International Journal of New Technology and Research (IJNTR)
ISSN: 2454-4116, Volume-5, Issue-10, October 2019 Pages 28-30

# An E-Exam Management System under E-Network Management Course

**Shaimaa H. Shaker, Noor M. Jaafar**

***Abstract*—in the methods of teaching the E-learning become the one of great invention, there are many assessment methods in E-learning, one of these assessment is the essay, when teacher want to correct manually it may take long time, so the automatic scoring of essay in recent years has been increasing rapidly. Using usually the method of cosine similarity, it depends on feature extraction in the texts of answers, no matter about the order of the words in the statement. In this study show how the text similarity can have evaluated the student's answers, so its need to be more accurate.**

***Index Terms*— E-exam, cosine similarity, TF, IDF, Text similarity, Tokenization, stemming.**

## I. INTRODUCTION

E-learning is widely used as educational innovation specifically in learning process and teaching these days. The examination could be any form. It can divide to more than one type like questions with multiple choice answers, calculation tasks or essays and true/false. A multiple choice type of examination is sometime best way for tests performed during the course. In the final exam the essays are more suitable. For essays type what need it is an approached to calculate the similarity between typical teacher answer and student answer. One of popular method of automatic scoring essay is the cosine similarity, most of time the cosine similarity presented best result for the string similarity [1] [2].

- **Text Similarity**

In information retrieval and automatic essay scoring or short answer grading the measuring of the similarity between (words, sentences, paragraphs and documents) is very important component in these diverse tasks. Words can be similar in two basic shapes (**Lexical based similarity (LBS) and semantically based similarity)** [3]. In LBS the degree of similar of two available string that their characters have a similar character sequence and words are similar, this was in general but also there are phrase level can be compute the similarity for it too by using something called phrase chain. *The* output of lexically way are 1 if the strings are identical if there is no common the output will be 0, semantically if they have the same thing, and each type of these shapes have different approaches inside. There are four main approaches for both as shown in figure (1) [4].

**Shaimaa H. Shaker**, Department of Computer Science, University of Technology, Baghdad, Iraq.

**Noor M. Jaafar,** Department of Computer Science, Iraqi Commission for Computers and Informatics/ Informatics Institute for postgraduate studies, Baghdad, Iraq.

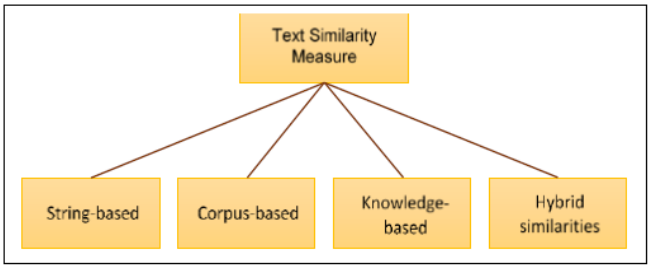


**Figure 1: Text Similarity Type**

➢ **Preprocessing of Text**

Data preprocessing used for extracting important and non-trivial and knowledge from text data. Preprocessing of text is deciding which documents or words in a collection should be retrieved to matching a user's need for information [5]. The standardization preprocessing steps are:

1. **Tokenization**

Its important step in information retrieval, the task of this a step is splits large strings made of texts into smaller pieces, which called tokens which is meaningful elements. The big chunks of text which is smaller than large strings can be process by tokenization and convert it into sentences, then at end these sentences can be tokenized into words. It's also in this step remove non letter character like (white space, special control characters) [6].

2. **Stop Word Removal**

is an important step in preprocessing techniques that should be used in Natural Language processing applications like Information Retrieval System, Text Analytics, and Question-Answering system, this step can improve the performance of system by reduce size of text by 20-30% of overall word counts in a specific text documents. In English language these words have been defined as conjunctions, prepositions, adverbs. The process of this step depend on compared a redefined list of stop words with the target text to specify which removal is required [5] [7].

3. **Stemming**

There are many algorithms of stem that have developed to optimize data, where stemming is the process of returning the different representation of word forms to the origin word [8]. This process of terms consolidated may reduce the size of index file in information retrieval or in question/answer systems. The aim of this procedure is to remove the affix (suffix, prefix). The most important purpose for stemming process is the words that sharing the exact root or stem will reduce the total number of words that need to process and save time and memory space [9].

➢ TF: Term Frequency

In this model its work to measures how frequently a term occurs in a document. Where each document is absolutely having some different in length of terms, sometime the possibility of a term would appear more number of times in document with long length than the shorter length.

$$\mathrm{Tf}(t.d) = \frac{term\ t\ count\ in\ d}{count\ of\ term\ T_d} \quad \ldots.1$$

Where TF is Term Frequency, t refers to term, and d is document. TF (t, d) = (Number of times term t appears in a document) / (Total number of terms in the document) [10].

