and university entrance examination questions were examined according to Blooms' taxonomy to identify the assessment levels of those exam papers. It was revealed that university level questions belong to higher levels whereas school questions belong to lower levels of the taxonomy [10]. Developing questions based on Bloom's hierarchy would be a productive way of ensuring the expected quality of student learning achievement. Higher skewness towards lower levels of Bloom's taxonomy can lessen the skill differentiation between a graduate and a first year undergraduate. Thompson *et al.* [11] noticed that in case of science courses there is a significant disagreement between academics in assigning questions into categories. For example, typical classroom tutorial problem 'to calculate' can fall into understanding, application or synthesis categories depending on the context. Therefore this research tries to provide an appropriate solution to assign the weights for each question using revised Bloom's taxonomy [7].

Main purpose of NLP is to convert human language in to a formal representation that computers can understand. NLP is used successfully in many fields such as information extraction, machine translation, text summarization, search and human computer interfacing. These research areas have used statistical NLP due to the easiness of interpretation [12]. Sentimental analysis is a field of NLP, which is used to identify and extract subjective information from sources. NLP preprocessing techniques such as tokenization, stemming, tagging, lemmatization, chunking and parse generation were used in education domain prior to applying semantic analysis techniques. Learning Management System (LMS) support for users, answering question and assessment generation, language learning and course preparation, subject evaluation, and exam paper evaluation are few areas of education that NLP was used extensively [13]. Most of the question categorization techniques depend on the usage of NLP preprocessing techniques. Question categorization methodologies such as use of regular expression, term weighting [14], Support Vector Machine (SVM) [15], and neural network techniques [16] have used NLP preprocessing techniques.

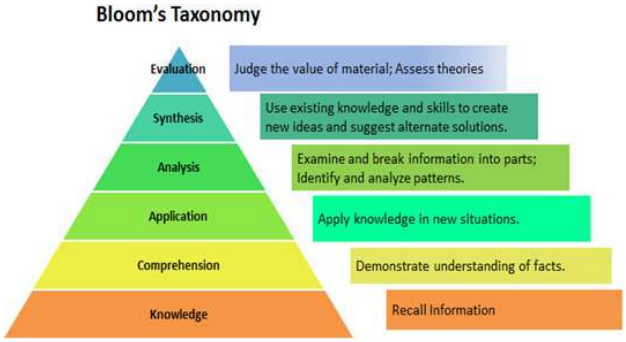

Figure 1. Bloom's Taxonomy

TABLE I.
ANDERSON'S REVISIONS ON BLOOM'S TAXONOMY

Category	Cognitive Verb list of Anderson Taxonomy	
	Description	***Verb list***
Categories	Recall or retrieve previous learned information.	defines, describes, identifies, knows, labels, lists, matches, names, outlines, recalls, recognizes, reproduces, selects, states
Understand	Comprehending the meaning, and interpretation of instructions and problems.	comprehends, converts, defends, distinguishes, estimates, explains, extends, generalizes, gives an example, infers, interprets, paraphrases
Apply	Use a concept in a new situation or unprompted use of an abstraction	applies, changes, computes, constructs, demonstrates, discovers, manipulates, modifies, operates, predicts
Analyze	Separates material or concepts into component parts	analyzes, breaks down, compares, contrasts, diagrams, deconstructs, differentiates, discriminates, distinguishes, identifies, illustrates
Evaluate	Make judgments about the value of ideas or materials	appraises, compares, concludes, contrasts, criticizes, critiques, defends, describes, discriminates, evaluates, explains, interprets
Create	Builds a structure or pattern from diverse elements	categorizes, combines, compiles, composes, creates, devises, designs, explains, generates, modifies, organizes, plans, rearranges, reconstructs

Chang [17] has extracted the verbs of a question to classify the question cognitive levels in which semantic similarity was not taken into consideration. Question categorization with just keyword mapping is not the appropriate solution for every scenario. Auto marking [18], a tool developed for a LMS was capable of marking the student answers submitted online. Student answers are often evaluated with the usage of semantic similarity algorithms available in WordNet.

1) WordNet based algorithms for semantic similarity: Semantic similarity is a way to check the similarity between documents, words and text by considering the distance between them. It is based on the likeliness of their meaning or semantic content as opposed to similarity, which can be estimated regarding their syntactical representation. It consists of a number of algorithms, which is used to measure the semantic similarity and relatedness between a pair of concepts (synsets). There are two main ways to calculate the semantic similarity between two ontologies: such as Edge-based and Node-based. Edge

based uses the edges and their type as the data source whereas Node based uses the nodes and their properties as the main data source [19]. Path similarity, Leacock-Chodorow Similarity, Wu-Palmer Similarity, The Jiang-Conrath Similarity and Lin Similarity are few of the semantic similarity algorithms. Out of those, path similarity was identified as the best algorithm in this context [20].

