ONLINE ASSESSMENT OF SIMILARITY BETWEEN SENTENCES IN QUESTION ANALOGUE SYSTEMS

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

NEHA KUMARI

11406169

Supervisor

LOVENEET KAUR

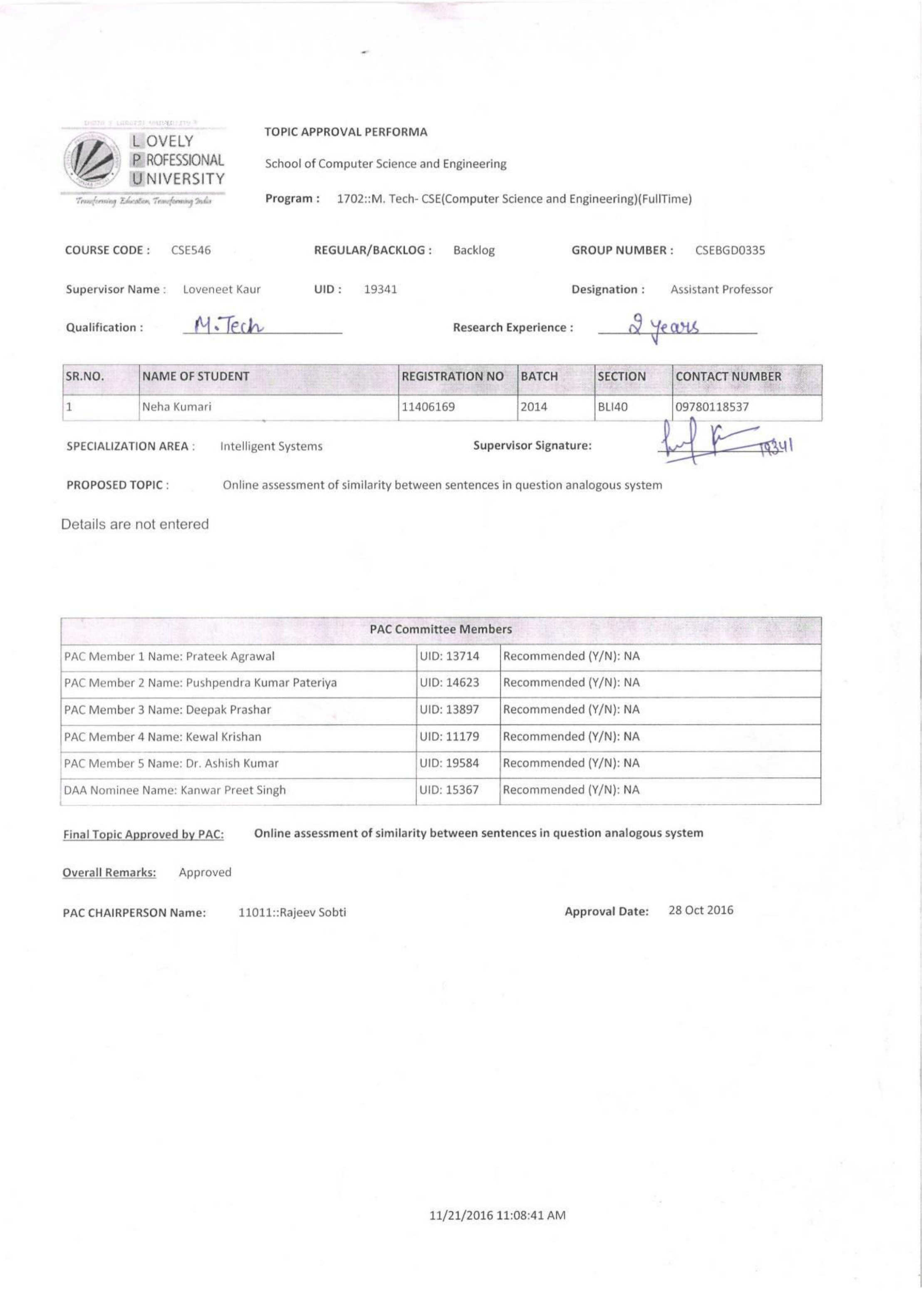
School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

November 2016

PAC FORM



LOVELY PROFESSIONAL UNIVERSITY
Transforming Education, Transforming India

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : 1702::M. Tech- CSE(Computer Science and Engineering)(FullTime)

COURSE CODE : CSE546 **REGULAR/BACKLOG :** Backlog **GROUP NUMBER :** CSEBGD0335

Supervisor Name : Loveneet Kaur **UID :** 19341 **Designation :** Assistant Professor

Qualification : M.Tech **Research Experience :** 2 years

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Neha Kumari	11406169	2014	BLI40	09780118537

SPECIALIZATION AREA : Intelligent Systems **Supervisor Signature:** 19341

PROPOSED TOPIC : Online assessment of similarity between sentences in question analogous system

Details are not entered

PAC Committee Members		
PAC Member 1 Name: Prateek Agrawal	UID: 13714	Recommended (Y/N): NA
PAC Member 2 Name: Pushpendra Kumar Pateriya	UID: 14623	Recommended (Y/N): NA
PAC Member 3 Name: Deepak Prashar	UID: 13897	Recommended (Y/N): NA
PAC Member 4 Name: Kewal Krishan	UID: 11179	Recommended (Y/N): NA
PAC Member 5 Name: Dr. Ashish Kumar	UID: 19584	Recommended (Y/N): NA
DAA Nominee Name: Kanwar Preet Singh	UID: 15367	Recommended (Y/N): NA

Final Topic Approved by PAC: **Online assessment of similarity between sentences in question analogous system**

Overall Remarks: Approved

PAC CHAIRPERSON Name: 11011::Rajeev Sobti **Approval Date:** 28 Oct 2016

11/21/2016 11:08:41 AM

ABSTRACT

Similarity is defined as a method which is used to measure or compare two things in order to find similar patterns between things like text, people, sentences, questions etc. It is the one of major time consuming process in which duplicacy or redundancy occurs in large sets of a data. But here our main focus is to compare or map two questions and measure similarity between them with the help of a technique called syntactic similarity. Syntactic similarity play an important role in the text analysis, text document, data mining and natural language process (NLP). Text analysis is used to find out the similar text between the two texts and it is an important text related research used in topic detection, topic tracking, question gernating, question answer system etc. Syntactic similarity is used for pattern matching algorithms to find the similarity between the sentences. Pattern matching is a novel technique that is being used in various areas such as, processing of signals, computer vision, video and image processing. The main idea behind the matching is to find one or more occurrences of a string in another string. Searching the database is one of the core problems in string matching. String matching has also been used as an integral tool for both theory and practice in various applications of artificial intelligence.

In this thesis, we are presenting multi pattern string matching algorithms like Rabin-Karp, Naive Based and Boyer-Moore which are used to find similarity using the tool MATLAB. Our proposed algorithm is better than the already defined algorithms in terms of accuracy and similarity and produced better results than the other algorithms.

Experimental result shows that our proposed algorithm generated more similarity and accuracy as compared to other algorithms that we have used for string comparison in our system. Proposed algorithms similarity and accuracy rate is little bit more that is useful to remove the duplicacy or similarity in the questions analogues system.

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation entitled "**Online Assessment of Similarity between Sentences in Question Analogous System**.” in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Ms Loveneet Kaur I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University’s Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Neha Kumari

11406169

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.tech Dissertation entitled "**Online Assessment of Similarity between Sentences in Question Analogous System.**" submitted by **Neha Kumari** at **Lovely Professional University, Phagwara, India** is a bonafide record of his original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

(Loveneet Kaur)

Date: 28-11-2016

Counter Signed by:

1) **HoD's Signature: ________________**

HoD Name: ___________________

Date: ___________________

2) Neutral Examiners:

(i) Examiner 1

Signature: _______________

Name: __________________

Date: ___________________

(ii) Examiner 2

Signature: _______________

Name: __________________

Date: ___________________

ACKNOWLEDGEMENT

I owe a debt of deepest gratitude to my dissertation supervisor, Assistant Professor, School of Computer Science And Engineering, for his guidance, support, motivation and encouragement throughout the period within which this work has been carried out. His readiness for consultation at all times, his educative comments, his concern and assistance even with practical things have been invaluable.

I am grateful to all faculty members who helped me by providing me the necessary opportunities for the completion of my work. I also thank the other faculty and staff members of my department for their invaluable help and guidance.

Last but not the least I am very lucky for having wonderful support of my friends and family members especially my parents. Thanks to almighty for providing me such invaluable gifts, support, enthusiasm and courage.

Place: Lovely Professional University Neha Kumari

Date: 11406169

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

API	Application Programming Interface
BF	Brute Force
BM	Boyer Moore
DB	Data Base
DNA	Deoxyribonucleic Acid
E	Entailment
ESL	English Second Language
FAQ	Frequently Asked Questions
FLFC	First Least Frequency Character
HLDA	Hierarchical Latent Dirichlet Allocation
H	Hypothesis
KMP	Knuth Morris Pratt
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
NLP	Natural Language Process
PCC	Person Correlation Coefficient
QA	Question Answer
RK	Rabin Karp
RTE	Recognizing Textual Entailment
RSMA	Recursive Based String Matching Algorithm
SSSR	Sentence Selection with Semantic Representation
SRA	Semantic Role Annotation

SVM	Support Vector Machine
SW	Stop Word
T	Text
WN	Word Net

\

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Pattern matching is the field of computer science that deals with matching the patterns and produces a good or exact result. Pattern matching tells about how to find the exact patterns in the strings and match particular one pattern with others and check the matching between the patterns. Patterns are generally of two types, one is sequence and other is tree structure. Sequences are described by some regular expression whereas tree structures are used in the programming language. Pattern matching has a wide range of applications that are used in the Natural Language Process (NLP), web search engine, Spam filter, word processor etc. [1]. Pattern matching has been used in various string matching algorithms, some of which are Rabin-Karp algorithm, Knuth-Morris, Boyer-Moore and Naïve-Based. Pattern matching algorithm is very useful of the algorithm to find out the similar pattern from the text these algorithms are widely used for many problems, string matching algorithms are good to matching. These algorithms are used to find out the similar patterns in the string and useful to solve the problems of similarity matching between the text, pattern or questions.

Similarity between the given string and pattern is called string pattern matching. Let take a string “What is computer?” and a pattern “What is it?”, now we have to find whether the words in the second string is obtained at some position in the first string or not. If the words in the first string and the second string is matched we will find the similarity between both the Strings. There are two type in which string matching is divided. Perfect pattern matching algorithms and string matching algorithms [2]. Instead of approximate pattern matching algorithms, exact pattern matching algorithms have various kinds of applications. These exact string matching algorithm will find out the similarity between both the given strings in an accurate manner, while the approximate pattern matching algorithms gives the nearest similarity value between both the strings. The perfect pattern matching algorithms are again divided into two types namely Single and multi-patterns of matching algorithms. Multi pattern matching have more realistic and practical applications in various fields such as, database search, intrusion detection and prevention.

There are multiple numbers of pattern matching algorithm that used to obtain the matching between the string and text. Pattern matching algorithms like KMP, BF, and RK etc.

Each of the algorithms to work on a differently bases and aim to have find the similar patterns in the strings.

String matching algorithms are work on a two things pattern and text in which to given text and search a pattern in the text.

Example of String Matching:

Text: ABCABCABCABCDEAB

Pattern: BCDE

In this chapter we will discuss about the introduction part in which pattern matching, similarity, sentence similarity, applications, drivers and database and MATLAB tool that used in this thesis work.

1.1 Similarity

Similarity is concept which has defined the similarities and we can say that it tells that how the two things are similar in each other's and also defined the relevant information between them.

Example:

- Similarity between two questions.
- Similarity between two sentences.
- Similarity between two texts.

1.2 Sentence Similarity

Sentence similarity is defined as in which similarity between the two sentences and we can say that it produce the same information when to compared the two sentences [3]. Sentence similarity is useful in various area like information system, text mining, question answer system etc. the diagram is show that how to calculate the sentence similarity. Sentence similarity is very useful to solve the problems in data mind that is related to in text and find the duplicacy in the data and also very useful in others area.

There is a figure of sentence similarity that shows that how to be calculate the similarity just a simple view given in the diagram and shows that how to display the output of sentences using similarity.

