VISVESVARAYA TECHNOLOGICAL UNIVERSITY,

JNANA SANGAMA, BELGAUM - 590014

A SRS project report on

ASSESSMENT OF QUESTION QUALITY USING BLOOM'S TAXONOMY

Submitted in partial evaluation of the 8th Semester Project Progress Review - 1

1PI13CS092	Mohit Surana
1PI13CS147	Shiva Karnad Deviah
1PI13CS150	Shrey Agarwal

Under the guidance of

Internal Guide:

Prof. Nitin V Pujari

HoD, PESIT - CSE

External Guide:

Dr. Anantharaman Iyer

Co-founder, JNResearch

Jan – May 2017

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

PES INSTITUTE OF TECHNOLOGY,

(AN AUTONOMOUS INSTITUTE UNDER VTU, BELGAUM AND UGC, NEW DELHI)

100FT RING ROAD, BSK 3RD STAGE, BENGALURU - 560085

Assessment of Question Quality using Bloom's Taxonomy

Mohit Surana, Shiva K Deviah, Shrey Agarwal

Under the guidance of Prof. Nitin V Pujari and Dr. Anantharaman Iyer

ABSTRACT

Bloom's Taxonomy has been in existence for over half a century. The main idea behind this model was to promote higher forms of thinking in education, such as analyzing and evaluating concepts, processes, procedures, and principles, rather than simple rote learning. In evaluating the usefulness of Bloom's Taxonomy, only the cognitive domain is usually considered. This is a well researched topic and much work has been done on the subject. Bloom's Taxonomy has proved to be a good pedagogic tool in the field of education. However, in evaluating information, even the cognitive processes and knowledge level must be considered. This version of the taxonomy considering both the cognitive domain as well as the cognitive processes dimension as called the Bloom's Modified Taxonomy. The aim of this project will be to explore the various techniques in which the modified taxonomic can be applied, and to analyse the accuracy of each of these applications.

TABLE OF CONTENTS

1. Introduction

Bloom's Matrix, in simple terms, quantifies the level of knowledge and skill required to carry out some task. The original version of Bloom's Taxonomy (Bloom, 1956) documents only the cognitive domain, which consists of the following:

1. Remember
    - Recall memorized information. Different keywords include define, describe, identify, recall, etc.
2. Understand
    - Understanding that enables one to explain in one's own words some problem or how to perform a task. Different keywords include estimate, explain with example, summarise, and interpret.
3. Apply
    - Use a concept to good effect in a new situation; application of knowledge. Keywords include apply, demonstrate, modify, solve, and use.
4. Analyze
    - Decompose a complicated concept into smaller components so that the organizational structure may be understood. Keywords include breakdown, distinguish, identify, illustrate, and infer.
5. Evaluate
    - Judgements or informed opinions about the value of a concept. Keywords include compare, contrast, justify, evaluate, and explain.
6. Create
    - Create a new concept/structure/pattern from existing elements. Keywords include devise, design, modify, plan, reorganise, rewrite, compose, combine, compile, and categorise.

A modified version of Bloom's Taxonomy (Anderson and Krathwohl, 2001) considers this knowledge domain in detail:

1. Factual
    - The basic level of knowledge that anyone must be acquainted with to solve problems
2. Conceptual
    - The interrelationships among the basic elements; the understanding behind how things work
3. Procedural
    - The knowledge to do something
4. Metacognitive
    - Knowledge of cognition in general as well as awareness and knowledge of one's own cognition

Together, these 2 domains form the knowledge matrix of Bloom's Modified Taxonomy. Currently, there exists no tool in the open source community that assesses question quality according to Bloom's Modified Matrix, considering both the cognitive level and cognitive processes dimension together. Furthermore, existing work in this domain considers levels to have a flat relationship; rather than hierarchical, which in actuality, they are. For example, in order to explain the rationale behind an a technique or concept, you need knowledge and understanding of the problem domain too. Our system will aim to capture the hierarchical structure of the levels during the process of question classification.

A successful implementation of this system will find many practical uses in the field of education. These include:

1. Assess the quality of lecture delivery; analysing students' doubts after a lecture to determine their understanding of the topic post lecture
2. Automated question paper setting; selecting questions based on their difficulty level.
3. Weighted GPA system; apply weightage to subject grade by analysing question papers set for that subject.

1. Problem Definition

The aim of this project is to assess the quality of questions by classifying them according to Bloom's Modified Taxonomy. For every question, a heatmap is to be generated indicating where the question lies on the matrix. This project also involves evaluating some of the applications of Bloom's Taxonomy.