2) Cosine similarity of question [21]: Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not the scaler magnitude.

Two vectors with the same orientation have the cosine similarity of 1, two vectors at 90° have the similarity of 0, and two vectors diametrically opposed (180°) have the similarity of -1, independent of their magnitude. Cosine similarity (in Eq. 1, A and B are vectors of which similarity is measured) is particularly used in positive space. Student question classification was improved with the usage of cosine similarity especially in Java programming classes [22]. Intelligent tutoring dialog text classification [23] is an instance where cosine similarity was used to improve the accuracy of the classification.

$$Similarity = \frac{A.B}{\|A\|\|B\|} = \frac{\prod_{i=1}^{n} A_i B_i}{\sqrt{\prod_{i=1}^{n} A_i^2}\sqrt{\prod_{i=1}^{n} B_i^2}} \quad (1)$$

C. Related Previous Work of the Research

Related work was carried out during the initial stages of this research primarily to identify most appropriate techniques and tools for each stage of question classification. After question string extraction NLP sequence starts with tokenization; as a preliminary work a set of tokenizers: Word tokenizer, Wordpunct Tokenizer, Regexp Tokenizer, Treebank Tokenizer, and Stanford Tokenizer were evaluated [20]. When tokenizing, it is important that the tokenizer breaks the sentences into words efficiently and effectively: the tokenizer should not break a sentence into unwanted tokens such as non-alphanumeric symbols: brackets, exclamation marks, period, etc. hence the stemmers and taggers receive reduced number of tokens for tagging, stemming and lemmatization, reducing the amount of processing. In the evaluation for a sample of common strings Word tokenizer, Wordpunct Tokenizer, Treebank Tokenizer and Regexp Tokenizer resulted in 68, 61, 69 and 34 tokens respectively [20]. Therefore Regexp Tokenizer was selected, for the tokenization.

The next important analysis carried out previously in this research was to identify the most suitable tagging technique as part of the natural language processing. Part-of-speech (PoS) tagging is the process of converting a sentence that is in the form of collection of words, into a list of tuples, such that each tuple is in the form (word, tag). Many taggers were tested with the tree bank corpus for appropriateness for the research and the tagger accuracy levels against the tree bank corpus are summarized in Table II [20]. With the highest accuracy of tagging, ClassifierBasedPOSTagger was selected, as the tagging technique.

As the next step similarity algorithms were compared to identify the most suitable algorithm in the context of the question classification. Path similarity (path), Resnik Similarity (res), Wu-Palmer-Similarity (wup) and Leacock-Chodorow-Similarity (lch), Jiang-Conrath Similarity (jcn), and Lin Similarity (lin) were used with Information Content (IC) in WordNet to identify the appropriate similarity algorithm [20]. The results obtained are given in Table III. Total value of semantic similarity algorithms for every verb identified from the question with all the verbs listed in Bloom's taxonomy categories was used to identify the main category of a question. This was further validated by a domain expert. The path similarity and the Wu-Palmer-Similarity algorithms gave the highest and second highest accuracies respectively against the sample of 26 questions that were tested with each category in order to identify the best WordNet similarity algorithm. Out of 26 questions 22 questions were accurately identified with the Path similarity algorithm.

TABLE II.
ACCURACY OF NLP TAGGERS TESTED WITH TREE BANK CORPUS

Tagger	Accuracy
unigram tagger	**0.7757**
unigram tagger with Default tagger(NN)	**0.8588**
Brill tagger	**0.8829**
Tnt tagger	**0.8756**
Tnt tagger with Default tagger(NN)	**0.8924**
Tnt tagger with N=2000	**0.8765**
Wordnet tagger with backofftaggers	**0.8848**
Classifier based tagger	**0.9309**

TABLE III.
WORDNET SEMANTIC SIMILARITY ALGORITHM COMPARISON WITH HUMAN ANNOTATED COMPUTER SCIENCE QUESTIONS

Algorithm	Accuracy Level (%)
jcn_similarity (brown)	62.0%
jcn _similarity (semcor)	62.0%
res _similarity (semcor)	57.0%
res _similarity (brown)	69.0%
lin_ similarity (brown)	76.0%
lin _similarity (semcor)	76.0%
lch_similarity	76.0%
path_similarity	84.0%
wup_similarity	80.0%

Further details about each of these evaluations and their important contributions to this work are elaborated in [20]. The methodology followed for question classification according to Bloom's Taxonomy categories presented in Section III briefly refers to these steps as part of the complete process; however, detailed discussion on these processing steps are not included to avoid repetition.