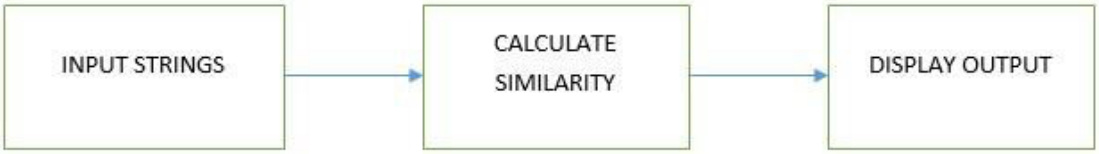


Figure 1.2 Calculating Sentence Similarity

1.2.1 Types of Sentence Similarity

There are three types of the sentence similarity measure:

(i). Statistical Measure

The statistical similarity calculated for a sentence depends on symbolic characters of the sentences and gives the data of structure. Statistical similarity measure the similarity between the sentences but one thing is that it takes only a statistical information of any sentences. It can be measure the similarity of word counts of the sentences. For Example:

Text 1: What is Computer?

Text 2: what is Robotics?

(ii). Semantic Measure

Semantic similarity defined to in which have a different structure information and symbolic information and could give the same meaning information. Semantic similarity in the sentence is based upon the meaning of the word that is in the sentences and syntax of the sentences. It works on the synonym and meaning of the words. This is very important of matching the similarity in several area of database.

For Example:

Text 1: Our sports teacher play Cricket very good.

Text 2: Our sports tutor play Cricket very good.

(iii). Syntactic Measure

Syntactical similarity is one part of text analysis, it might be misunderstand what is actually means. The text can have many information hidden into itself. Syntactical structure have obtain information's that are hidden. Syntactical means structure of the words and phrases.

A common analysis type (much more complex though) is lexical analysis which is analyzing meaning of the text.

Example 1:

Text 1: LPU is the good university to study in.

Text 1: Studying in LPU is really good.

Example 2:

Text 1: India is the great city to live in.

Text 2: India is great city to live.

These two example clearly defined that both of the texts are same in meaning if check the similarity then these two examples produce a same meaning because in the syntactic similarity always work on the word to word it does not check the synonym of the words so these examples shows the syntactic similarity between the two sentences.

1.3 Syntactic Similarity

Syntactic similarity is the concept of measuring the similarity to words to words. Syntactic similarity does not use the synonym of the words it only focus to measure the word in the sentences, but there are many small word used in the sentences like an, the, this, there etc. then the syntactic similarity not count these all the words and to used stop word (SP). Syntactic similarity to very useful in various applications to text mining, data mining, and information retrieve etc. There are the example of the syntactic similarity. First is where i can put on? And second question is where i can put in? In below the explanation of the diagram and the syntactic similarity how to compare as shown in the figure and one most important thing is that the syntactic similarity is worked on the word to word similarity and to compare the string on the bases of word to word and ignore the repeated, an etc. words in the sentences. It is very useful to in finding the similar objects in the data and always useful in real life applications and solve the problems related to similarity. In this thesis is similarity is calculated on the bases of syntactic similarity between the two questions that find out the relevant information between two questions and accuracy too. In the diagram to calculate the syntactic similarity.

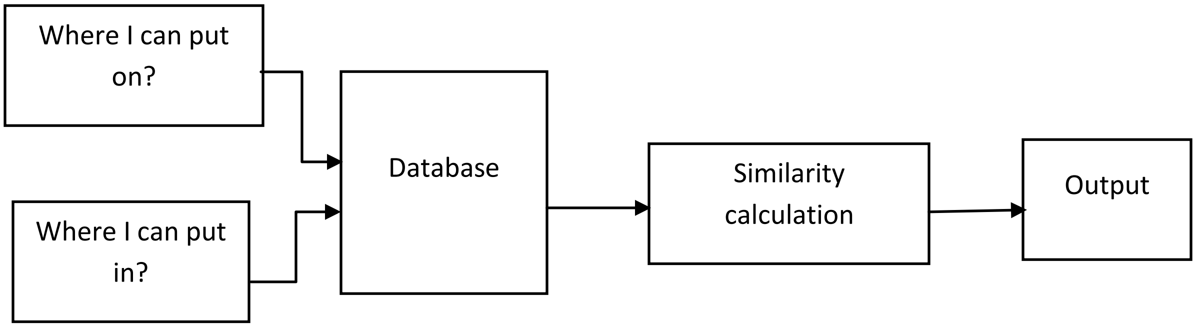


Figure 1.3 Calculating Syntactic Similarity

This diagram calculates the syntactic similarity between the two questions. There are two questions Present here and then they goes to the data base after data base to calculate the similarity between two questions that how much these two questions are similar and here similarity is calculate on the bases of word to word. And finally to produce an output that shows that how much these two sentences are similar in nature and also to find that how much these two questions are accurate.

1.3.1 Syntax and Semantics Similarity

Syntax is the symbolic representation whereas the semantics means the meaning of the given statement. In other language, if we implement the two programs written in the different language, could work the same thing is called the semantic but the symbols which are used to implement a program would be different is called the syntax. The role of the compiler is check the syntax i.e. compiles time error and derive the semantics from the language rules but don’t find all the semantic errors There are three level of syntax:

Lexical level, Grammar level and the context level which determine that what the variable name and the object name define to and check that whether the types are valid or not? In the computer language semantics are used to define what they actually program work or compute. Then those semantics will one to one mapping between how the user interface wants and how actually it work.

1.3.2 Syntactic vs. Semantic Similarity

Semantic similarity is the term in which the meaning of the given sentences are same that mean it work on the meaning of the sentences. And it is find the similarity on meaning and synonym of the two words.

Example:

"Our sports teacher play Cricket very good"

"Our sports tutor play Cricket very good"

In the given example here the meaning of both the text are same.

On the other hand syntactic similarity is the part of text analysis. It means the structure of the words that have to be given or the phrases. In the syntactic the meaning doesn't matter, here only similarity will occur when the word to word is match.

Example:

"I am staying at home"

"I am staying at house"

In the several field in which phase to the problem is duplicacy of the data. These fields like data mining. Information retrieve and text mining etc. And to obtain the similarity between two documents of sentences. There is a big advantage of the similarity in the frequently asked question (FAQ) system. Firstly to find the question sentences from the questions and answer data and to be return the correct answer for a user.

1.3.3 Use of Similarity in Data and Text Mining

- **Data Mining**

It is a method in which to used and examine a large of data and obtain hidden and important data that have to enhance business competence. Many industries to take advantage though data mining in a business development. Application of data mining are widely used in the market, bank, and health department transportation etc. The big of the aim to use the similarity in a mining concept that mine the data. Similarity is based here on the base of distance of small and large in high similarity in a small distance and low in the big of distance.

- **Text Mining**

Text mining is the process of computerized analysis of one text or a number of documents (corpus) and extracting unimportant information from it. The main importance of Text Mining is to absorb the method of transforming unstructured textual data into structured data representation.

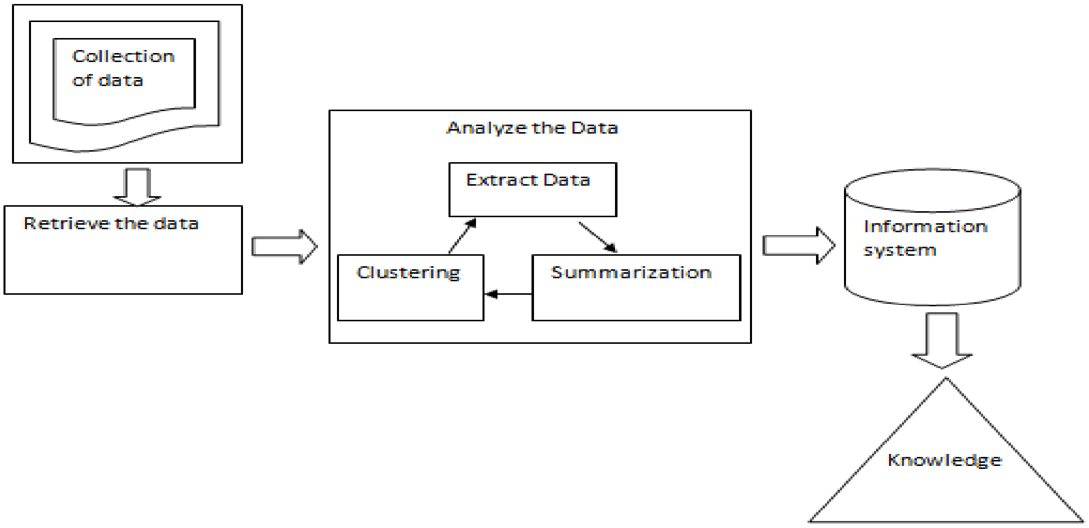


Figure 1.4 Text Mining

The results can be analysed to determine useful knowledge, some of which would only be establish by a human reading and analysing the data.

Text similarity is divided into the five steps: Collection of data, Retrieve the data, Analyse the data, clustering and summarization, Information system, Knowledge.

First to collect the data into the various sources and then retrieve the data after to be analyse the data that will be done with two step first to clustering the data and then summarization. After this process the data is send into the information system in which to have a useful of the data that find with in all pre-processing and finally to in a knowledge in which to be have a knowledge able data is to be presented and analyse the useful knowledge. These all step the text similarity is to be find out. All the step are necessary to complete the process to analyse the data and for a useful information and knowledge.

1.4 Applications of Syntactic Similarity

Syntactic similarity has various kinds of applications that are used in different kind of real life scenario [4]. Some of the applications following as:

- **Biometric Informatics**

Semantic similarity has been applied in biomedical informatics. In this application semantic similarity is used in gene ontology. Semantic similarity in biomedical informatics is used to obtain or compare the similarity of gene or also to used proteins and sequences that are similar in nature. Semantic similarity to find out the similarity in between the gene and proteins.

- **Natural Language Process**

In a field of computer science NLP is widely used process. It can be used in various areas like sentiment analysis. In this similarity is found by the semantic web that provides semantic extensions that find similarity by content not by arbitrary descriptions.

- **Geo-Informatics**

In this application semantic similarity is used to obtain the similar geographic features or their types. SIM-D, OSM semantic serve and similarity calculator used in the Geo-informatics to calculate the similarity.

- **Computational Linguistics**

Database that are constructed manually and always have human supervision those type of database are not automated and cannot measure the similarity between two terms in the data base. In this Word net is used for compare the strings.

1.5 String Matching Algorithms

Here to present the algorithms and tell about their application, theses algorithms namely Naïve Based, Rabin-Karp and Boyer Moore that are used for a string matching between the sentences and also very easy and simple to use in pattern matching.

1.5.1 Naive Based Algorithm

The naïve based approach for string matching is a very basic approach. Naïve based is easy to understand and implement, but in some of the cases, the naïve based algorithm works too slow. It will take the worst case complexity of iterations (n*m), if the text length is "m" and pattern length is "n" for completing this work. The idea behind the Naïve-Based string matching is just to compare each character of a text T [s...s + m-1], pattern P [0…m-1]. It performs various shifts and returns all the shifts which are valid. Naïve based algorithm is like a brute force algorithms and used for a string matching there are many field in which algorithms is used and find out the similarity between them that is most important of the task. So it is very simple and easy to use. Naive-Based algorithm have based upon string matching that always perform a good results and applied in the real time applications like real time predictions, sentimental analysis so on.

It have many fields in which naive based is used and play a good role to find the strings into the patteren.it also used for a text classification and have to find the string in the text.

These are the applications of the NB algorithm:-

- Real time prediction,
- Multi class prediction
- Recommendation system
- Text classification
- Spam filtration
- Sentimental analysis

1.5.2 Rabin Karp

Rabin – Karp is a string matching variant algorithm in which hashing is used for string search. For finding any of the set of patterns in the given string it uses hashing technique.

For length of one string “n”, for a set of patterns the length of “m”, the average time complexity for the algorithm is *O* (*N+M*) with a space complexity of O (P). The best case running complexity of this algorithm is also same as that of the average case whereas, the worst case running complexity can be O (NM). By using Rabin-Karp, for a pattern P [0…m-1] we will calculate a hash function *h(x)*. By using the obtained hash value we will find the match of every substring and length is m-1 of the string. Rabin-Karp algorithm used for a multiple of pattern search and used the hash function and using hashing for shifting substrings search that is useful technique to find out the patterns in given string.