The project can be divided into three main phases:

1. Dataset collection

Since this is a first attempt at assessing questions based on Bloom's Modified Taxonomy, we will be collecting data by using the Stack Exchange API to retrieve questions based on keywords extracted from the mini-world reference, and label the questions accordingly.

2. Training

Train 2 classifiers for both the dimensions separately and get the softmax probability for each question. Combine the results from the above 2 classifiers using regression techniques (by assigning weights).

3. Use case evaluation

I. Evaluating lecture delivery based on the doubts raised by students.

II. Automating the setting of question papers by cherry-picking questions based on their difficulty level as identified by our system.

III. Determining weighted GPA. Deriving a relationship between the difficulty level of the question paper answered by a student, their marks scored, and their obtained GPA to get a weighted GPA that truly reflects the effort the student has put in order to get their marks.

2. Generic Proposed Solution

Our approach to solve the problem at hand will involve the following steps:

1. Extraction of keywords from the mini-world reference (textbook / online articles), followed by validation of the keyword by checking the context attached to it (this has been done using DBPedia / Wikipedia).
2. Question Retrieval:
    a. Extraction of questions based on the keywords obtained in Step 1, through the use of

the StackExchange API. In order to speed up this process, either mapreduce or threads can be considered. Through this process, collect a large number of questions (around 1-2 million).

b. Filter out noise from the dataset (such as questions that do not explicitly apply to the current problem domain but were picked up due to the nature of the keyword used to query that question).

3. Labelling questions on difficulty as per different levels in Bloom's Matrix. Two approaches are possible:

   a. Either follow a rule based approach by applying weights to classes and assigning questions to one or more classes based on the keywords

   OR

   b. Identify / hand-label a small set of different kinds of questions using human recall. These questions become the seed for subsequent labelling through clustering of questions.

4. Train classifiers to accurately output a probability distribution of the various levels a question can belong to, based on the seeded data. Again, two approaches are possible:

   a. Assume an independence between the cognitive domain and the cognitive processes dimensions. This enables us to train two classifiers in parallel, and a third classifier that learns how to combine the output of the two classifiers to obtain the final probability distribution possibly using linear / logistic regression. (This is the approach we have chosen to follow for now.)

   OR

   b. Treat the two dimensions together during classification, and use only one classifier to learn. This is a little harder compared to the first proposed methodology as it has to work with more classes, learn complex hierarchies and requires more careful labelling.

5. Display the predicted difficulty for unknown questions not seen in the dataset. This should be shown in the form of a heatmap.

6. The system described in Steps 1 through 5 should be used to evaluate at least one of the three use cases described above.

3. When the Project is Considered Done

The project will be considered done once we are able to classify questions with an accuracy of 65% or higher. This accuracy is determined, not by our system, but by human recall as an effort at validating our system.

2. Literature Survey

As part of literature survey, we first looked at Bloom's Taxonomy, what the motivation for it was, various improvements made to the original model and its applications. Then we looked at past work done to replace manual intervention in the task of classification of these questions and the reasons to use the same.

Bloom's Taxonomy came into existence in 1956 and slowly it began to be used as a pedagogical tool to assess the various levels of knowledge or skill that was required to answer questions. This helped evaluators to design papers to suit the level of their target audience and also helped them to assess the weak areas in need of attention. This was used in conjunction with specialized exercises that were shown to help improve the different cognitive levels of knowledge of a given person with the aim to eventually lead to their better performance.

In this Taxonomy, the Knowledge category included both noun and verb aspects. This brought unidimensionality to the framework at the cost of a Knowledge category that was dual in nature. This anomaly was eliminated in the Revised Taxonomy. Introduced in 2001 by Anderson, Krathwohl, et al, the Revised Taxonomy allowed the two aspects, the noun and verb, to form separate dimensions — the noun providing the basis for the Knowledge dimension and the verb forming the basis for the Cognitive Process dimension.

Although the possible benefits of using this are plentiful, its application is not widespread due to the requirement of intensive and skilled manual labour in order to classify the content. Many of the studies that were conducted required the support of domain experts to label the data and even amongst them there was often dissensus. This has prevented scalable adoption of the Bloom's Taxonomy and ultimately the loss of the opportunity to reap its benefits.

In the recent past,

- Van Hoeij et al (2004) applied Bloom's taxonomy on essay-based questions but they recorded very low accuracy.

