1. Introduction

Cognitive domains in Bloom’s Taxonomy (BT) are widely used in educational field to analyze student understanding and knowledge (Anderson, Krathwohl & Bloom, 2001). Cognitive level consists of six (6) levels, which start from a lower level of learning to a higher level. The lower level is known as knowledge level, which purpose to recall data from previous lesson. Second, comprehension level requires students to have an ability to explain the knowledge learns. Third, application level, student should able to apply the knowledge learned into an action. Fourth, the analysis level address student to investigate information they studied. Fifth, synthesis level, student needs to link all the information and integrate it into something new. The final level is the evaluation level where the student may achieve when they able to stand for any opinion. To produce a good written exam questions, it must consists all of six cognitive levels. The lectures might have to consume time in classifying the question since they also need to follow the guidelines of constructing exam questions which are JSU and common problem like error typing when construct the questions might happen. The time allocates in classifying the exam question might extend and causes replication of exam questions. Thus, it could delay the development of exam questions since the process is done manually. Even though, the lectures need to follow the BT level, however, there is a high probability of wrongly classify the cognitive level of exam questions. Furthermore, the exam questions constructs are having a lot of bloom’s keywords which somehow two (2) keywords can belong to more than one level of cognitive or one (1) keywords belonging to two (2) cognitive levels. The problems with keywords belonging cause difficulty for educators to identify the correct level for each exam questions. Each exam questions generate must follow the bloom’s taxonomy guidelines. Thus, as constructors, they need to classify the exam question generates, according to

Abstract

This project is about analyzing the effects of classifying the written exam question into cognitive level of Bloom’s taxonomy. Correctly analyze and classify the written exam questions into correct cognitive level can generate a good set of exam questions. As known by many educators, classifying exam question into its cognitive level is a tedious task and required full attention by educators. Moreover, there are situations where one keyword of cognitive level belongs to more than one level which could be an issue of difficult to determine the correct cognitive level of questions. To solve the problem of classifying exam question faces by educators, the techniques information retrieval of text mining were implements in this project. Before that, question bank are required to perform text preprocessing to generate the clean data. The activities done under text preprocessed are such as data transformation, tokenization of question and stop word removal. The effects of classifying the clustered data being analyzed to study the possible hidden pattern of classifying based on Bloom’s Taxonomy.

Keywords: Classification, Clustering, Text-Mining
its level of bloom’s taxonomy by recognizing the variables that suitable. In order to gives solution of that problem, classification techniques are one of approaches that widely used in a large data set to find hidden information. In addition, Natural Language Processing (NLP) was implemented to analyze the cognitive level for each question as it is helpful to analyze data in text forms.

2. Literature Review

According to Krathwohl (2002)\(^2\), Benjamin Bloom was the one who introduced the BT in the year 1956. Bloom’s Taxonomy consists of three (3) domains which are Cognitive, Affective, and Psychomotor. The first domain which is cognitive domain is focused on intellectual skills includes the ability to call up previous learning. While affective domain refers to the feeling which includes the way people handle their emotions. Last domains, the psychomotor relates to the physical movements. Many researches have been done researches on Bloom’s Taxonomy especially in educations. The Taxonomy of Educational Objectives (TEO) or known as Bloom’s Taxonomy is a framework aimed to classify learning objective, according to the cognitive level gained by students. Basically, the focal points of cognitive domain are on psychological skill which related to the mental thinking. Bloom’s Taxonomy is an important element when developing exam questions. According to Asshaari, Othman, Bahaludin, Ismail and Nopiah (2012)\(^3\), the crucial phases in constructing examination questions are to categorize the question itself refers to its Bloom’s Cognitive. Text mining is a process of discovering knowledge from textual data which are also known as text data mining (Fayyad, Piatetsky-Shapiro & Smyth, 1996;\(^4\) Hears, 1997). As known by people data mining were an extraction of useful information from structured databases of facts, while text mining are looking for patterns from natural language text (Weiss, Apte, Damerau, Johnson, Oles, Goetz & Hamp, 1999)\(^5\). Generally, the text mining task starts with selecting the document collection, which doesn’t differ with the data mining tasks. (Feldman, Fresko, Kinar, Lindell, Liphstat, Rajman, Schler & Zamir, 2006)\(^6\).

Text mining is actually part of data mining, Artificial Intelligence (AI) and Machine Learning, Computational Linguistics, Statistics, Databases and Library and Information Sciences. Since the exploration of mining in forms of text, many techniques have been discovered to fulfil today’s need in the context of text mining (Nagarkar & Kumbhar, 2015)\(^7\). As mentioned by Miner, Elder, Fast, Hill, Nisbet and Delen (2012)\(^8\) in their books, text mining consists of seven (7) types which are document classification, document clustering, Information Extraction (IE), Information Retrieval (IR), web mining, NLP and concept extraction. Basically, information retrieval is about finding information. In details, it is the process of retrieving information from a large number of text-based documents related to information search, access, need and retrieval. Normally, information retrieval cater problem of locating relevant documents in a document collection based on a user’s query.