These are the application of RK algorithm:-

- Text Processing
- Bioinformatics
- Compression

1.5.3 Boyer Moore Algorithm

In Boyer-Moore algorithm the pattern matching have to work right to left. By using BM, we can skip number of characters as compare to previous algorithms. if we take the example that in which character of first matched with text that is not in the pattern P [0...m- 1], here to skip the m character we can do this again and again. As the Knuth-Morris-Pratt (KMP) algorithm, pre-processes the string to find a table in which have a knowledge to leave a character to each of the pattern Boyer-Moore (BM) to maintains alphabets in the table which contains as many as characters in the string.

These are the application of BM algorithm:-

- Search engine
- Bad character heuristic
- And good character heuristic

1.6 Drivers

1.6.1 JDBC Driver

Java Database Connectivity (JDBC) is used a programming language like java and it is an application of API. JDBC is a client side adapter not a server side that is install by a client side and to convert the request from a java program protocol that to DBMS understand it. There are the four types of JDBC in which type1 call native code of locality available ODBC driver.type2 that call the data base native vendor from the client side have java driver is talk with the server middleware and that then talk to data base type4 pure java that use data base native protocol.

1.6.2 ODBC Driver

Open Database connectivity, is an important application (API) for used in the Data Base Management System for accessing the data, the aimed of designer ODBC is to make the system independent of data base system and operating system. With the few changes of a data base and application written using the ODBC can be written in both side of client and server and provide for other platform only a few changes of data access code.

1.7 Microsoft Access

Microsoft Access is a database management system (DBMS) given from Microsoft that mixes the relational Ms-Jet Database Engine with a visual user based interface and software-developing tools. It is a member of the Ms-Office stack of applications, included at the Professional level and higher editions of it or sold independently. Ms-Access stores information in its own format depending on the Access Jet Database. It also imports or links directly to a data stored in other application and database. A Software developer, data architect and power user can also use Ms-Access to develop applications soft- ware. Like other Ms-Office applications, Access is given with graphical Basic for Application (VBA), an object-oriented language that can refer varieties of objects which include DAO (Data Access Objects), ActiveX Data Objects, and many other ActiveX components. Graphical object used in a form and a report expose the method and property in the VBA programming environment, and a VBA code module may call operating system operation.

1.7.1 Uses

Instead of using its owned database system for storage, Ms-Access may be used as the 'front end or graphical interface' of a system while other programs act as the 'back end or storage' table, like My SQL Server and non-Microsoft product like Oracle and Sybase. Multiple backend sources can also be used by an Ms-Access Jet Database (ACCDB and

MDB formats). Similarly, other application like as VB, ASP.NET, or Visual Studio .NET will use the Ms-Access databases format for table and query. Ms-Access can be part of a more complicated solution, where it can be combined with other technology like Microsoft Excel, Microsoft Outlook, Microsoft Word, Microsoft Power- Point and ActiveX controls. Access table and also support a lot of standard attribute type, indices, and referential integrity including cascading updates and deletes. Access also has a interface for queries, form for displaying and entering information, and report to print. The underlying database, that is contained these object, is multi-users and handle record-locking. Repetition of task can be generated automatically by using macros with point-and-click options. It is very easy to put a database system on a networking and can be used by multiple user share and up-date information instead of overwriting other works. Information is locked at the lower level which is acts differently from Excel which lock down the entire spread sheet.

1.7.2 Features

A User can make a table, query, form and report, and connect with each other with the help of macros. An Advanced user can also make use of VBA to make other solution with information of high quality for manipulating and controlling users. Ms-Access also has a property of generating reports for creating feature that work with any information sources that Ms-Access can use.

The main idea of Ms-Access was for users so that they will be able to use information from any target. Other feature including: the import and export of information to many formats that includes Excel, Outlook, ASCII, and dBase, Paradox, FoxPro, SQL Server and Oracle. It has also the ability that links information in its existing locations and uses it for viewing, querying, editing, and reporting. This allows the main information to changes while enabling that Ms-Access use the new information. It also performs different join between information set stored differently platform. Ms-Access is used by people for downloads information from low level database for manipulating, analysing, and report locally. There is another Jet Database format (MDB or ACCDB in Access 2007) which contains the applications and information in files. This makes it very easy to send the whole application to other users, who can use it in disconnected environments.

One of the advantages of Ms-Access are a programmer’s perspective is that its relates compatibility with SQL (structured query language) — query can be viewed visually or edited as SQL statements, and SQL statements can use directly in Macros and VBA Modules to manipulate Ms-Access tables. User can combine and uses VBA as well as “Macros” for programming form and logics and offer object-oriented concepts. VBA can also be included in queries.

1.8 Software’s and Database

We are using MATLAB for programming and development of the GUI, storing strings we are using Microsoft access database and we connect the database with MATLAB using JDBC, ODBC driver.

CHAPTER 2

LITERATURE SURVEY

This chapter presents an overview of literature pursued for the study. It describes what work has been done for the present aim. It helps in understanding what is the lacking or missing in the techniques that can be improved.

Yuhua Li et.al. [1], introduced the concept of semantic similarity. It is used in the area of a text mining, information extraction, and dialogue systems. Previously similarity was measured in the form of long text but here similarity is measure in form of short text. Firstly semantic similarity is obtained from lexical database and the corpus. Lexical knowledge is based on the knowledge of the word in human language. The corpus shows the actual use of language and word. We focus not only on the common human knowledge but using corpus applications also. Secondly we consider impact of the order of the word on sentence meaning. Different word and number of word pairs in a different pair.

Wanpeng Song et.al. [6], proposed a method for measuring the similarity which is measured using statistic similarity & semantic similarity. Experimental results shows that the similarity measured by the proposed technique is better than existing techniques

MuthukrishananUmamehaswari et.al. [7], proposed a technique for measuring the similarity between sentences using semantic techniques for calculating similarity based on reformulation between two sentences. The experimental results show using semantic techniques depending on reform helps to enhance the working capability of Question Answer systems.

Zhong Min Juan [8], proposed a technique in which the technique of word co-occurrence corpus is used to get better performance for the purpose of comparing questions and answers. Firstly a knowledge base based on semantics is built called the “Word co-occurrence corpus”, then count up frequency for the sentences is calculated using statistical and semantically driven techniques.

Jun Sheng Zhang et.al. [9], Proposed two techniques, first of which is aimed at calculating statistical likeness among sentences depending upon symbolic information and structural information. The second one is that the sentence similarity is calculated based on set of words and similarity of sentences as the word order can capture extra local information about the sentence pairs.

Palakorn Achananuparp et.al. [10], proposed a technique that measures the similarity between the sentences. There are large number of applications like question answering, text mining and text summarization. Sentence similarity is to be calculated using Word overlap measures, simple word and IDF overlap, jaccord technique, phrasal Overlap measures, TF-IDF Measures TF-IDF Vector Similarity and Linguistic Measures, semantic similarity measures for sentences, word ordering similarity**,** the Combined Semantic and Syntactic Measures.

Prathvi Kumari et.al. [11], suggested a technique to measure the semantic similarity of two words. Information is available on the internet and to use the techniques that make usage of page-count and snippet to calculate the semantic similarity. Large number of word co-occurrences are defined using the page count and are integrated in the lexical pattern extracted from the text snippets. Pattern extraction and clustering methods are used for a numerous semantic relation between the two or more words.

Partha Pakray1 et.al. [12], suggested a technique of textual entailment that recognizes the systems that use lexical and syntactic features. TE is a rule based system. Textual Entailment is a relationship of pairs and textual expressions, entailing “text” (T) and entailed “hypothesis” (H). T is entailing H if the means of hypothesis H can be obtained from the means of text T.

Enrique Alfonseca et.al. [13], suggested that the previously proposed system had presented time constraint and had an incomplete prototype. So we present a system using the syntactic and semantic similarity that verifies the syntactic analysing of QA system and test other semantic distances metrics to churn out more accurate results and integrated system for the future.

Kai Wang et.al. [14], suggested a technique for defining the simple question. It depends on the syntactic tree based structure and is capable of solving the problems of similar matching questions. Yahoo answer, question matching, syntactic structure, QA keywords are used for this.

Wael H. Gomaa et.al. [15], proposed a technique for measuring the text similarity that divides the text similarity into three approaches 1. String based 2. Corpus based 3. Knowledge based similarity. Text similarity is very important for text based research and related applications, like information retrieval, document clustering, topic detection, topic tracking etc.

Anterpreet Kaur et.al. [16], suggested that Syntactic similarity is an important area of text document, data mining, and natural language process. Proposed technique is to be introduced in the system in which it is not possible to change the order of the word and languages are not dependent, to calculate the similarity between the questions in two questions paper. But it may happen that questions are related to each other. So we ignore this type of problem in proposed system in which our system may be able to know the similar question in the paper and will be able to find those questions. So the possibility of similar questions is decreased in the future.

Ercan Canhasi [17], suggested a technique that uses to measure the similarity of short English texts, specifically of sentence length. The proposed technique is used to measure semantic and word order similarities of two sentences. In order to perform this, it uses a structured lexical knowledge base and statistical information from a corpus knowledge base. The suggested technique performs well in determining sentence similarity for most of the sentence pairs, consequently the proposed technique will be used in computer automated sentence similarity measurements and other text based mining problems.

Zhao jingling et.al. [18], proposed a technique for calculating the sentence likeness which is partitioned into a three parts. In first part word semantic similarity is obtained and in second part semantic resemblance among sentences that is based on sentence structure and semantic similarity of the words is obtained. Finally the order of words for sentences similarity combined semantic similarity plus word ordering similarity is calculated as the absolute similarity among the sentences. To use word similarity techniques which is divided into two group corpus based technique and dictionary based technique.

Xiao-Ying-Liu et.al. [19], proposed a technique which is used to map the two applications with existing one. Sentence semantic structure is used to overcome the problem of variability language expressions. Pair of verb arguments represents a sentence except frames that are smaller structure of frames. So combining the verb - argument pair and words similarities count which depends on Word Net from which total sentence similarity is measured, removing the result of semantic gap. These two approaches for calculating the similarity between two sentences are superior as compared to existing one. In future will carry out other applications such as text summarization and question answering.

U.L.D.N Gunasinghe et.al. [20], proposed a technique for measuring the sentence similarity. This technique depends on semantic and syntactic methods of sentence similarity. The technique takes into consideration a vector space model for calculating the sentence similarity, the vector space model is discovered at the nodes in the sentence. This technique has two parts in first part we consider relation between verbs and in the other we consider relation between nouns in the sentence.

Chi Zhang et.al. [21], proposed a technique namely sentence selection with semantic representation (SSSR). SSSR uses well developed procedures to consider summary of sentences. The procedure for selection used for SSSR is to consider sentences that reform the initial documents with negligible distortion with well-planned combinations. This model makes use of two selection procedures weighted mean of word's embedding and deep coding.

Asli Celikyilmaz et.al. [22], proposed two techniques Latent Dirichlet Allocation (LDA) and Hierarchical LDA (HLDA) to discover the hidden concept and introduced a set of methods based on LDA to count the similarity of questions and candidates passages that are used for ranking results. Result of this article shows that retrieving information from secret concepts that enhance the results of a classifier – based Question Answer model. In this method we use a small sub set because of a computational cost. Increasing the number of training sample to find the more accurate results. In approaching times rather than using IBM model one to study advanced methods that increased the accuracy of the systems and also make a plan to employ the translation probabilities learned from the QA archive for document extracting experiments.

Megha Mishra et.al. [23], Proposed a method which combines the three methods that is semantic, syntactic and lexical and it uses a SVM classifier, with the help of this classifier it is used to advance the accurateness of a system.

Shashank et.al. [24], suggested a technique to measure the similarity and jaccord technique it is used with the help of this to improve the accuracy with compare of previous technique.

Wan-Yu Lin et.al. [25], have introduced a framework that is used to identify the online plagiarisms detection. Three features are used viz. syntactic, lexical and semantic that consists of POS, reordering, word alignment, duplication-gram, Phrase tags and semantic similarity sentences. To compare these feature with online plagiarism detection system and produced result that is more sophisticated and how human used the online plagiarism source. The system is language independent that mean work both English and Chinese version for evaluation.

Hiranya Jayathilaka et.al. [26], proposed a new method that automatically analyses the API similarity and quantifies application porting effort to use in a simple system and Python language in which API document developer syntactic and semantic aspect of API operation. To present an algorithm that consume and analyses the feature and automatically detect that two API feature and tell that is syntactically god or not or what difficulty to port of application among them. Our approach use both randomly generate and real word API. Our metric capture the difficult that developer associated with porting the application from one API to other.