- Chang and Chung (2009) achieved high accuracy but they used an extremely small dataset.
- Yusof and Hui (2010) implemented an Artificial Neural Network (ANN) approach using multiple feature methods but were unable to get good accuracy.
- Haris and Omar (2012) used a rule-based classifier which was time consuming, expensive to make and maintain, limited to computer subjects, and performed badly in other domains.

Apart from these papers, there were a few that showed promising results:

- Exam Questions Classification Based on Bloom's Taxonomy Cognitive Level using Classifiers Combination (Dhuha Abdulhadi Abdul Jabbar, Nazlia Omar)
    - Explored different concepts such as SVM, NBC, and K-nearest neighbours with good results: 80-90% accuracy.
- Automatic Classification of Questions into Bloom's Cognitive Levels using Support Vector Machines (Anwar Ali Yahya, Addin Osman)
    - Used SVM with good results - 80-90% accuracy.
- Automatic Classification of Answers to Discussion Forums According to the Cognitive Domain of Bloom's Taxonomy using Text Mining and a Bayesian Classifier (Jhonny Pincay, Xavier Ochoa)
    - Used NBC, but achieved a moderate degree of precision.
- Classifications of Exam Questions using Linguistically-Motivated Features: A Case Study Based on Bloom's Taxonomy (Addin Osman, Anwar Ali Yahya)
    - Used SVM, NBC, Logistic Regression and Decision Trees and used a combination voting strategy to pick the class from the predictions made by the classifiers.

All work that we found involved assessment on the cognitive domain and was, thus, restricted to one-dimensional analysis only. This prevents us from having a more fine grained assessment. We wish to tackle this by evaluating on a two dimensional grid formed by cognitive and the knowledge domain. Most of the past work had very low values of recall and we wish to improve upon that by making use of better algorithms and representations.

3. System Requirements Specification

Below is a high level flow chart indicating the procedural flow of our system.