III. METHODOLOGY

The main challenge we discuss in the research problem is that the exam questions are not properly categorized and correct weights are not assigned for each category in the mid or final exam questions. To address this challenge this research followed its methodology as shown in Fig. 3 using revised Bloom's taxonomy. There are few steps used to categorize the questions automatically as described below.

Following are the steps of the question processing with NLP techniques:

1) Question Extraction: Pypdf package was used to extract exam questions from PDF documents. Questions were identified with a set of regular expressions describing the text string patterns that we are interested in [20]. For example, following regular expression was used to identify the main description of the question.

r"\[Q" + str(QMain) + r"\](.*?)\[Q" + str(QMain+1) + r"\]"

The question number was given as (Q1) to QMain and QMain +1 was coded as (Q2). The expression part (.*? extracts the entire text string in between the questions.

2) NLP Processing: Once questions were stored in MySQL database, each question is then tokenized. Regexptokenizer was developed to token based on spaces within the sentences and it produces a less number of appropriate tokens to proceed with the subsequent steps in the process, as identified in previous work of this research. Then question correction was developed with Enchant to correct the word after tokenization. Lemmatization is more appropriate as identified previously in this research [20]. Unlike stemming, here we are always left with a valid word with the same meaning as in the original sentence. Wordnet Lemmatizer in WordNet always tries to find a matching valid root word. Therefore it is effective to lemmatize the word before tagging to find the verbs than applying stemming techniques [20].

Part-of-speech tagging is the process of converting a sentence, in the form of a list of words, into a list of tuples, where each tuple is of the form (word, tag). Many taggers were tested for appropriateness. Classifier based tagger was used with the tree bank corpus sentences, which resulted in accuracy of 0.9309 [20]. Since the classifier based tagger with the usage of ClassifierBasedPOSTagger has given the highest accuracy it was selected for the tagging.

3) Verb extraction: After the completion of tokenization, word correction, lemmatization and tagging, verbs were extracted for each question and stored in the database. The tag starting is matched with letters 'V', and 'W', selected as related words for Anderson taxonomy and stored in the database. Therefore VB, VBD, VBG, VBN, VBP, VBZ, WDT, WP, WP$ and WRP types of tag words were extracted [20].

Algorithm in Fig. 2 was used on several semantic similarity algorithms to identify the most suitable algorithm. The 3rd step extracts all the verbs of a question. In step 4 taxonomy word list was taken from the database for each category. After that Wordnet synsets lists were taken for each word of the taxonomy category, for each word the similarities for the verbs and the taxonomy verbs were identified.

Tag patterns for Bloom's category: Tag patterns are unique for a given taxonomy. Classifier based tagger was used to tag the identified tags for each question pattern and stored in the database. Table IV illustrates few tag patterns that were stored with their question pattern and shown for each taxonomy category.

Fig. 3 shows the high level modular architecture of the developed tool for question classification; it also summarizes the important process steps and their order of operation during the question classification process. Processing activities explained previously are carried out in each of the relevant process steps.

1. For the entire question in question list

2…..Wordlist= Apply tagger (Apply lematizer (Spellcorrector(Correct token(question))))
3…..Verb extraction
4…….Loop for all the taxonomy
5……..Loop for all the verbs
6……For verb in Verblist
7……..Loop with all the synsets
Usage of semantic similarity comparison
8…… …Feature Extraction

Figure 2. Wordnet semantic similarity algorithm

TABLE IV.
DIFFERENT QUESTION PATTERNS FOR EACH CATEGORY IN BLOOM'S TAXONOMY

Category	Question	Pattern
Kn	What does it mean What is it What is the best one Who was the	<WP><VBZ><PRP><VB> <WP><VBZ><PRP> <WP><VBZ><DT><JJS><CD> <WP><VBD><DT>
Co	Who do you think What was the main idea Can you distinguish between	<WP><VBP><PRP><VB> <WP><VBD><DT><JJ><NN> <NNP><PRP><VBP><IN>
Ap	How could you develop Judge the effort of What was the main idea	<WRB><MD><PRP><VB> <NNP><DT><NNS><IN> <WP><VBD><DT><JJ><NN>
An	How was this similar to Can you compare Can you distinguish between	<WRB><VBD><DT><JJ><TO> <NNP><PRP><VBP> <NNP><PRP><VBP><IN>
Sn	How many ways can you What would happen if	<WRB><JJ><NNS><MD><PRP> <WP><MD><VB><IN>
Ev	How would you prioritize Rank the importance of	<WRB><MD><PRP><VB> <NNP><DT><NN><IN> <WP><MD><PRP><VB>

Tag pattern Identification and cosine module: Extracted questions were broken into individual sentences. Based on the identified tag patterns grammar rules were generated for each tag pattern in Table IV (i.e., every category in Bloom's taxonomy: Knowledge, Comprehension, etc.). Parse tree was generated with regular expression parser and the tree was tested to identify that a particular tag pattern was appeared under each taxonomy category. If the tag pattern was identified then it was stored under the relevant category. Identified question tag pattern and the matching tag pattern in the database were tested to identify the cosine similarity of the pattern.