➢ **IDF: Inverse Document Frequency**

Which measures how important a term is. While computing TF, all terms are considered equally important. However, it is known that certain terms, the relationship is reversible the more frequent the word the less the importance of the word. Thus need to weight down the frequent terms while raise the rare ones, by computing the following equation:

$$\mathrm{IDF} = \log\frac{N}{Ni} \quad \ldots. 2$$

Where: N denotes the total documents number in the group of documents. Ni denotes in how many documents the word (i) occurred in the group of documents. IDF (t) = log_e (Total number of documents / Number of documents with term t in it) . IDF may be used within TF model as a coefficient such that TF of each term is multiplied by IDF: [11].

$$\text{TF-IDF (Di,t)= TF(Di,t)} * \text{ IDF(t)} \quad \ldots.3$$

➢ **Cosine Similarity**

The similarity between two texts is derived from cosine value between two texts' term vectors. Cosine similarity work in better way in lexical where it's still can't handle the semantic meaning of the text perfectly. Cosine similarity this process can be defined as measuring the similarity between the two specific vectors related to n dimensions through finding the angle between them. Assume A and B (vectors of attributes).

$$\cos(\vec{a},\vec{b}) = \frac{\vec{a}\cdot\vec{b}}{|\vec{a}|\cdot|\vec{b}|} = \frac{\sum_i(\vec{a_i}\cdot\vec{b_i})}{\sqrt{\sum_i(\vec{a_i}^2)\sum_i(\vec{b_i}^2)}} \quad \ldots. 4$$

The resulting similarity equation which must ranges from zero that meaning exactly opposite, to 1 meaning exactly the same, and the values in-between referring intermediate similarity or dissimilarity [12]

## II. RELATED WORK

**Mochammad Ali Fauz, Budi Darma Darma Setiawan [2017]** proposed two ways for evaluated the student's answers based on n-gram and cosine similarity, where each method depend on specific steps, n-gram used as feature extraction and modified with cosine similarity to split by words rather than the letters [1].

- **Hebah Rababah, Ahmad T. Al-Taani [2017]:** Design system for Arabic essay scoring system to evaluating the answers, using cosine similarity measure based on string-based similarity [13].
- **Phil M, Arunesh. K [2016]:** design system to measure between two documents using cosine similarity algorithm which can be used in essay scoring or information retrieval [14].

## III. PROPOSED METHOD

In this proposed method will illustrated the way to evaluated the students answers in essay assessment type Most of E-learning website systems do not have the examination with short answers type which are frequently limited to multiple choice questions, so in this part will discuss the proposed way of how the system will exams the students and evaluated them based on text similarity. Where all words are subject to specific step of prepossessing, which the words of both students and teacher typical answer goes to remove stop word and stemming after tokenization, then these word must be weighted by using TF-IDF model which is represent in vector model, and to measure the similarity and the relevant between student vector and teacher vector the cosine similarity took the role and measure the result.

**The Proposed E-Exam System for E-Network Course**

**Step1:** the words in the document of student and typical answer which are define as texts are separated into number of tokens by using the space as a character define as a splitter to recognize each single word from another. Tokens might be defining numbers, symbols and words. These words are consider as input tokens to later stages.

**Step2:** Take list of terms from the step1 and set of selected stop words prepared as samples to be removed in this system which that do not have positive effect in weighting of evaluates and meaning of the sentence for both typical teachers answer and student answer.

**Step3:** The Porter stemming was applied as automatic stemming method, which removing suffix is implement like removing ED, ING, ION, IONS, which reduce the size of text and also the returned word to it origin will be very helpful in unifying because it is deepened on the base of the word. As example (connected, connecting, connection, connections) are all stemming to connect.

**Step4:** In this step after return list from stemming, the system should unify each word in stemming list and return it to original word based on data set which contain 30259 records, which is contain two columns, first column is the original words and the second contain the synonyms which may hold a lot of synonyms separated by comma that the English language dictionary match.

**Step5:** To specify the lexical similarity between vocabularies in students' answers and the typical answer, the preprocessed text used in this system as an input matrix in TF-IDF model, where each term in the document is related to specific weight decide the importance of the term in the document, where the document defined as D and represent in $W_i$= ($W_1$, $W_2$, $W_3$,…$W_i$), where $W_i$ is define as the weight of the term in each single document D. TF-IDF is a weight function for

terms resulted from the TF where in TF is extraction the frequency for both prepressed text and multiply by IDF matrix to result TF-IDF vectors.

**Step6:** This is the final step of evaluated for the examination where is measure the extent of correctness answer of student and comparing it with the typical answer by using cosine measure algorithm, which is calculate the distance between the two vectors that result it from TF-IDF model, where the first vector represents the student answer and second vector represents the typical answer where both are after manipulate in early steps.