Ludmila Cherkasova et.al. [27], have used the four syntactic algorithms, in which three algorithms are based upon Border Shingling and fourth is based upon content-based chunking. Performance studies disclose that similarity report of all four algorithms is extremely susceptible to the size of sliding window and frequency sampling parameter. So the work made use of a different version of a traditional sliding window algorithm except calculating the hash value of whole chunk here to use the numerically least fingerprint sliding window within this chunk. And to improve the performance for small document .in shingling-based algorithm sampling is requires that introduced the degree of vagueness and probability for false positive. So the basic sliding window algorithm produced a compressed file signature without sampling by using the chunk of hashes. The signature of file must have correct information concerning the file.

Akhtar Rasool et.al. [28], Author have introduced the different string matching algorithms and observed their performance .These algorithms are Naïve string matching, Knuth-Morris-Pratt, Rabin-Karp, Brute Force, Boyer-Moore, Aho-Corasick and Commentz Walter algorithm .Main idea of these algorithm is analysed and also compared the matching efficiency with time and speed. So, authors have analysed that performance is purely based upon the algorithms selected and bandwidth used. Result shows that multiple string matching algorithms are compared and in which some of the algorithms produced a good result and those are BM, Aho-Corasick, KMP algorithms are efficient and BM is fast for a large alphabet and KMP decrease the time as compare to brute force algorithms.

Mr. Rahul B Diwate et.al. [29], have studied about different algorithms' of pattern matching KMP, Naive string search algorithm, Knuth-Morris Pratt and Boyer-Moore pattern search algorithms. Nowadays everything is performed on the internet. Searching is a one of the operation that is performed by the user pattern matching is the one of the technique for searching each of these algorithms have their own characteristics. The BM, KMP are more effective algorithms Fast DTW algorithms are best for all images audios and videos patterns processing. KMP and BM algorithm have produced good result and complexity. KMP algorithms have lesser time complexity and BM have less pre-processing time complexity or Fast DTW have a linear time and space complexity.

Kenji Sagae et.al. [30], proposed an approach for derive the word cluster based on the syntactic similarity and also tells how these words cluster can be used in transition depending dependency parser. And improve parsing accuracy by using the two parser and unlabelled text that use the different way to use the syntactic structure experimental result shows that syntactic structure are efficient in leveraging cross framework and improved the accuracy of parsing. In future can improve the performance and accuracy with the help of other framework and others natural language process (NLP) task.

Su-Youn Yoon et.al. [31], proposed a method that measure ESL (English as second language) using syntactic competence and show that how to combine the core and find the use of the ESL with NLP techniques for purpose of automated scoring. Feature measure the range of grammatical expressions that is based upon the POS tag distribution. Corpus consisting of a large number of learner results were collected and divided into 4 groups. Syntactic competence of testing reaction was measured by identify the most of same group from the learner corpus. Speech recognition error results in a minor gap in performance that is an important advantage of our method.

Diarmuid O Seaghdha et.al. [32], proposed a novel method for incorporating syntactic information in the Probalistics latent variable models of the lexical choice and contextual similarity. Result of this approach captures the effects of context on the interpretation of the word and replacing that word with similar one. Two data sets are required in the demo but these two are potentially applicable in a range of application where the semantic disambiguation is required. In future can adapt this approach for word sense disambiguation as well as related domain-specific tasks.

Yasher Mahdad et.al. [33], proposed a method based upon the off-the shelf parser and semantic resource for recognizing textual entailment (RTE) model is used syntactic and semantic similarity for an RTE system without require the large automatic rule acquisition and hand coding in this lexical similarity produce lexical-syntactic rules automatically that is derived from the supervised learning rule. Syntax is encoded in parse tree and similarities are defined by WorldNet similarity measure. To experimentally show that Latent Semantic Analysis (LSA) derived lexical semantic embedded in syntactic structure is a good approach and the model is presented here is one the best system in RTE challenges. Compared to other method does not required the large set of handcrafted or corpus extracted lexical syntactic rules. For a better performance experimental result I compare with a previous one and this approach improve the baseline model.

Rafael Ferreira et. al. [34], proposed a technique for sentences similarity that helps to solve the problem by taking into lexical, syntactic and semantic analysis of sentences. In previous systems Word Net was used to determine the semantic word which gives improper result. In this work Semantic Role Annotation (SRA) [35] is used to retrieve the semantic word and two traditional method of Pearson's correlation coefficient (PCC) and Spearman's rank correlation coefficient (SRCC) is used and gives the better results.

Jehad Q. Odeh et.al. [36], proposed two techniques first one is the lowest frequency character algorithm (FLFC) and other one is recursive based string matching algorithm (RSMA). FLFC is an improved version of scan meant for low frequency characters proposed by Horspool [37]. FLFC Proposed technique was implemented, tested, compared and analysed with boyer-moore and naive brute force using a different data and size. Different techniques were tested using the same machine. The result was average. RSMA-FLFC algorithm enhanced the execution time as compared to brute force and boyar moor. Testing to calculate the effectiveness of proposed RSMA compared to FLFC without implementing the recursive techniques applying FLFC is more useful if it is merged with recursive matching techniques.

Jiwoon jeon et.al. [38], proposed a technique to automate collection of semantically similar questions pair from existing QA collection. Then taking the collections of bilingual and runs the IBM machines translations model 1 [39] to gain knowledge about word translations probabilities. To provide with a new questions, a translation based data extracting model exploited the word relation to extract similar question from QA archives. Different type of methods are used to resolve the mismatch problems between the questions that are knowledge based [40] which is machine readable dictionary, Employee manual rule and template [41], statistical technique develop the information extracting and natural language processing [42].

Nimisha Singla et.al. [43], author have introduced various string matching algorithm and their use in various application those applications are Bioinformatics, chemistry informatics, text editors in computer, DB query, wide window pattern matching, matching DNA sequence, digital library, search engine and others applications and string matching algorithms are Boyer Moore, BM Horspool, Brute Force, KMP, Quick Search, Rabin-Karp, Approximate String matching, Smith waterman, Needleman and wunsch. Each algorithm is applied on an application and one application is explained with an optimal algorithms. In a result to find out that Boyer Moore algorithms have a less time complexity and BMH; KMP algorithms have a less pre-processing time complexity. Others algorithms are depend upon the input and those are good for particular applications.

2.1 LIST OF VARIOUS TECHNIQUES USED FOR CALCULATING SYNTACTIC SIMILARITY

This table describes comparison of various techniques used for calculating syntactic similarity between two word and sentences or between two questions in an analogous system.

S. No	Paper Name	Author Name	Technique Used	Description	Conclusion
1	An Effective Similarity Measurement for FAQ Question Answering System [8]	Zhong Min Juan	Semantic and Statistical methods	A method is proposed by combing semantic and statistical techniques that build semantic knowledge base, called as co-occurrence words corpus, then count no. of question sentence by using statistic method.	Combine the semantic and statistical method and compare with proposed method result show that proposed method is gives better performance.

2	A Syntactic Tree Matching Approach to Finding Similar Questions in Community-based QA Services [14]	Kai Wang, Zhaoyan Ming, Tat-Seng Chua.	Syntactic Tree Matching	The proposed method uses Syntactic Tree Technique to improve the accuracy rate as compared to the previously used methods for finding the accuracy.	8.3% accurate from previous methods BoW or plain tree kernel and 50% accurate if semantic features are used.
3	A Novel Approach For Syntactic Similarity between Two Short Text [16]	Anterpreet Kaur	LCS, Edit Distance and Bi-gram algorithms	The proposed method finds similar questions and removes the possibility of relevant question in future time.	Compare the proposed method with existing techniques. Result show that proposed method improve 70% accuracy rate.
4	Sentence similarity measuring by vector space mode [20]	U.L.D.N Gunasinghe, W.A.M De-silva, N.H.N De-silva, A.S Perera	Semantic and Syntactic methods	The proposed system can be used for variable length strings means the size of question is not fixed for calculating similarity.	Technique uses semantic and syntactic methods of sentence similarity that is more accurate than previous system.

5	LDA Based Similarity Modelling for Question Answering [25]	Jehad Q. Odeh	FLFC and RSMA algorithms.	In proposed system FLFC and RSMA are compared with Boyer-Moore and Brute Force and FLFC is found to be superior if merged with recursive matching technique.	RSMA-FLFC algorithm enhanced the execution time as compared to brute force and boyar moor and 50% improvement in accuracy rate from previous work.
6	The mathematics of statistical machine translation [28]	Megha Mishra, Vishnu Kumar Mishra and Dr. H.R. Sharma	Linear SVM Support Vector Machine	The proposed technique combines three features namely Semantic, Syntactic and Lexical with SVM classifier to improve the accuracy.	91.1 % for fine grain and 96.2 % accuracy rate for coarse grain.
7	Statistical Measure to Compute the Similarity between Answers in Online Question Answering Portals [33]	Shashank , Shailendra Singh	Jaccard Technique	The proposed system shows that the Jaccard technique is more efficient than any other statistical techniques.	Jaccard method for statistical measure gives efficient and accurate result than other techniques.

8	Online Plagiarism Detection through Exploiting Lexical, Syntactic, and Semantic Information [34]	Akhtar Rasool, Amrita Tiwari, Gunjan Singla,Nilay Khare	Naïve string matching, Rabin-Karp ,Brute Force, Knuth-Morris-Pratt ,Boyer-Moore, Aho-Corasick and Commentz Walter algorithm	Result shows that multiple string matching algorithms are compared and in which some of the algorithms produced a good result.	BM, Aho-Corasick, KMP algorithms are efficient and BM is fast for a large alphabet and KMP decrease the time as compare to brute force algorithms.
9	Using syntactic and semantic similarity of web apis to estimate porting effort[35]	Nimisha Singla, Deepak Garg	Boyer-Moore,BM Horspool, BF,KMP [44], Quick Search RK,Approxi-mate String matching, Smith waterman, Needleman & wunsch.	Applied on an application and one application is explained with an optimal algorithms.	Boyer Moore algorithms have a less time complexity and BMH, KMP have a less pre-processing time complexity. Others algorithms have good applications.

10	Applying Syntactic Similarity Algorithms for Enterprise Information Management [36]	Kenji Sagae and Andrew S. Gordon	Natural language process (NLP)	Proposed an approach to derive the word cluster based on the syntactic similarity and also tells how these word clusters can be applied in evolution based dependency parser.	Using the two parser and unlabelled text that use the different way of syntactic structure, Experimental result shows that syntactic structure are efficient in leveraging cross framework and improved the accuracy of parsing.

Table 2.1 Technique Comparison Table

These are the several of the paper that study in literature review which gave an idea about the similarity and string matching. These papers are basically depend upon the several of similarity, pattern matching and string matching between the two text, question answer system etc. In this study to analyse that there is no more work on the syntactic similarity that is totally different to the semantic similarity and other one.it does not work on the synonym of the word it is totally based upon word to word similarity.

CHAPTER 3
PRESENT WORK

3.1 Problem Formulation

An algorithm meant for string matching means to locate one or numerous occurrences of one string in another string. Searching the database is one of the core problems in string matching. String matching has also been used as an integral tool for both theory and practice in various applications of artificial intelligence**.**

String matching between the text, questions, sentences etc. has a big problem to study several of existing one paper and to find out the problem. Duplicity is the one of the problem in the data so with the help of syntactic similarity so solve this string matching problem. Question analogues system.

Previous works in this domain have focused that there is a problem of duplicity the data and pattern matching. Our work has been motivated by these problems and we have resolved these issues.

3.2 Objectives of study

String matching is the one of the important aspects in which to find the similar patterns. In many of the fields it is a big issue.However, there were some problems which exist and needed revision. Inspired and forced by that and after a comprehensive literature survey, our wok resolved some issues. This work achieved certain defined objectives and that are as follows:-

- Implements a novel approach for Syntactic similarity.
- To find the better similarity and accuracy rate.
- To calculate the syntactic similarity between two questions.
- To implement the proposal algorithm using MATLAB R2013a.
- To study various method of semantic and syntactic similarity.
- Use the three algorithms and compare with proposal algorithms.
- Check the similarity between two question that how much similar they are.
- Use the string matching algorithm for calculate the similarity between questions.