Cosine similarity of the questions and the question patterns were used to identify more features for the exam question classification. Several steps were used to develop the algorithm, related pattern of questions were identified for each category. Then a procedure was written to identify the tag pattern of all these question stems and to write it in to the database. Once the patterns were written grammar rules and parsers were generated for each tag pattern. Those parsers were used to identify the tag patterns of exam questions. Based on the tag patterns that were matched, highest match question stem was taken for each

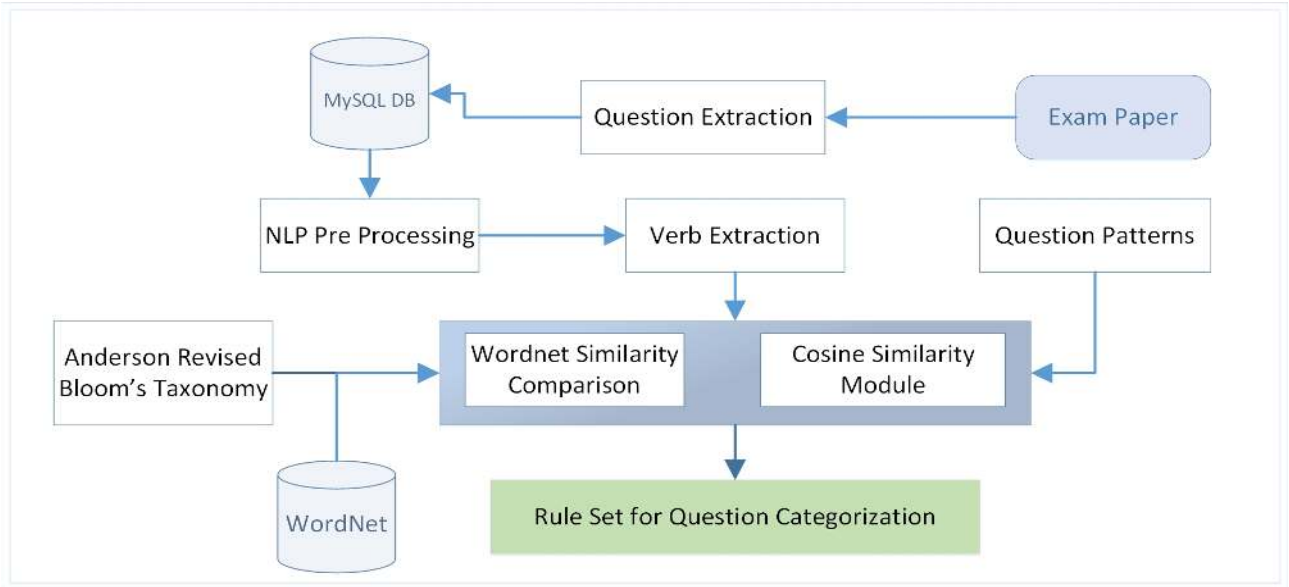

Figure 3. Proposed Architectural Diagram for question classification

taxonomy category and it was used with the original question through the cosine algorithm to check for the accuracy. Highest match value was taken for each map tag pattern and stored in the database. Apart from that ranking algorithm was used to identify the rank order of the cosine values of each taxonomy level.

As Figure 4 presents, functions were developed according to the cosine equation to find the cosine similarity of the two stems, which were taken out in the previous process step. As the first step question pattern and the related tag pattern stem that was extracted were converted to a vector as in line 1 and 2 of the algorithm shown in Fig. 4. Then the intersection was identified in between vector 1 and vector 2. Once the intersection was identified sum of all the intersections were calculated and stored in the variable Numerator. In lines 6 and 7, the sum square value of all the words in vector 1 is calculated and in lines 8 and 9 the sum value of all the words in vector 2 is calculated. Then in line 11, the shown ratio was used to calculate the cosine similarity between two patterns and the result was produced as the output. Those values given for each instance of use were stored under each taxonomy category. The probability of giving a high value for the matching pattern is higher when compared with other patterns. Generally for category where the question belongs, cosine value would be high. That feature was used to improve the exam question classification accuracy with the help of the lemma similarity and WordNet similarity.