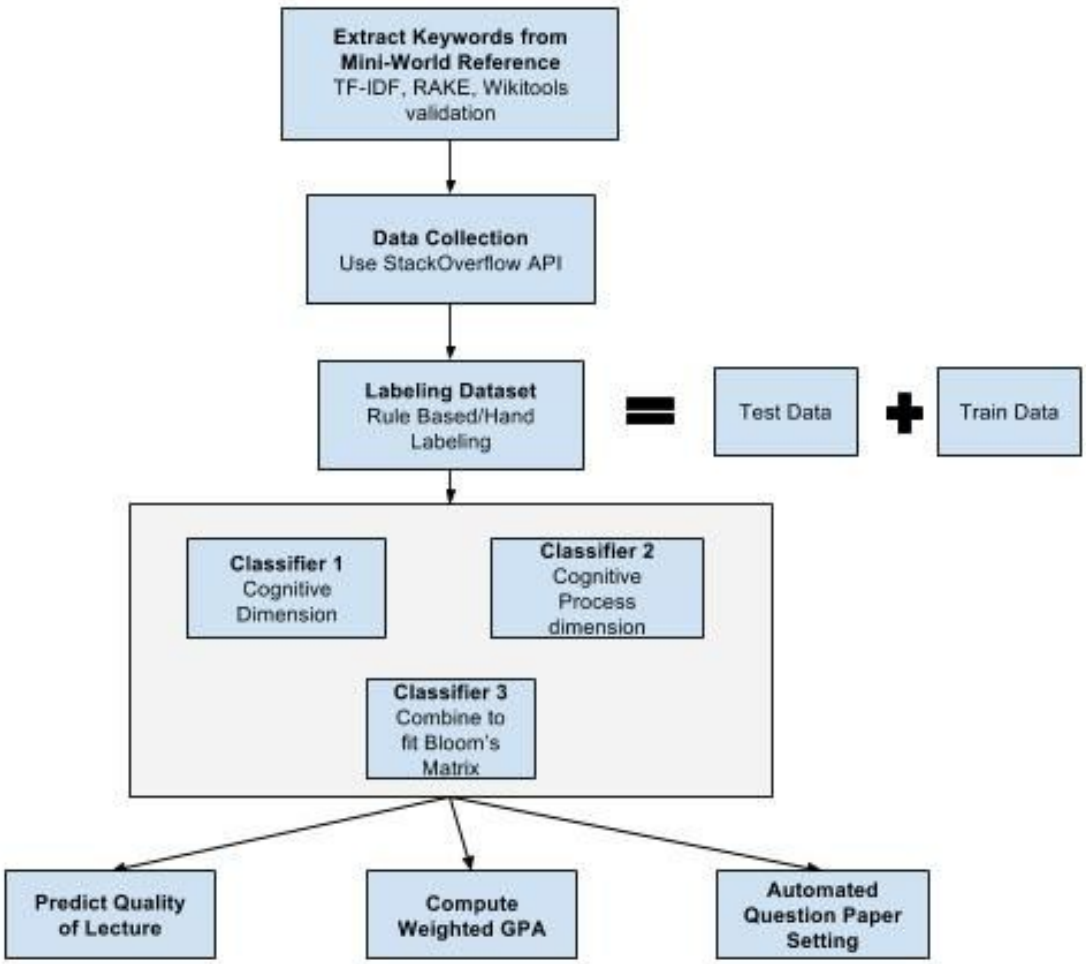

Fig 3.1: High level overview of the system

1. Development Environment Requirements

1. Software Requirements

Listed below is a list of all the software tools required to complete this project:

1. *python2.7* and *python3.6* for performing dataset collection)
2. *Orange* tool (for data mining on the collected data and visualization)
3. *Mapreduce frameworks* (to enable dataset collection to be scalable):
    a. *Hadoop and HDFS*
    b. *pyspark*
4. Python ML frameworks:
    a. *scikit-learn*
    b. *keras*

2. Hardware Requirements

In order to exploit the full power of the mapreduce framework for dataset collection, a server cluster is required. This an optional requirement. In the absence of a cluster, i.e., in standalone mode, a threaded approach to dataset collection will be faster due to less overhead.

2. Project Requirements

1. Functional Requirements

F1	Keyword Extraction	Given a mini-world reference, the system should adequately extract the most meaningful keywords in that context.
F2	StackOverflow Question Extraction	The system should leverage the APIs exposed by StackOverflow and partner websites to fetch a number of questions associated with the provided keywords.
F3	Question Paper Extraction	The system should also be able extract questions from a given question paper in a particular format.
F4	Assessing Question Quality	The system should be able to classify questions according to Bloom's Modified Taxonomy, displaying a heat map.
F5	Lecture Quality Analysis	The system should analyze doubts asked by students after a lecture and provide some statistics about the lecture.
F6	Question Paper Analysis	The system should analyze a given question paper and provide a heatmap indicating the overall level of the paper.
F7	Weighted GPA	The system should support a fair scoring system that takes into account the level of difficulty of a question paper while calculating the GPA of a student.

2. Project UI

U1	Interactive User Experience	A Graphical User interface should be provided where a user should be able to enter a question (or a list of questions).
U2	Heat Map	When the submit button is clicked, a new window should appear which shows a heatmap on the Bloom's Matrix for the question.

3. Non-Functional Requirements

NF1	Aesthetic frontend (usability)	The front end for the application must not suffocate the users with too much information, and should be intuitive and user-friendly.
NF2	Documentation	Should be provided with a simple yet comprehensive usage manual. Code should have sufficient comments so that future development and maintenance is easier.
NF3	High backend performance	Should implement the most efficient and appropriate algorithms for each and every sub task.
NF4	Maintainability	All features should be designed in a modular form to allow easy maintenance.
NF5	Response time	Response time should not be too long on an average.
NF6	Reusability	The classifier when developed for a particular subject, should be migratable to suit the requirements of other similar subjects without much hassle.

3. Constraints and Dependencies

The StackExchange API has a daily limit of 300 requests per day and a rate limit of 30 requests/second.

Dependencies include the following:

1. Upstream dependencies:
    - StackExchange and partner websites
    - All of the software mentioned in Section 3.1.1.
2. Downstream dependencies:
    - End users such as teachers, students and educational institutions

4. Assumptions

- For the purpose of training, we will assume independence between the two dimensions on Bloom's Matrix. This makes for easier classification and identification of hierarchical structure amongst classes which can be captured easier if they are treated independently.
- We also assume that the software mentioned in Section 3.1.1 being used is bug free.

5. Use Case Diagram

Below is the use case diagram for our system.