3.3 Research Methodology

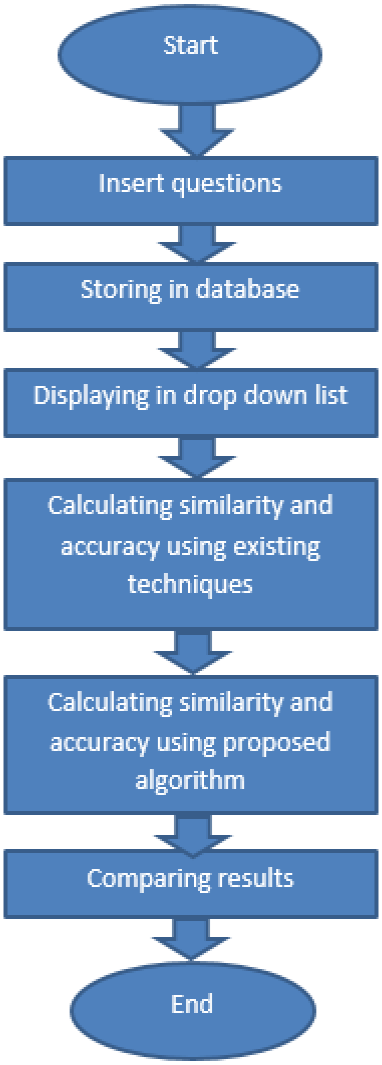

Start
Insert questions
Storing in database
Displaying in drop down list
Calculating similarity and accuracy using existing techniques
Calculating similarity and accuracy using proposed algorithm
Comparing results
End

Figure 3.3 Research Methodology

In my research work Naïve-Based, Rabin-Karp and Boyer-Moore algorithms are used that are based upon syntactic similarity. In this thesis, the proposed technique is performed following steps. These steps defined the proposed algorithm step by step and also explained the existing algorithms that are used for comparison with proposed method. Research methodology steps are explained all the study work that are helpful to achieve the desired result.

The steps of research methodology are described as follows:-

1. **Insert Questions: -** In this step of research methodology the input questions are the questions that the users enter into the designed GUI for the purpose of comparison and similarity.
2. **Storing in Database: -** In this step the questions users enter into the text boxes of the GUI are stored in this database using MICROSOFT ACCESS database. The questions are stored into two different tables designed for them respectively.
3. **Displaying Questions in Drop Down List:-** In this step the question that are stored in the database are retrieved from the database using JDBC and ODBC drivers and are displayed in the drop down boxes in the GUI. The feature helps the users to match any two questions that are present into the database and any newly entered question with any question present in the database
4. **Calculating Similarity and Accuracy using existing Techniques: -** In this steps the questions that the user enters in the GUI are passed to the existing algorithms like Naïve- Based, Rabin-Karp and Boyar-Moore algorithm and similarity and accuracy is calculated using these algorithms. These three algorithms are explained here:-

   - **Naive-Based Algorithm**

The naïve based approach for string matching is a very basic approach. Naïve based is easy to understand and implement, but in some of the cases, the naïve based algorithm works too slow. It will take the worst case complexity of iterations (n*m) if the length of text is "m" and length of the pattern is "n" for completing the task. The idea behind the naïve based string matching is just to compare each character of a text T [s...s + m-1] and the pattern P [0…m-1]. It performs various shifts and returns all the shifts which are valid. The figure shows the example of naïve based comparison of strings. [45]

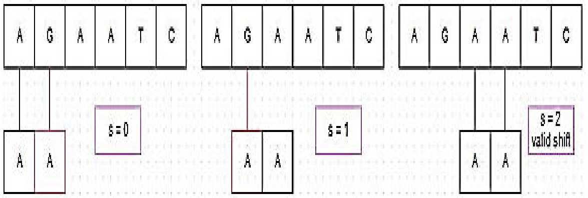


Figure 3.6 Naïve Based comparison string

The implementation of Naïve-Based string matching is:

```
NaiveMethod (string1, string2)
{
n = length (string1);
m = length (string2);
Limit = n-m;
j = 0, k = 0;
Numofshifts [];
For(i = 0; i <= limit; i++)
{
j = 0;
k = i;
For (j = 0; j <= m && str1 [k] == STR[j]; j++)
K++;
If (j >= M)
Add i to arrayOfValidShift;
}
Return arrayOfValidShift;
}
```

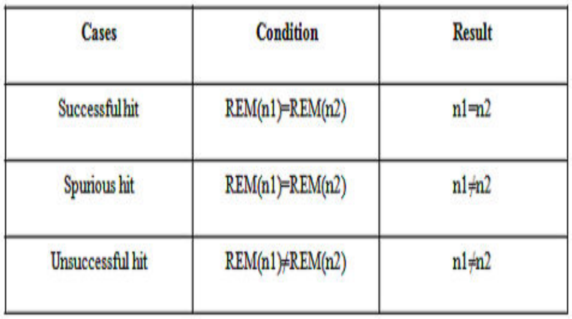
- **Rabin-Karp Algorithm**

Rabin – Karp is a string matching variant algorithm in which hashing is used for string search. For finding any of the set of patterns in the given string it uses hashing technique. For a string of length "n", for a set of patterns length of "m", the average time complexity for the algorithm is $O(N+M)$ with a space complexity of O(P). The best case running complexity of this algorithm is also same as that of the average case whereas, the worst case running complexity can be O(NM). By using Rabin-Karp, For a pattern P[0…m-1] we will calculate a hash function $h(x)$. By using the obtained hash value we will find the match for each substring of length m-1 of the string. Like the KMP, Rabin-Karp also uses the pre-processing before the pattern searching process. By using that pre-processing operation it obtains the hash value which is used to compare the string and pattern. The complexity for the pre-processing stage is can be calculated as O(M). So, the time complexity for running the program can be shown as *O(M* x *(N-M+1))*. Here, we consider the following notations while calculating the similarity for search using hash value: [46]

h(p): it denotes the obtained hash value of the pattern P.

$h(t_s)$: it denotes the hash value of substring [s…s+M-1]

The computation of *$h(t_{s+1})$* can be shown as : (h(text[j], text[j+M], hText, d). There are three cases after division they are:



Cases	Condition	Result
Successful hit	REM(n1)=REM(n2)	n1=n2
Spurious hit	REM(n1)=REM(n2)	n1≠n2
Unsuccessful hit	REM(n1)≠REM(n2)	n1≠n2

Figure 3.7 Three cases of Rabin Karp algorithm

- **Boyer Moore Algorithm**

In Boyer-Moore algorithm the string matching is performed from right to left. By using this, we can skip the number of characters than the previous algorithms. For example, if the first character matched of the text is not contained in the pattern P [0..m- 1], we can skip m characters immediately. As the KMP algorithm, this algorithm pre-processes the pattern to obtain a table which contains information to skip characters for each character of the pattern. Boyer-Moore algorithm also maintains a table of alphabets which contains as many as characters in the string. The advantage of BM algorithm over KMP and the naive one, we only need four attempts to and the valid shift. In this case, the time complexity of the BM algorithm is sub linear: O(N/M).

Our work includes three algorithms of string matching that match the string and our proposed algorithms also gives the similarity and accuracy rate and compared with these three algorithms and produce better result [47].These all are the existing techniques that used by other authors in the papers so to study these algorithms and got idea these all that further to use for proposed system.

Finally to have an implementation of the proposed algorithms is involved that is based upon the string matching. And used the syntactic similarity. There are the step that involved in the implementation of the algorithm: Firstly we took a two question one in first text box and other is second text box. Then we have a five push button and two drop down boxes. Write the two questions in the text boxes and click on the database then database stored the questions and add to the drop down box that we used. After then click on the first algorithms that show the similarity and accuracy rate.

Calculate the Similarity as:

$$S = \frac{(\boldsymbol{Common\ words*2})}{(\boldsymbol{total\ word\ in\ string1+total\ words\ in\ string2})}*100$$

Then second algorithm and third that show the similarity and accuracy rate. Accuracy is defined to compare each character to pattern n and string length m.Finally to click on proposed algorithm that show the similarity and accuracy rate and compared with three one the comparison result show that our proposed algorithm have better accuracy rate to others.

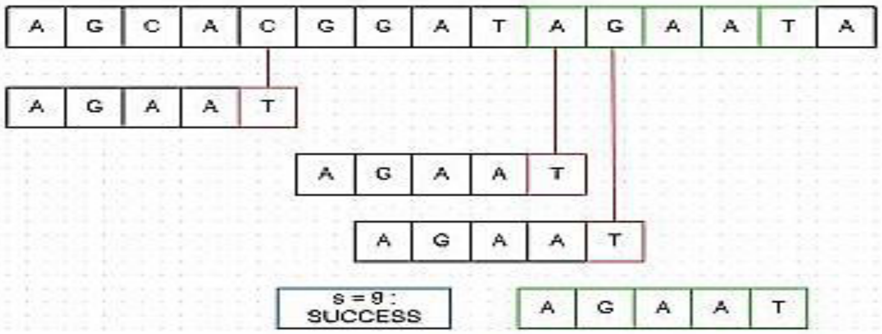


Figure 3.8 Boyer Moore Comparison string

5. Calculating Similarity and Accuracy using Proposed Algorithm: - In this step the questions are now passed to the algorithm we have designed in order to find the similarity and accuracy.

- **Proposed Algorithm**

For the proposed fast algorithm, we will calculate the hash value for the pattern first. Later, the given string will be divided into multiple small strings which will be considered as patterns for which the hash values will be generated individually. Now, we will compare the hash of pattern with the hash values of individual string patterns of the string. If the hash value matches, then we will find the similarity index between the patterns. The algorithm can be shown as:

```
Procedure FastAlgoSet (set of string subs[1..n], pattern string ,m):
  set hsubs :=emptySet
  for each sub in subs
calculate hash(pattern,m);
  insert hash(sub[1..n]) into hsubs
hs: = hash(s[1..m])
for i from 1 to n-m+1
   if hs∈hsubs and s[i..i+m-1] ∈ subs
  return i
  hs := hash(s[i+1..i+m])
  return not found
```

Our proposed fast algorithm uses multi pattern search along with Boyer-Moore and KMP hashing technique. This provides the worst case time complexity of O(n/m) which is very less when compared to other algorithms. Our algorithm is implemented in MATLAB environment with MICROSOFT ACCESS as a backend.

6. Comparing Results: - In this step we take results from all the existing algorithms and our proposed algorithm and we compare them in order to find out which algorithm is more accurate in calculating similarity and accuracy our proposed algorithm or the existing ones. The diagram for calculating similarity and accuracy in our proposed system is described below in which questions are passed to the algorithms after they are stored in database and displayed in drop down lists.

The diagram below describes the the process for calculating similarity and accuracy as per the algorithm we have designed. In this system we first take two questions from the users and store them into the database. After storing questions we display the questions and pass them to the all four algorithms naïve based, rabin karp, booye mooore and our proposed algorithm in order to measure similarity and accuracy. At last we will be having results from all algorithms and the we compare the results and check out which is the best algorithm to use for string matching.

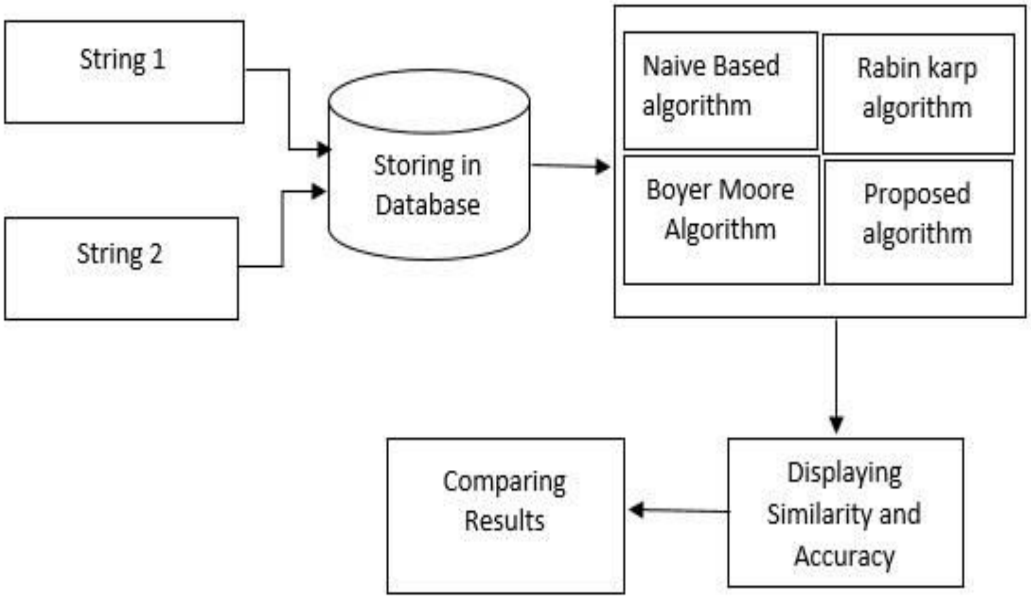


Figure 3.5 Proposed System

CHAPTER 4

IMPLEMENTATION

This chapter discusses the implementation of the proposed model. In the beginning tool and drivers used for implementing the proposed model that is MATLAB tool, JDBC, ODBC drivers and Microsoft access database is discussed. In the end graphical user interface is discussed with the functioning of every component along with the result window.