Fig. 5 shows a working instance of the completed tool for processing exam questions and classifying them as explained above. The tool was developed as a web solution based on client-server architecture considering easy access and usability across different types of key stakeholders of assessment processes within an educational institute. Moreover it can be easily plugged into an e-Learning environment such as Moodle [24] for wider usage and convenient deployment within the higher education institution.

IV. RULE GENERATION AND WEIGHT ASSIGNMENT

As we mentioned previously the combination of WordNet similarity value and the cosine values was used

1. Vector 1 = Convert tag pattern in to a Vector
2. Vector 2=Convert question pattern in to a Vector
3. Intersection = intersection between the vector 1 and the vector 2
4. For each text in vector:
5….Numerator = the sum value of all intersection
6….For each text in vector1:
7……..sumsquareQuestion= sumsquareQuestion + the square root value of question stem
8. …For each text in vector2:
9. ……sumsquarePatternstem= sumsquarePatternstem + the square root value of pattern stem
10 denominator = squareroot(sumsquareQuestion) * square-root(sumsquarePatternstem)
11 Return numerator/denominator

Figure 4. Algorithm to generate cosine value for question stem and stag pattern stem

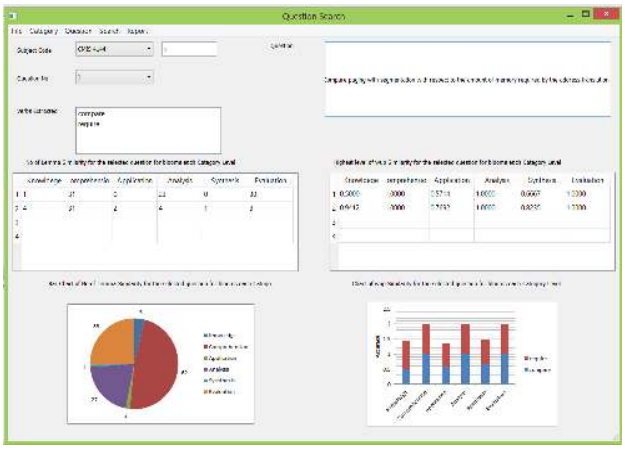

Figure 5. User Interface of the developed system

to generate and identify the rules. We have used the following assumption in order to generate the rules. Usually the proportion that each question belongs to a category can be different. It is understandable that for certain exam questions, which are complex and cover more than one category of learning levels in the taxonomy, we see potential classification of multiple categories. To resolve this, the values of the highest final value category were consid-

ered as the category for the question. Highest category was assigned 50% weight.

The six levels of Bloom's taxonomy can be grouped into two broader groups as lower order thinking level and higher order thinking level of learning [25]: Knowledge, Comprehension, and Application are considered within lower order of learning level whereas Analysis, Synthesis and Evaluation are considered within higher order of learning level. For a given question since we have already assigned 50% to the highest taxonomy category the remaining 50% is divided among the questions that are in the same broader group of leaning i.e., lower order or higher order of learning.

Based on that identification the top three highest Bloom's taxonomy levels were selected; if they are within the same group the remaining 50% of the weight was equally distributed among them. If the second and third highest taxonomy categories are not within the same group as of the highest category, then the highest category is assigned with the remaining 50% as well, i.e., 100% score. In case where only one of the 2nd or 3rd highest categories are present in the same group then remaining 50% is assigned equally as 25% to the highest and the rest to the 2nd or 3rd highest whichever is present in the group; hence the weights will be for the highest 75% (50% + 25%) and 25% for the other category. If all three highest categories are in the same group then the final weights are: for the highest 66.6% (50%+16.6%), 2nd highest 16.6% and 3rd highest 16.6%. This means for each scenario the highest applicable level of the Bloom's taxonomy will be assigned with weights 100%, 75% and 66.6%, respectively.

V. RESULTS AND DATA ANALYSIS

The questions used for evaluation are a collection of examination questions obtained from the Faculty of Applied Sciences, Wayamba University, Sri Lanka. The training set consists of 53 examination questions and the test dataset comprises of 35 questions. According to the output most of the questions were solely identified to be in one category. Highest total value of WordNet similarity value can be used to identify the main question category of the question [20]. Since action verbs are not included in some of the questions, WordNet similarity algorithm alone could not improve the correct categorization of these questions. That makes weight assignment with lemma similarity a difficult task for. Therefore Wordnet similarity algorithm, and cosine similarity algorithm was used to improve the accuracy of the question classification and weight assignments. Following are two sample questions that were analyzed based on our proposed methodology (summarized in Fig. 3).