Text categorization (TC) is the process of assigning a collection of documents into predefined categories (Genkin, Lewis & Madigan, 2007)\(^9\). Many applications related to automatic text categorization such as document retrieval, detecting topics of the contents, automatically extract information and much more (Sebastiani, 2002)\(^10\). Up until 1990s, the techniques of “knowledge engineering” was nominated text categorization that needed to bring out categorization rules by expertise and apply the rules automatically to new documents by code them into a system (Witten, 2005)\(^11\).

3. Methodology

Data gathering is an initial step of development where all the data and information related to this process are collected. Any boundaries or limitations are identified to prevent from work done over scope. Based on the researches of environments, the current process is being analyzed. A second phase in this methodology phase is data analysis. Once the data is gathered, techniques or solution was identified and the preprocessed data were held to get the valuable information which needed to use on classifying process. Final phase, which is knowledge extraction, is where the rules were generated from the analyzed data. The final step is design and implementation.

3.1 Data Preprocessing

Data transformation is being conducted. Text-form
data is transformed into numerical representation. Each question occupied with difficulties level being transformed into a simpler keyword to increase the performance of database in retrieving the data.

Data cleaning then take place where the missing value being reduced. Some of the questions could have many missing attributes such as the difficulties level, the keyword and the knowledge level in bloom taxonomy. The missing value problem can be minimized by delete the tuple or fill-up the cell with default value. For some attributes which consist of too many missing data could lead to inaccurate analysis later on.

Data smoothing conducted to normalize the questions set according their attributes. Binning and normalization create the initial cluster of the sets. Smoothing by means and normalized using z-score normalization is performed into the data.

$$z = \frac{x - \mu}{\sigma}$$  \hspace{1cm} (1)

Equation (1) shows the z-score normalization where:
- $\mu$ is the mean of the population;
- $\sigma$ is the standard deviation of the population.

Stratified sampling then being conducted to select group by group of data to undergo correlation analysis and then perform clustering. Datasets also divided into 70 per cent of training sample and 30 per cent for testing respectively.

Correlation analysis was conducted into the datasets to discover the pattern and relationship of one attributes to another and determine cause and effect involved in the data presence. Two formulas involved in this stage which are chi-square and regression analysis. Equation (2) shows the selected chi square:

$$\sum \frac{x_i - y_i}{y_i}$$  \hspace{1cm} (2)

Where,
- $X$ is observed questions keyword;
- $Y$ is expected questions keyword.

Regression analysis performed to reduce outliers which have minimum or no impact towards the attribute or parameters set. Some of the regression analysis also turned to indicate the correlation among datasets. The regression determined the dependence among variables. Some of the attributes in a set of questions may have negative correlation but still used to calculate causal effect to others.

### 3.2 Clustering Processed Data

Determine hard-NP and soft-NP to sort the priority of which attribute being set as main attribute to cluster. Take the sample of training data first to perform the analysis. Weight-value for each attributes and parameters is being analyzed. Weight value given after thorough analysis in each parameter is being adjusted to its optimal points. Next is finding the centroid for the data to perform K-means by using Euclidean Distance.

Cluster the data based on the result of Euclidean Distance and repeat the calculation until the graph of cluster reach its optimal point. The K-Means analysis is repeated by using different value of centroid. Then the process of measuring the distance between data using Euclidean Distance being calculated using the same datasets of questions. Once the model of the result is obtained, rules for cluster the data are generated. The generated rules from the clustered data being tested by using 30 per cent of sample data which being hold for testing purpose.

### 3.3 Classify using Text Mining

The data are divided into a training set and testing set. Training set consists most of the data and a small portion of the data is used as testing set. The testing set is used to build up a model while test set is to validate the performance of model built. Before classifying the training set, the strong and valid attributes are chosen based on identified attributes. Those identified attributes were keywords of question and marks provided for each questions.

At the same time, list verbs of cognitive domain from bloom’s taxonomy of were checked with the previous exam questions to choose only the suitable verbs. Then model built based on training set were compared with the testing set. The deliverables of this activity are data were classified according to its class label.

$$p(a | b) = \frac{p(b | a)p(a)}{p(b)}$$  \hspace{1cm} (3)

Based on the preprocess question, tokenize is performed to analyze the sentences. Every token of words were matched with verbs from of cognitive level. Whenever the question matched the verbs, it were classify into its cognitive level based on Naïve Bayes probability of usage stored in the database. Rules are extracted after classified the data into its class.
Studying the Effects of Performing Text Mining to Improve Classification of Clustered Questions based on Bloom Taxonomy

Where,
\[ P(a|b) \] is posterior probability for the keyword
\[ P(b|a) \] is likelihood
\[ P(a) \] is taxonomy level prior probability
\[ P(b) \] is taxonomy level predictor prior probability

Classification rules are ‘IF-THEN’ rules. ‘IF’ indicates condition over the data, whereas ‘THEN’ part refers to a class label. Based on the classification of data into its probability class label, rules are generated. Rules identified were implemented into the clustered data.