4.1 Matlab

MATLAB is a short name for matrix laboratory. In the current scenario for the implementation of proposed methodology MATLAB version (R2013a) is used. MATLAB provides an environment for development of algorithms, analysis of data, visualization, and numerical computation. It performs many computational intensive tasks with considerable high speed. It provides a high level technical computing language and interactive programming environment. MATLAB is used in the areas like signal and image processing, communication, control design, test and measurement, financial modeling and analysis, computational biology etc.

In MATLAB, variables are present in a “workspace” that correlates variable names and their values. A global workspace has defined global variables. MATLAB provides for two types of reusable code units i.e. scripts and functions. Scripts take no particular input or parameters, operating directly on the caller’s workspace. The caller can either be a function or the global workspace. On the other hand, functions have several input/ output parameters. These parameters remain bound to the function’s workspace. A symbol unbound in a function is still evaluated but the global workspace. MATLAB is the tool of choice for high-productivity research, development, and analysis. It has a rich toolbox which is the collection of various functions. The above stated features are the main reason for opting MATLAB. There are some driver and MICROSOFT ACCESS that is used in the work.

4.2 Drivers & Microsoft Access

Java database connectivity is the very useful application of programming and it is a used in the programming interface like java. That is defined that how a client may access the database. JDBC is a client side adapter not a server side that is install by a client side and to convert the request from a java program protocol that to DBMS understand it. JDBC is allow multiple of implementation and same application is used, JDBC Connection support for crating and executing the statement.it used for update statement such as SQL, to all the command that simply used to create data, insert the data, delete the data, update and also for statement of query and select the JDBC represent the statement using the classes such as statement, prepare statement and callablestatement. There are the four types of JDBC in which type1 call native code of locality available ODBC driver.type2 that call the data base native vendor from the client side and java driver is talk to server side and then talk to data base type4 pure java that use data base native protocol.

4.2.1 ODBC and JDBC

Open Data Base connectivity is an standard application programming interface (API) for used to accessing the Data Base Management System (DBMS) the aimed of designer ODBC is to make the system independent of data base system and operating system. With the few changes of a data base and application written using the ODBC can be written in both side of client and server and provide for other platform only a few changes of data access code. ODBC is depend upon driver model where the driver encapsulate the logic and need to convert multiple of commands and function into specify calls by the system. ODBC is universal it is very common and simple used by and available for most database and platform.

4.2.2 Microsoft Access

Microsoft Access is a database management system (DBMS) given from Microsoft that mixes the relational Ms-Jet Database Engine with a visual user based interface and software-developing tools. It is a member of the Ms-Office stack of applications, included at the Professional level and higher editions of it or sold independently. Ms-Access stores information in its own format depending on the Access Jet Database. It also imports or links directly to a data stored in other application and database. A Software developer, data architect and power user can also use Ms-Access to develop applications soft- ware. Like other Ms-Office applications, Access is given with graphical Basic for Application

(VBA), an object-oriented language that can refer varieties of objects which include DAO (Data Access Objects), ActiveX Data Objects, and many other ActiveX components. Graphical object used in a form and a report expose the method and property in the VBA programming environment, and a VBA code module may call operating system operation.

4.3 Graphical User Interface

GUI is an interface that is user friendly. And also to very easy to use simply to have the click on the given buttons and everything is clear on the GUI. An interface user friendly interface is created, so that it can be easily used by one.it performs all the functioning by clicking on the buttons. In the GUI that created for implementation having used the two text boxes and two drop down boxes and five push buttons. First text box have a one question that user will be write and second text box have write a second question by user defined and drop down box have a list of the question that entered the text box. There are five push of button and each of have a different role. First push button for a Naïve Based algorithm, second push button for a Rabin Karp algorithm, third push button for the Boyer Moore algorithm, fourth push button for an our proposed algorithm and five push button for a data base in which when we enter the question in the first and second box then click on the database and question stored in the data base. First four push button for find out the similarity between two question on the base of syntactic approach and also find the accurate rate between them. The GUI consists of 2 text boxes with 2 drop down boxes and 5 buttons for adding strings to the database and for comparison of different algorithms.

We have taken five push buttons in our system. In first push button we have set naïve based, in second we have set Rabin-Karp. In third we have set Boyer Moore, in fourth we have set proposed and in last push button we are storing the question into the database. We have also taken two text boxes for two question that we will take input from the user and store them in database. We also have took two drop down lists in which we are displaying the questions after storing them into the database. When we calculate similarity we are displaying the results in the dialogue boxes which will show both accuracy and similarity.

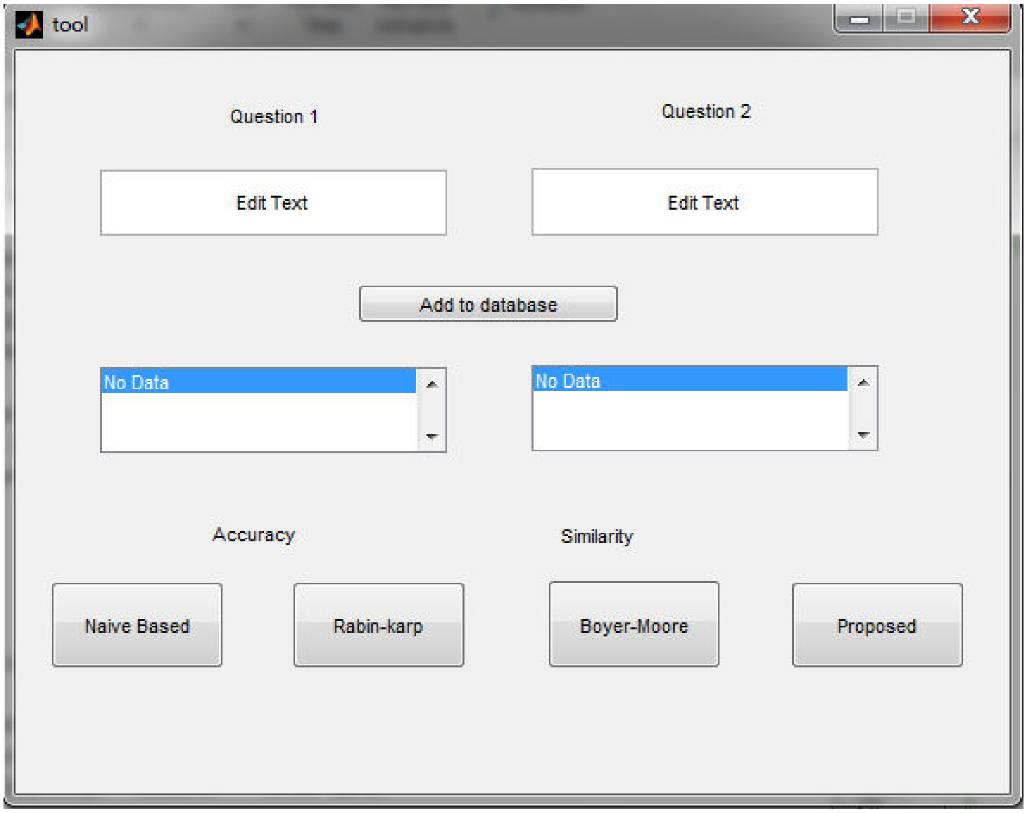


Figure 4.1 GUI

CHAPTER 5

RESULTS AND DISSCUSSION

The chapter present the result and discussions. In first section discuss the result, comparison with other methods and brief about the create database, JDBC, ODBC, MICROSOFT ACCESS and results.

5.1 Results

Our result is based upon the string matching algorithms and include the three algorithms of matching namely Naïve-Based, Rabin-Karp and Boyer-Moore. These three algorithm to produce the similarity and accuracy rate when compared with the two questions. Accuracy rate of these three algorithms is change each algorithm gives the different result when compare the two question. Our proposed method also find out the similarity and accuracy rate. The result is compare with these three algorithms and find the output. The output of proposed algorithms is little bit more accurate.

Comparison study give us that our technique is little bit good to find the similarity and remove the duplicity in between the two questions. When compared with naïve based and Rabin-Karp this algorithm provides a better speed of string search. Complexity of these algorithms in base of time complexity and run time complexity. Naïve-Based will take the worst case complexity of iterations (n*m). Rabin-Karp the average complexity of time for the algorithm is *O* (*N+M*) with a complexity of space O (P). BM algorithm is sub linear: O (N/M), KMP complexity for running the program can be shown as *O (M* x *(N-M+1))*. This provides the complexity to worst case time c O(n/m) which is very less when compared to other algorithms.so it clear that when proposed method is compare with existing one then it have a better result and good to remove the redundancy of data.

5.2 Similarity

Determine the similarity between the two questions and find out the similarity that how much they are similar in nature. There are two question first is what is computer? Second is what system is? Similarity is find out the given formula. This is measure the similarity between the given two question firstly see the common words between the two questions and multiply by 2 after to divide the words in which to have total words in string1 plus total words of string2.the formula is**:**

Similarity formula:

$$S = \frac{(\boldsymbol{Common\ words*2})}{(\boldsymbol{total\ word\ in\ string1+total\ words\ in\ string2})} *100$$

5.2.1 Comparison Table

There is table that show result of the proposed algorithms is compare toprevious one algorithm. Result shows that our method are better as compare to other because it increase the speed rate little bit as compare to others algorithms. Table show the comparison of other algorithms:

ALGORITHM	SIMILARITY	ACCURACY RATE
Naïve-Based (NB)	70%	60%
Rabin-Karp (RK)	75%	68%
Boyer-Moore (BM)	78%	69%
PROPOSED Algorithm	80%	73%

Table 5.1 Algorithm Comparison Table

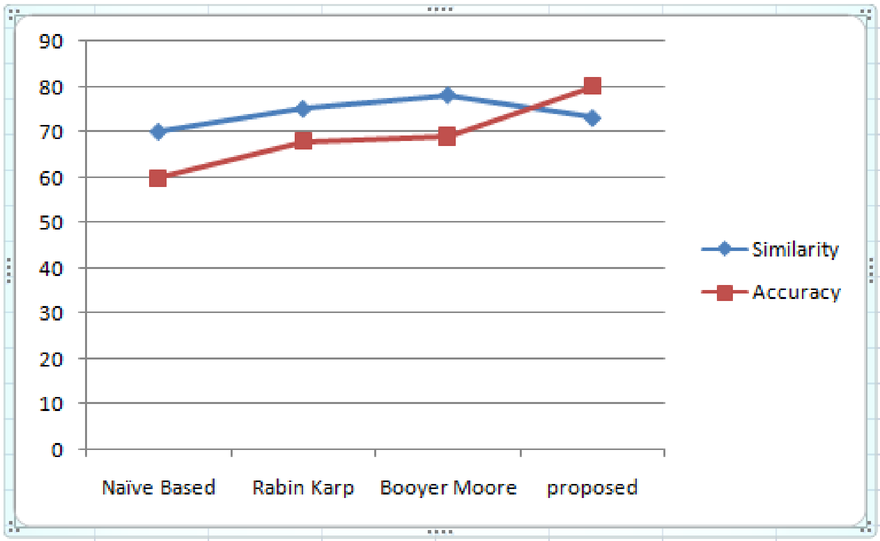
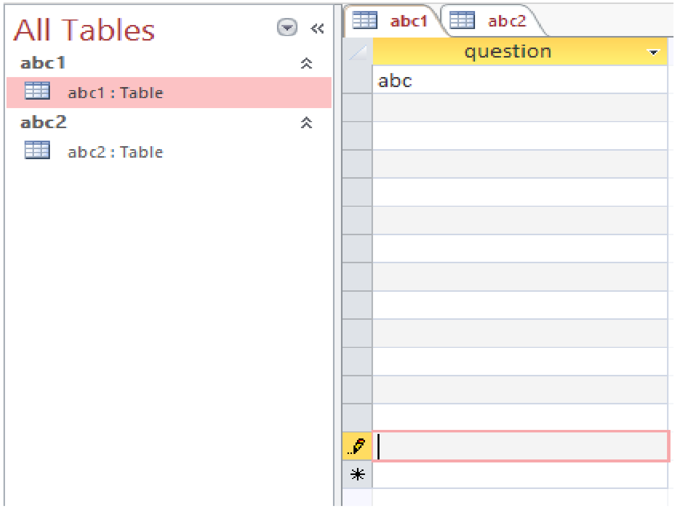


Figure 5.2 Accuracy Graph

This graph above represent the accuracy and time space complexity of designed our algorithms and previous used algorithms. This graph shows the accuracy results of our algorithm and other algorithms and this also shows time and space complexity of our algorithm is better than the other algorithms.