A. *Q1: Some of architectural designs are given below. Rank the importance of those designs.*

This question belongs to a higher level of Bloom's taxonomy; a level where students evaluate or reconstruct the knowledge they have learnt before. As a result of summation value of the path similarity WordNet algorithm it was categorized as synthesis question as below. Only the given word was extracted. According to Table 2 it is evident that tag patterns were represented in each of Bloom's Taxonomy categories except for Analysis category. In Table V cosine value was calculated with (1 + *Cosine* θ). Highest matching total value was given by the Evaluation category, which is equivalent to 2.0. Final value in Table V was generated when cosine value was multiplied with the total sum of WordNet similarity values for each Bloom's taxonomy category. Based on the final value Evaluation was identified as the main category to classify the assessment level of the question.

TABLE V.
TOTAL VALUE IMPROVEMENT OF WORDNET WITH COSINE SIMILARITY OF QUESTION 1

Category	Total WordNet	Cosine Similarity	Final Value
Knowledge	2212.135	1.3535	2994.125
Comprehension	2111.182	1.3535	2857.485
Application	2462.240	1.3525	3330.180
Analysis	1997.166	1.0	1997.166
Synthesis	2700.875	1.2886	3480.348
Evaluation	2019.645	2.0	4039.290

As explained above the two groups of Bloom's levels, lower and higher levels of learning, were considered for weighting. According to the analysis this question Q1 belongs to the higher level of learning group. Categories were selected out of the higher level group in which the final value is the combined highest. According to Table 2 Synthesis category appears as the 2nd highest and it was selected for additional weight assignment. Every time the highest category is allocated 50% and the rest was divided equally among the top most category and the identified highest categories within the group. Therefore, the final weight that was assigned for Evaluation was 75% and Synthesis was 25%.

B. *Q2: Explain what is meant by an irreducible functional dependency set.*

According to the manual categorization by the expert this question belongs to Comprehension category. Once we observed the total values that were assigned with WordNet similarity algorithm it is evident that this question was categorized as Application category (Fig. 6). However once the cosine similarity was applied the question category was correctly identified as Comprehension as shown in Table 6.

When the question was checked for the tag pattern matching <WP><VBZ><VBN><IN> pattern was identified as a matching pattern. This pattern was appeared only in comprehension category. Therefore according to the cosine similarity, only the comprehension value was increased up to 2. Multiplication of cosine and the sum of WordNet similarity value was used to generate the final value in Table 6. Observing the final value, it was evident that this question was correctly categorized as a Comprehension level question.

Fig. 7 clearly depicts how the value was increased to identify the correct question category with the usage of cosine similarity. Red colour was used to display the line after using the cosine similarity and the blue colour line represents the values before applying the cosine similarity.

According to the analysis the above question belongs to the lower level type of questions as per Bloom's taxonomy. The 1st, 2nd and 3rd highest final values appeared in the lower level category; hence the weight distribution was as: Knowledge (16.66%), Comprehension 66.6% (50% + 16.66%) and Application (16.66%).

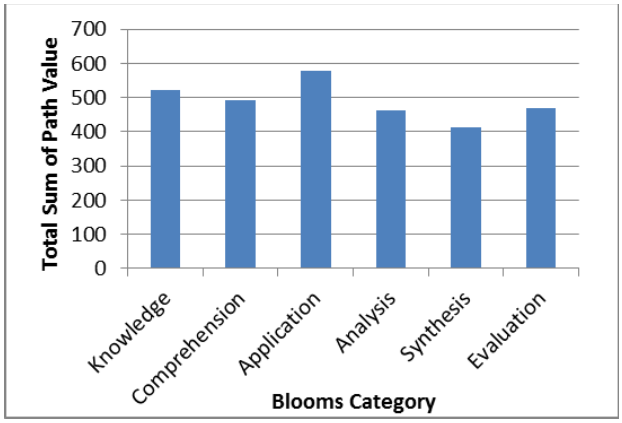

Figure 6. Sum of WordNet similarity of Anderson Taxonomy for question 2

TABLE VI.
TOTAL VALUE IMPROVEMENT OF WORDNET WITH COSINE SIMILARITY OF QUESTION 2

Category	Total WordNet	Cosine Similarity	Final Value
Knowledge	520.83	1	520.83
Comprehension	493.24	2	986.48
Application	577.65	1	577.65
Analysis	460.94	1	460.94
Synthesis	411.18	1	411.18
Evaluation	468.68	1	468.68