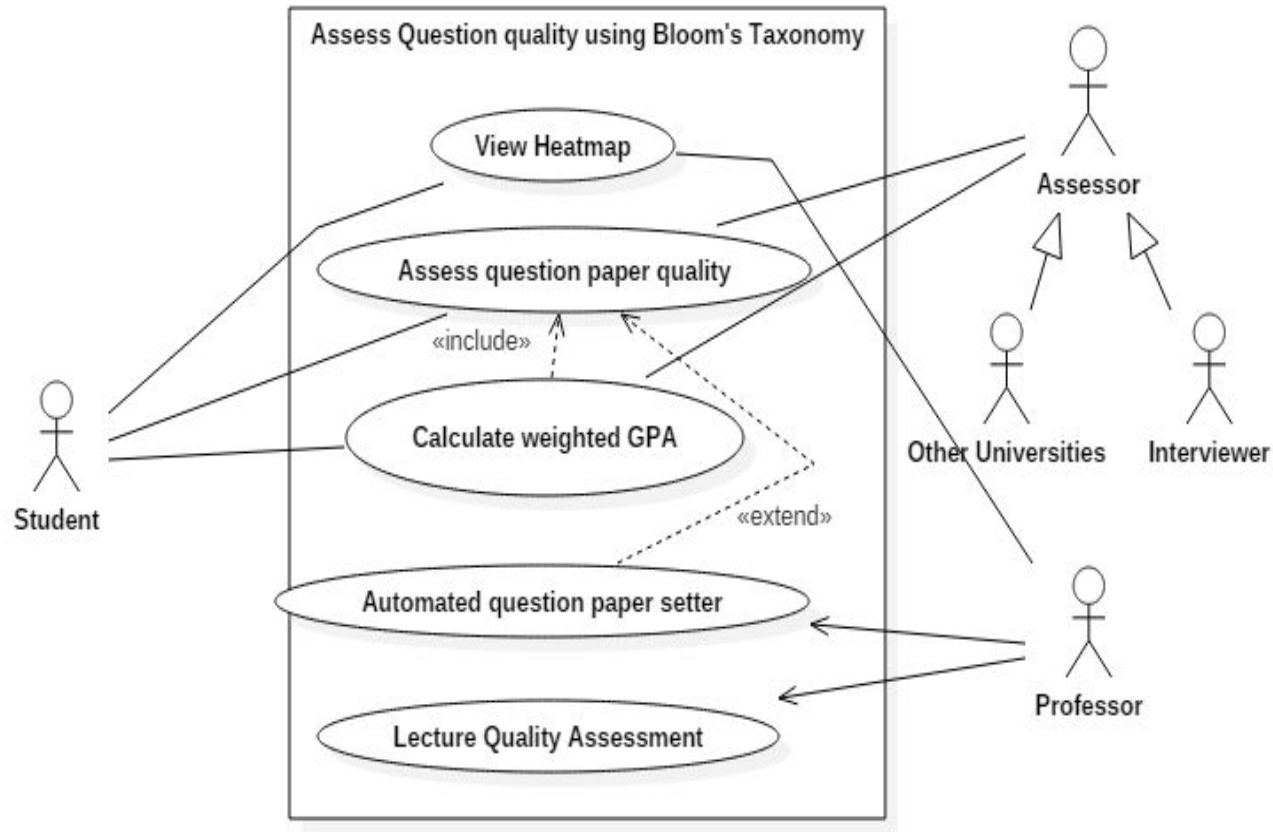

Fig 3.2: Use Case Diagram for the question quality assessment system

6. Requirement Traceability Matrix (RTM)

Req ID	Description	Assignee	Test Plan	Test IDs
F1	Keyword Extraction	Shiva, Shrey	Will verify the quality of keywords extracted using the glossary of the textbook.	T1
F2	StackOverflow Question Extraction	Shiva	A list of keyword - question ID pairs will be maintained that can be used to check if there are any issues.	T2
F3	Question Paper Extraction	Shiva	The question papers provided by the institute will be parsed by us and with the help of exception handling, we will detect parse errors.	T3
F4	Assessing Question Quality	Mohit, Shiva, Shrey	A labelled dataset (BCLs) that has been created by the authors of one of the papers cited in the Literature Survey will be used for the initial training. Further training will be done by using data labelled by us after discussion with subject teachers.	T4
F5	Lecture Quality Analysis	Shrey	Students with and without prior knowledge of a topic will attend a class and their views will be used to benchmark the answers provided by the system.	T5
F6	Question Paper Analysis	Mohit	The question paper setter will be asked to provide some indicative labels and other teachers for the same subject shall discuss and once consensus has been reached, we shall use that as the standard.	T6
F7	Weighted GPA	Shiva	The GPAs for the last academic year and the papers for a few select subjects shall be used to showcase the difference that the new scoring system brings.	T7

U1	Interactive User Experience	Shiva, Mohit	Shall be tested by asking various users with differing levels of comfort in the usage of computers and their feedback will be used to improve or modify the provided user interface.	T8
U2	Heat Map	Mohit, Shrey	A comprehensive set of self generated cases shall be used to see if the right regions are being highlighted.	T9

4. Schedule

Shown below is a Gantt Chart documenting the timeline of our project.