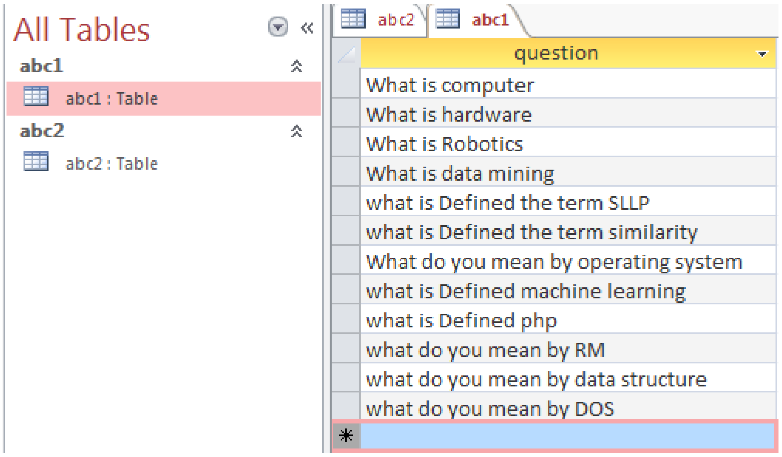
5.2.2 Steps for calculating similarity and accuracy are as follows:-

1. First we create database in MICROSOFT ACCESS using create table in Microsoft access. In the table we create entries for the data that are going to store through the interface that users are going to enter.



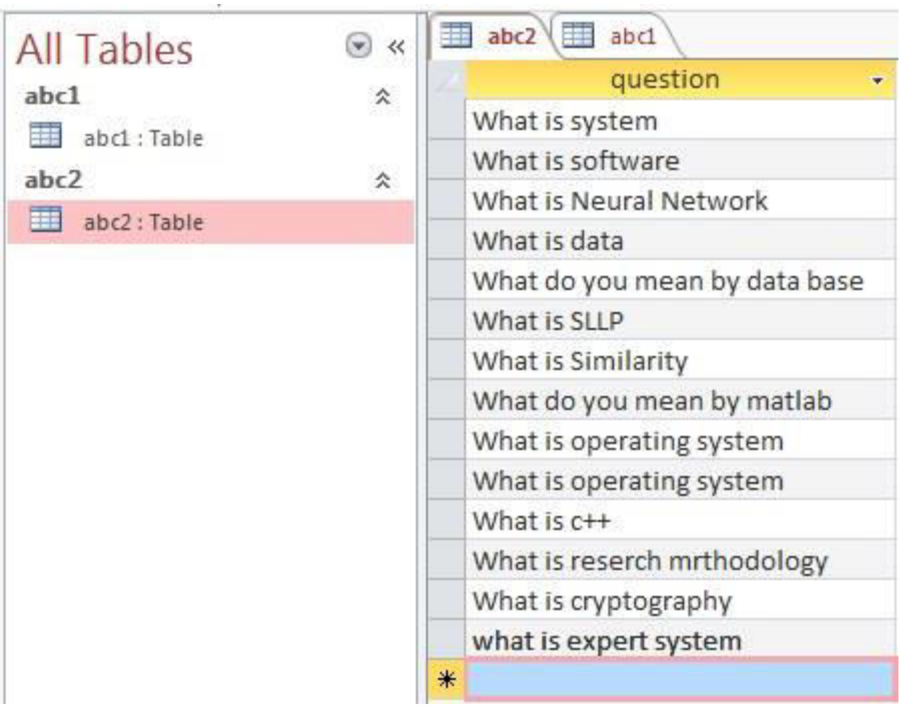
question
abc

Figure 5.3 Empty database



question
What is computer
What is hardware
What is Robotics
What is data mining
what is Defined the term SLLP
what is Defined the term similarity
What do you mean by operating system
what is Defined machine learning
what is Defined php
what do you mean by RM
what do you mean by data structure
what do you mean by DOS

Figure 5.4 Table 1 in database



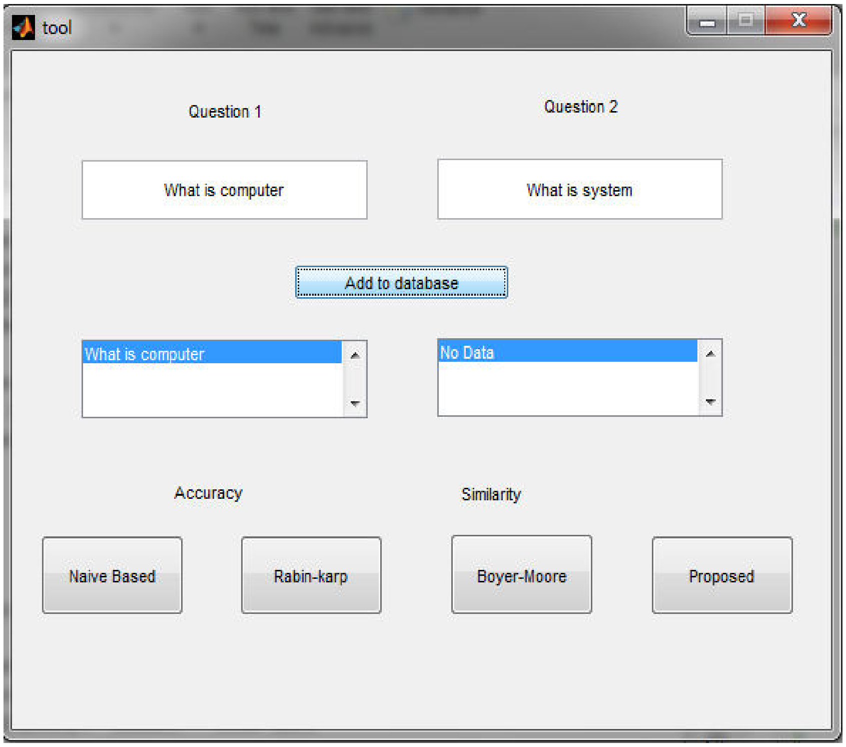
question
What is system
What is software
What is Neural Network
What is data
What do you mean by data base
What is SLLP
What is Similarity
What do you mean by matlab
What is operating system
What is operating system
What is c++
What is reserch mrthodology
What is cryptography
what is expert system
*

Figure 5.5 Table 2 in database

2. When a user enters a question into the given fields on the interface they are stored into the Microsoft access database by clicking the push button on interface named as add to database and these questions will be stored into the database respectively question one of first edit box will be stored into the table 1 and question of the edit box 2 will be stored into the table number 2 of the database.

3. In this phase we configure ODBC (open database connectivity driver) and JDBC (java database connectivity driver) for connecting the GUI with the MICROSOFT ACCESS database. When the connection is established successfully then we are able to store our data that is questions that the user enters into the text boxes into the database with the help of these Drivers and are able to retrieve them from database and displaying them into the drop down lists.

4. In this phase we made the users to enter questions into the proposed system through the GUI. These questions are first stored into the database using “ADD TO DATABASE “button and questions stored in database and drop down list in which shows all of question that we have entered. So that users will be able to any calculate similarity and accuracy of a question with any other question he/she wants.