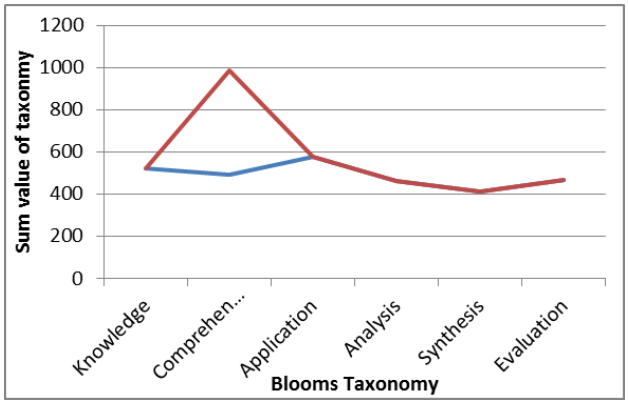

Figure 7. Sum of WordNet similarity of Anderson Taxonomy for question 2

VI. DISCUSSION

Based on above analysis rules were identified to assign the weight for the category of the question and assign the question to the top most category as well. Since there are no widely accepted weights for each category of questions by academics the automatic weight assigning provided partially accurate results. According to the generated rule set, 32 questions out of 45 were identified with correct category (accuracy is 71.0%), against the expert classification of the same.

The number of question patterns is very important to identify the question category separately. In this research around 75 question patterns have been used in each category. If the number of similar question patterns were increased then the similarity value of the similar question pattern and the gap between the highest related and unrelated question patterns can be increased.

Since human experts often are primarily able to identify only the main category of a given question, our weight assignment for each related category will give another perspective to evaluate the quality of the teaching, learning and assessment in course modules.

A. Study Limitations

The level or the depth of a question depends on the focus of the course curriculum. If a subject topic was already taught previously, when it is taught again the level of the same question goes down from previously top level to a lower level. For example a question can be considered as Evaluation in its first delivery course module, let us say, at First year but the same type of question can be considered as in the level of Knowledge when it is used for assessment within a related course module in a subsequent year (let us say Second year). Another limitation occurs when there are images as part of the question description. The information given in those images are not considered for the process we presented in this paper. Apart from that some question tag patterns were not identified correctly since the tagger is not performing with 100% accuracy. Therefore some of the tag patterns were missed and the cosine value becomes 0.0 for those categories.

VII. CONCLUSION

Our study proposed a rule based exam question classification model with the help of the NLP pre-processing techniques, WordNet similarity algorithm and the cosine similarity algorithm. WordNet similarity algorithm accuracy was mainly based on the verbs that were appeared in the question paper. Some of these verbs, which are not appeared in any level of Bloom's taxonomy made the exam questions categorized into a different category. By combining the question pattern with the cosine similarity it was evident that the question can be categorized under the correct category. Cosine similarity value categorization process can be further enhanced for accuracy by improving the question patterns for each category.

Research work presented in this paper has been successful in achieving the research objectives. Apart from that there are few important future research areas to be explored. At present all the question patterns are stored in the database. Since there are different ways to ask the same question, new method should be adopted to identify the question patterns automatically. Apart from that verb extraction process should be improved to gain higher levels of accuracy of the WordNet algorithm. Moreover it is suggested that the outcome of this research should be improved further after analyzing a large set of exam questions in different disciplines. With the generic nature of the proposed rule-based classification method, it is fair to say that the process can be extended into different educational fields and disciplines; however, appropriate modifications and extensions to the classification process need to be incorporated. Another vital future research can be to examine the efficacy of other popular educational taxonomies instead of Bloom's to classify questions.

REFERENCES

[1] B. S. Bloom, Taxonomy of Educational Objectives: The Classification of Education Goals. Cognitive Domain. Handbook 1. Longman, 1956.

[2] J. B. Biggs and K. F. Collis. Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome). Academic Press, 2014.

[3] A. J. Swart, "Evaluation of final examination papers in engineering: A case study using Bloom's Taxonomy." IEEE Transactions on Education, vol 53, no. 2, 2010, p.257-264. http://dx.doi.org/10.1109/TE.2009.2014221

[4] S. Çepni, "An analysis of university science instructors' examination questions according to cognitive levels." Educational Sciences: Theory and Knowledge, vol 3.1, 2003, pp. 78-84.

[5] J. Moon, How to use level descriptors. Southern England Consortium for Credit Accumulation and Transfer, 2002. ACM SIGCSE Bulletin vol 35, pp. 124-136

[6] R. Lister and J. Leaney, Introductory programming,criterion-referencing, and Bloom. In Proceedings of the 34th SIGCSE technical symposium on Computer science http://dx.doi.org/10.1145/611892.611954

[7] Anderson, W. Lorin, D. R. Krathwohl, P. W. Airasian, K. A. Cruikshank, R. E. Mayer, P. R. Pintrich, J. Raths, and M. C. Wittrock. "A Taxonomy for Learning,"Teaching, and Assessing-A Revision of Bloom's Taxonomy of Educational Objectives",(Eds.) Addison Wesley Longman.", 2001.