4. Analysis and Findings

Based on the results from correlation analysis conducted earlier, the K-Means analysis performed with 14,000 questions available in question bank. Fig. 1 shows the result yielded with the application of Euclidean Distance measure for every keyword in each question involved. For each question set, calculation performed with at least three different values of K, being one the values equal to the number of existing classes in the data. Each data set, nondeterministic, ran 100 times of iteration for each testing question.

Table 1. Classified data without Text Mining

| Level | Means Squared Error | Correctly classified (%) | Misclassified/undefined (%) |
|-------|--------------------|--------------------------|----------------------------|
| C1    | 3.1                | 42.13                    | 57.87                      |
| C2    | 5.6                | 44.7                     | 14.47                      |
| C3    | 1.25               | 51.78                    | 48.22                      |
| C4    | 1.91               | 52.9                     | 47.1                       |
| C5    | 4.21               | 56.8                     | 43.1                       |
| C6    | 2.19               | 64.15                    | 35.85                      |

Table 1 indicated the classified questions to their class label. Performing classification without tokenize the and clustered the keywords only yielded at most 56.8 per cent accuracy at class C5. Due to class level C5 consist of high clarity of keyword compared to other classes.

Table 2. Classified data using Text Mining

| Level | Means Squared Error | Correctly classified (%) | Misclassified/undefined (%) |
|-------|--------------------|--------------------------|----------------------------|
| C1    | 0.44               | 71.52                    | 28.48                      |
| C2    | 0.58               | 70.64                    | 29.36                      |
| C3    | 0.92               | 69.18                    | 30.82                      |
| C4    | 0.26               | 80.63                    | 19.37                      |
| C5    | 0.19               | 82.14                    | 17.86                      |
| C6    | 0.33               | 76.44                    | 23.56                      |

After performing text mining in tokenize the keywords and calculate the probability of belonging by using Naïve Bayes, the result as in Table 2 is yielded. The accuracy of classification model eventually increased due to the involvement of the text mining. The increment provide clearer gap margin between correctly classified and misclassified questions. With the assistant of the Naïve bayes probability function, the belonging percentage clarified.

Figure 1. Initial cluster by using K-Means.

Each clustered data being used an initial class to determine the probability of the questions belong to. Table 1. shows the average classified data before text mining is performed while Table 2 demonstrate the after-used findings.

Figure 2. Classified data using text mining.
To evaluate the accuracy, efficiency and scalability of classified data into its high probability class label, an extensive performance study have performed. In general, if the classification performed solely based on a single keyword in a sentence, some effective rules may be missed. Thus, to some extend, the whole sentence of a question may lead to higher class label considering hidden meaning of the supporting words after or before the specified keyword. The results are shown in Figure 2, where the classified questions are mostly above 50 per cent accurate. However, the results may be varied depending on the data volume and the inference engine which stored the specified keywords.

5. Conclusion

This research focusing on studying the effect of applying text mining into clustered data. Based on our scope classifying structured of text in aligning with class label in Bloom Taxonomy, we found that it is a success experiment in hybrid a cluster and classification approach. Some of the keywords may contributes to the several class labels that may produce vague measurement. For further analysis, this research could be enhanced with adding several uncertain parameters and perform Fuzzy Inference technique.

6. Acknowledgement

I would like to take this opportunity to express my profound gratitude to Universiti Teknologi MARA and Ministry of Higher Education of Malaysia (KPM) for funding this research. This research is under the Research Acculturation Grant Scheme (RAGS) with file number 600-RMI-RAGS 5/3 (1/2014).

7. References

1. Anderson WL, Krathwohl DR, Bloom BS. A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Allyn & Bacon, 2001.
2. Krathwohl DR. A revision of Bloom's taxonomy: An overview. Theory into practice. 2002; 41.4:212–8.
3. Asshaari, Izamarlina. Appraisal on Bloom's Separation in Final Examination Question of Engineering Mathematics Courses using Rasch Measurement Model. Procedia-Social and Behavioral Sciences. 2012; 60:172–8.
4. Fayyad U, Piattetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI magazine. 1996; 17.3:37.
5. Weiss SM, Apte C, Damerau FJ, Johnson DE, Oles FJ, Goetz T, Hampp T. Maximizing text-mining performance. IEEE Intelligent Systems. 1999; 4:63–9.
6. Cherf H, Napoli A, Toussaint Y. Towards a text mining methodology using association rule extraction. Soft Computing. 2006; 10(5):431–41.
7. Nagarkar SP, Kumbhar R. Text mining: An analysis of research published under the subject category 'Information Science Library Science' in Web of Science Database during 1999-2013. Library Review. 2015; 64(30):248–62.
8. Miner G, Delen D, Elder J, Fast A, Hill T, Nisbet RA. The Seven Practice Areas of Text Mining. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. 2012; 29–41.
9. Genkin A, Lewis DD, Madigan D. Large-scale Bayesian logistic regression for text categorization. Technometrics. 2007; 49(3):291–304.
10. Sebastiani F. Machine learning in automated text categorization. ACM computing surveys (CSUR). 2002; 34(1):1–47.
11. Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.