## IV. RESULTS

Here in this section the results of proposed system are illustrating, which the system returns good results where here in this question as example the answer of student get 0.81 from final step in cosine similarity that because the answer of the questions does not consider the order of words and detect the synonyms.

| Question | Typical Answer | Student Answer |
|---|---|---|
| What are the types of Communication channel? | "Guided media twisted pair, coaxial cable, fiber optic, the electromagnetic waves are guided along a physical path. Unguided media also called wireless the transmitting electromagnetic waves are not guided along with a physical path. They are radiated through air/vacuum/water" | "Guided media twisted pair, coaxial cable, fiber optic, and the electromagnetic waves, Unguided media also called wireless the transforming electromagnetic waves are not guided along with a physical path." |

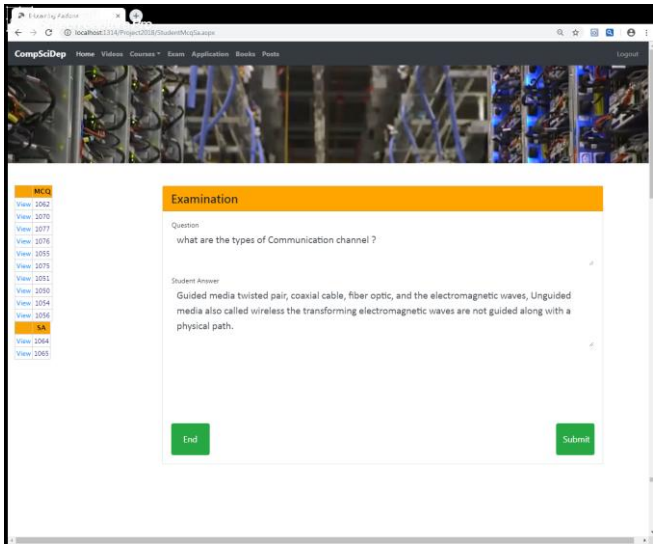


**Figure 2: Automatic Scoring Essay System**

## V. CONCLUSION

Depend on the experimental result and analysis of the system performance, we concluded that the system of automatic scoring essay examination is gives a good result when it relies on synonyms, the accurate process of preprocessing steps, and the order of words don't take in the account all these factor is strength the answer degree depend on the cosine similarity.

## REFERENCES

[1] M Ali Fauzi, Djoko Cahyo Utomo, Budi Darma Setiawan," automatic essay scoring system using n-gram and cosine similarity for gamification based e-learning", International Journal of Mechatronics, Electrical and Computer Technology (IJMEC), vol. 5, no. 15, pp: 2026-37, 2017.

[2] Aljameel, Sumayh S., James D. O'Shea, Keeley A. Crockett, and Annabel Latham. "Survey of string similarity approaches and the challenging faced by the Arabic language." In 2016 11th International Conference on Computer Engineering & Systems (ICCES) IEEE, pp: 241-247, 2016.

[3] V. Gurusamy, and S. Kannan,"Preprocessing Techniques for Text Mining", In Conference Paper. India, 2014.

[4] Singh, Vikram, and Balwinder Saini. "An Effective tokenization algorithm for information retrieval systems." Departement of Computer Engineering, National Institute of Technology Kurukshetra, Haryana, India (2014).

[5] Raulji, Jaideepsinh K., and Jatinderkumar R. Saini. "Stop-word removal algorithm and its implementation for Sanskrit language." International Journal of Computer Applications, vol. 150, no. 2, pp: 15-17, 2016.

[6] R. Lourdusamy and S. Abraham, "A Survey on Text Pre-processing Techniques and Tools," International Journal of Computer Sciences and Engineering, vol.6, no. 3, pp:148–157, 2018.

[7] Gunawan, D., C. A. Sembiring, and M. A. Budiman. "The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents." In *Journal of Physics: Conference Series*, vol. 978, no. 1, p. 012120. IOP Publishing, 2018.

[8] George, S., and Shibily Joseph. "Text classification by augmenting bag of words (bow) representation with co-occurrence feature." *IOSR Journal of Computer Engineering, vol.* 16, no. 1, pp: 34-38, 2014.

[9] Iadh Ounis, "Inverse Document Frequency", Reference Work Entry Encyclopedia of Database Systems, pp:1570-1571, 2009.

[10] Sindhuja, B., and Veena Trivedi. "Usage of cosine similarity and term frequency count for textual document clustering." *International Journal of Innovative Research in Computer Science & Technology (IJIRCST)*, vol. 2, no. 5, pp: 9-12, 2014.

[11] Wang, Meng-Jiao, and Yong-Zhen Li. "Hash function with variable output length." In *2015 International Conference on Network and Information Systems for Computers*, pp. 190-193. IEEE, 2015.

[12] Bruce Schneier,"One-Way Hash Functions." Applied Cryptography, Second Edition: Protocols, Algorthms, and Source Code in C, 20th Anniversary Edition, pp:429-459. 2015.

[13] Hebah Rababah, Ahmad T. Al-Taani "An Automated Scoring Approch for Arabic Short Answers Essay Questions*" International Conference on Information Technology (ICIT)* pp.697-702. IEEE, 2017.

[14] R. Kohila, Dr.K. Arunesh "Text Similarity Measure for News Articles Based on String-Based Approach." *Global Journal of Engineering Science and Research Management* pp.2349-4506, 2016.