[8] M. Forehand, "Bloom's taxonomy." Emerging perspectives on learning, teaching, and technology, 2010, pp.41-47.

[9] J. Rutkowski, et al. "Application of Bloom's taxonomy for increasing teaching efficiency–case study." In Proc. of ICEE2010, 2010.

[10] A. Azar, "Analysis of Turkish high-school physics-examination questions and university entrance exams questions according to Blooms' taxonomy."Journal of Turkish Science Education vol 2.2 2005, pp. 144-150.

[11] T. Lord and S. Baviskar. "Moving Students From Information Recitation to Information Understanding-Exploiting Bloom's Taxonomy in Creating Science Questions." Journal of College Science Teaching, vol 36(5), 2007, p. 40.

[12] Christopher Manning, Hinrich Schütze, Foundation of Statistical Natural Language Processing.

[13] P. Smrž, "Integrating natural language processing into e-learning: a case of Czech." In Proc. of the Workshop on eLearning for Computational Linguistics and Computational Linguistics for eLearning, 2004, pp. 1-10. http://dx.doi.org/10.3115/1610028.1610029

[14] X. Quan, W. Liu and B. Qiu, "Term Weighting Schemes for Question Categorization", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, No. 5, 2011, pp. 1009-1021 http://dx.doi.org/10.1109/TPAMI.2010.154

[15] K. Hacioglu and W. Ward, "Question Classification with Support Vector Machines and Error Correcting Codes", Proceedings of HLT-NAACL, Vol 2, pp. 28-30, 2003. http://dx.doi.org/10.3115/1073483.1073493

[16] N. Yusof and J. H. Chai, (2010). Determination of Bloom's Cognitive Level of Question Items using Artificial Neural Network. 2010 10th International Conference on Intelligent Systems Design and Applications. 866-870. http://dx.doi.org/10.1109/ISDA.2010.5687152

[17] W. Chang and M. Chung. "Automatic applying Bloom's taxonomy to classify and analysis the cognition level of English question items." 2009 Joint Conferences on Pervasive Computing (JCPC). 2009.

[18] L. Cutrone and C. Maiga. "Automarking: automatic assessment of open questions." Advanced Learning Technologies (ICALT), 2010 IEEE 10th International Conference on. IEEE, 2010. http://dx.doi.org/10.1109/icalt.2010.47

[19] P. Resnik, (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language, J. Artif. Intell. Res., vol11 p.95-130

[20] K. Jayakodi, M. Bandara and I. Perera, Automatic Classifier for exam question in Engineering:A process for Blooms Taxonomy, In Proc of 5th TALE conference, IEEE Press, Zhahai,China

[21] J. Vembunarayanan, "Tf-Idf and Cosine similarity", [available at] https://janav.wordpress.com/2013/10/27/tf-idf-and-cosine-similarity/

[22] Heiner, Cecily, and J. Zachary. "Improving Student Question Classification." Educational Data Mining 2009. 2009.

[23] Landauer, K. Thomas, P. W. Foltz, and D. Laham. "An introduction to latent semantic analysis." Discourse processes vol 25.2-3 (1998): p. 259-284.

[24] Moodle, [available at] http://www.moodle.org/

[25] Ateneu, "Introduction to Bloom's Taxonomy: LOT and HOT: Cognitive Engagement and Levels of Thinking", [available at] http://ateneu.xtec.cat/wikiform/wikiexport/cmd/lle/clpi/modul_3/apartat_1

AUTHORS

Kithsiri Jayakodi is a lecturer at the Faculty of Applied Science, Department of Computing and Information System, Wayamba University, Sri Lanka (e-mail: itkith@yahoo.com).

Madhushi Bandara is a lecturer at the Dept. of Computer Science and Engineering University of Moratuwa, Sri Lanka (e-mail: madhushi@cse.mrt.ac.lk).

Indika Perera is a senior lecturer at the Dept. of Computer Science and Engineering University of Moratuwa, Sri Lanka, (e-mail: indika@cse.mrt.ac.lk)

Dulani Meedeniya is a senior lecturer at the Dept. of Computer Science and Engineering University of Moratuwa, Sri Lanka, (e-mail: dulanim@cse.mrt.ac.lk)

This article is an extended and modified version of a paper presented at the 2015 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE2015), held 10-12 December 2015, United International College, Zhuhai, China. Submitted, 15 March 2016. Published as resubmitted by the authors in April 2016. Submitted 16 February 2016. Published as resubmitted by the authors 16 March 2016.