Question 1: What is computer

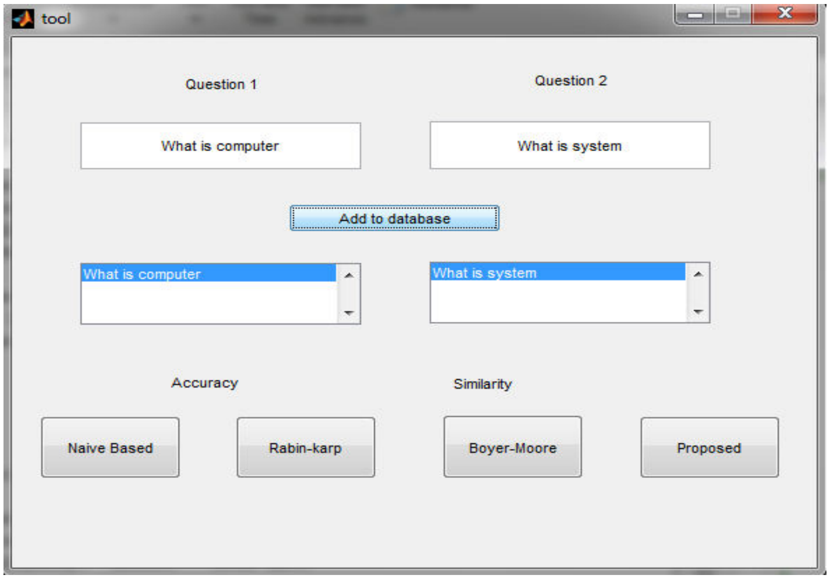
Question 2: What is system

What is computer

No Data

Figure 5.6 User Enter Questions

5. In this step we Store questions in the data base and display in the drop lists.



Question 1: What is computer

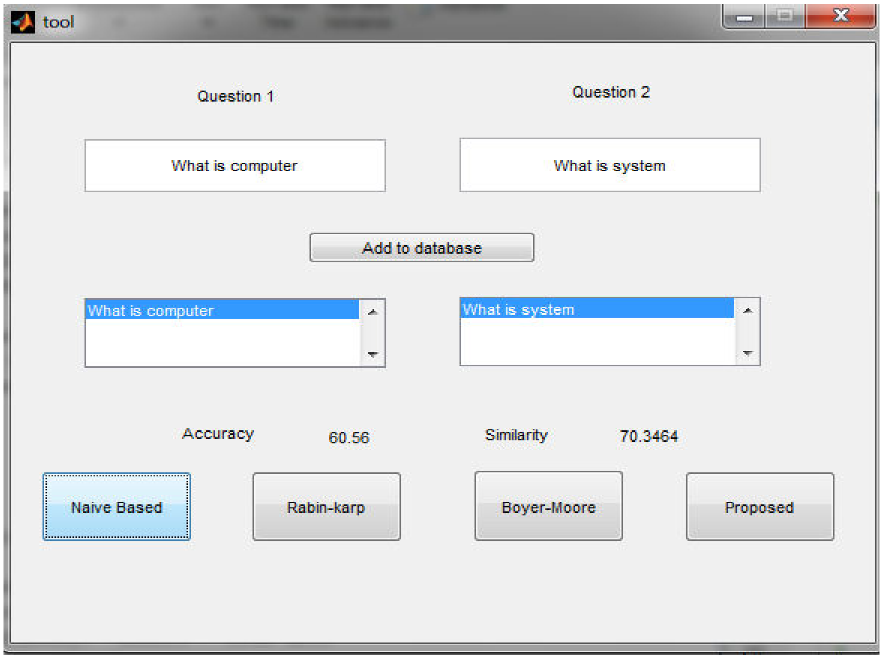
Question 2: What is system

What is computer

What is system

Figure 5.7 Stored in database and displayed in drop down list

6. In this step we calculate Similarity and Accuracy using Naïve Based algorithm.



Question 1

What is computer

Question 2

What is system

What is computer

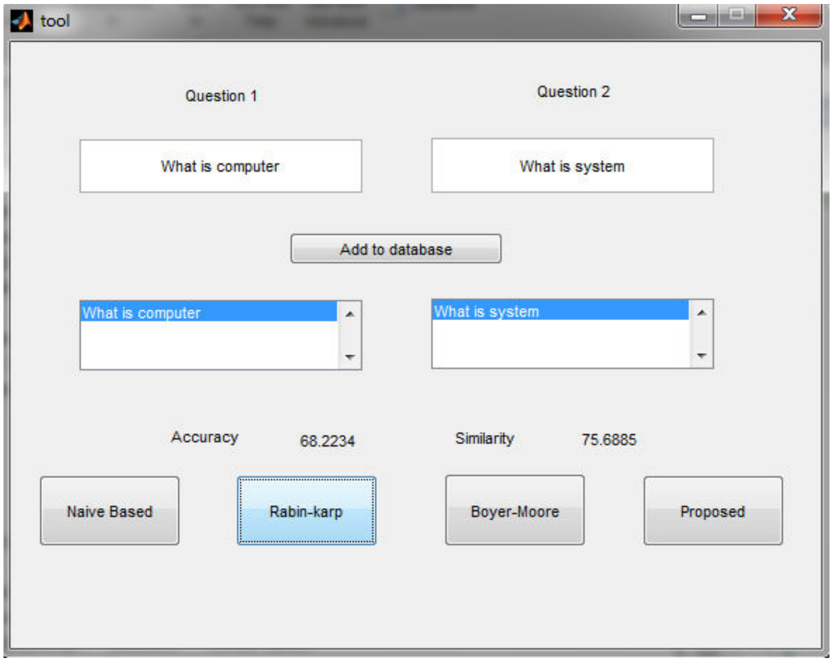
What is system

Accuracy 60.56

Similarity 70.3464

Figure 5.8 Similarity and Accuracy Calculated using NB Algorithm

7. In this step we calculate Similarity and Accuracy using Similarity and Accuracy using Rabin Karp algorithm



Question 1

What is computer

Question 2

What is system

What is computer

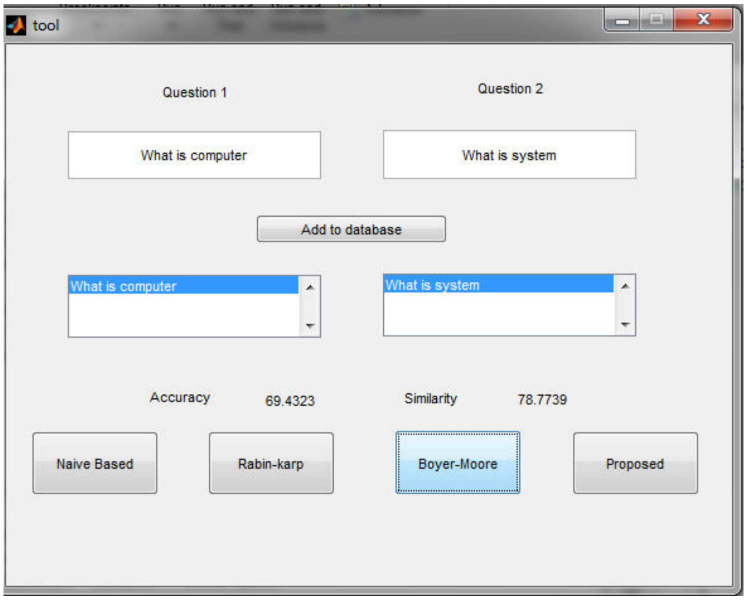
What is system

Accuracy 68.2234

Similarity 75.6885

Figure 5.9 Similarity and Accuracy Calculated using RK Algorithm

8. In this step we calculate Similarity and Accuracy using Similarity and Accuracy using Boyer-Moore Algorithm.



Question 1

What is computer

Question 2

What is system

What is computer

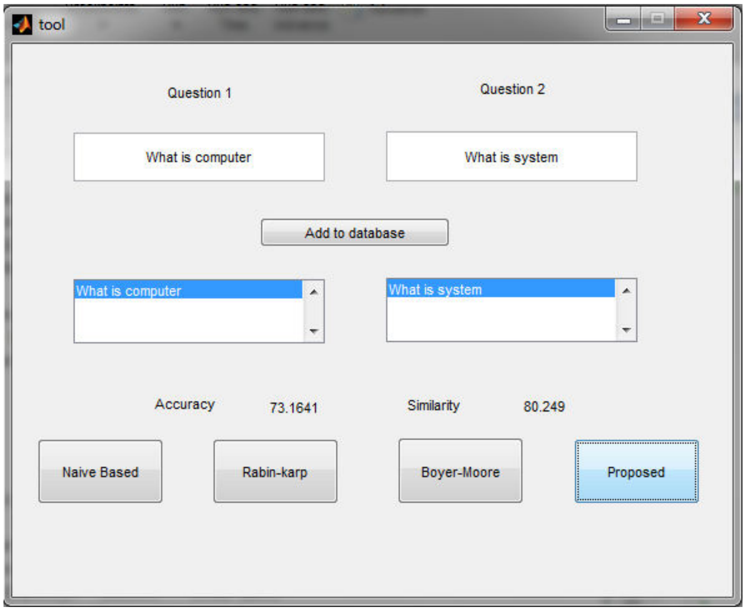
What is system

Accuracy 69.4323

Similarity 78.7739

Figure 5.10 Similarity and Accuracy Calculated using BM Algorithm

9. In this step we calculate Similarity and Accuracy using Similarity and Accuracy using Proposed Algorithm.



Question 1

What is computer

Question 2

What is system

What is computer

What is system

Accuracy 73.1641

Similarity 80.249

Figure 5.11 Similarity and Accuracy Calculated using Proposed Algorithm

5.3 Discussion

The performance of the proposed algorithms is better than the existing algorithms and speed up the accuracy rate the proposed algorithm have a multi pattern search combined with the fastest boyer-moore algorithms. Our proposed algorithm contain a little bit accuracy rate as compare to others and time complexity is also less when compare to others. The proposed method is compare with the existing method and find out the accuracy rate. The accuracy rate of the proposed system is better than other three. Result of the proposed system is compare with the three existing algorithms naïve based, Rabin Karp and Boyer Moore. The result of comparison study give the results and show them proposed method is more accurate than other method but the performance can be improve further. Also the proposed method is a better to produce a more accuracy rate so it can be improve more in the future.

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

Similarity have include various type in which similarity is to be measures. Each of have a different meaning and focus is to be find out the similarity. Syntactic similarity is the one of the best similarity in which similarity is depend upon the words of the sentences it play an good role in data mining, text mining, information retrieve etc. There are multiple of the string matching and pattern matching algorithms that match the patterns and find out the similarity or duplicity of the data. In this study to be analyse the Boyer-Moore algorithm, Naive-Based algorithm and Rabin-Karp algorithms and implement all the algorithm and compared with our proposed one.

In this work, to do string matching with the help of syntactic similarity that give a good result because it always focus on the structure of words and it does not depend upon the synonym of the word. Proposed algorithm has multi pattern search and combined with the others that gives the better result as compare with other that mean little bit accuracy rate.

6.1 Conclusion

It has been observed and analyzed from the implementation of three algorithms and our technique that the algorithms have been better result and little bit good accuracy rate.The proposed algorithm has the multi pattern search combined with the fastness of Boyer-Moore algorithm. We have shown that our proposed algorithm contains a little bit accuracy rate when compared to other algorithms. The time complexity of the algorithm is also less when compared to others. When compared with naïve based and Rabin-Karp this algorithm provides a better speed of string search.

6.2 Future Scope

In future work, researcher can propose new scheme which can reduce the included similarity of this algorithm. And the algorithm can be further developed by reducing the time complexity for calculating the hash value of an algorithm.

REFERENCES

[1] “en.wikipedia.org/wiki/Pattern matching”

[2] Koloud Al-Khamaiseh and Shadi ALShagrain,”A Survey of String Matching Algorithms, *“International journal of Engineering Research and Applications*, July 2014, Vol.4, pp.144-156.

[3] YUNTONG LIU and YANJUN,”A Sentence Semantic Similarity Calculating Method Based On Segmented Semantic Comparison, *“School of Computer and Information Engineering,* 10th February 2013.Vol.48 no.1.

[4] Harispe S., Ranwez S. Janaqi S., Montmain J,”Semantic Similarity from Natural and Ontology,*” international conference of computational linguistics*, 2014 vol. 4 pp. 34-45.

[5] Yuhua Li, David McLean, Zuhair A. Bandar, James D. O’Shea, and Keeley Crockett,” Sentence Similarity Based Semantic Nets and Corpus Statistics,”*IEEE international conference of semantics similarity,*2014 vol.4 pp.45-50.

[6] Wanpeng Song, Min Feng2 Naijie Gu1and Liu Wenyin,“Question Similarity Calculation for FAQ Answering,”*Third International Conference on Semantics, Knowledge and Grid IEEE,*2007, vol.3, pp. 1-9.

[7] MuthukrishananUmamehaswari, MuthukrishnanRamprasath, and Shanmugasundaram Hariharan, “Improved Question Answering System by semantic reformulation,”*IEEE- Fourth International Conference on Advanced Computing, ICoAC 2012 MIT*, Anna University, Chennai. December 13-15, 2012.

[8] Zhong Min Juan, “An Effective Similarity Measurement for FAQ Question Answering System, *“International Conference on Electrical and Control Engineering IEEE*, 2010.

[9] Jun sheng Zhang, Yunchuan Sun, Huilin Wang and Yanqing He,“Calculating Statistical Similarity between Sentences, *“Journal of Convergence Information Technology,* February 2011, Vol.6, no.2.

[10] Palakorn Achananuparp, Xiaohua Hu, and Shen Xiajiong,” The Evaluation of Sentence Similarity Measures*,”International Conference of Computational Linguistics*, Vol. 6, pp.25-32.

[11] Prathvi Kumari, and Ravi Shankar K*, ”*Measuring Semantic Similarity between Words using Page-Count and Pattern Clustering Methods,*“International Journal of Innovative Technology and Exploring Engineering (IJITEE),* July 2013, Vol.3, pp.2278-3075.

[12] Partha Pakray, Sivaji Bandyopadhyay and Alexander Gelbukh,” Textual entailment using lexical and syntactic similarity,”*International Journal of Artificial Intelligence & Applications (IJAIA)*, January 2011,Vol.2,no.1.

[13] Enrique Alfonseca, Marco De Boni, José-Luis Jara-Valencia, Suresh Manandhar, “A prototype Question Answering system using syntactic and semantic information for answer retrieval,”Department of Computer Science The University of York ,2014 vol.7, pp. 1-9.

[14] Kai Wang, Zhaoyan Ming and Tat-Seng Chua, “A Syntactic Tree Matching Approach to Finding Similar Questions in Community-based QA Services,”School of Computing National University of Singapore2009.

[15] Wael H. Gomaa and Aly A. Fahmy,” A Survey of Text Similarity Approaches,”*International Journal of Computer Applications,* April 2013, Vol.68, no.13, pp.0975 – 8887.

[16] Anterpreet Kaur,” A Novel Approach for Syntactic Similarity between Two Short Texts,*“INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, June2015,Vol.4,issue 06, pp. 2277-8616.

[17] Ercan Canhasi,“Measuring the sentence level similarity,” Faculty of Computer Science University of Prizren, Kosovo ISCIM 2013, pp. 35-42.

[18] Zhao Jingling, Zhang Huiyun and Cui Baojiang,” Sentence Similarity Based on Semantic Vector Model, *“Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing IEEE, 2014.*

[19] Xiao-Ying Liu and Chuan-Lun Ren,” Similarity measure based on sentence semantic structure for recognizing paraphrase and entailment, “Hindawi Publishing Corporation Mathematical Problems in Engineering,Vol. 2015, Article ID 203475, 8 pages July 2013.

[20] U.L.D.N Gunasinghe, W.a.m de silva, N.H.N.D de silva, A.S Parera and W.A.D Sashika,” Sentence Similarity Measuring by Vector Space Mode, *“International conference on Advances in ICT for emerging regions,*2014,pp. 185-189.

[21] Chi Zhang, Lei Zhang, Chong-Jun Wang, and Jun-Yuan Xie," Text Summarization Based on Sentence Selection with Semantic Representation,"*IEEE 26th International Conference on Tools with Artificial Intelligence*,2014,pp.1082-3409.

[22] Asli Celikyilmaz, Dilek Hakkani-Tur and Gokhan Tur, "LDA Based Similarity Modelling for Question Answering, "*Proceedings of the NAACL HLT 2010 Workshop on Semantic Search*,Los Angeles, California, June 2010,pp.1-9.

[23] Megha Mishra, Vishnu Kumar Mishra and Dr. H.R. Sharma,"Question Classification using Semantic, Syntactic and Lexical features,"*International Journal of Web & Semantic Technology (IJWesT)*,Vol.4, no.3, July 2013.

[24] Shashank and Shailendra Singh, "Statistical Measure to Compute the Similarity between Answers in Online Question Answering Portals," *International Journal of Computer Applications,*Vol.103, no.15, Octover 2014, pp.0975 – 8887.

[25] Wan-Yu Lin, Nanyun Peng Chun-Chao and Yen Shou-de Lin," Online Plagiarism Detection through Exploiting Lexical, Syntactic, and Semantic Information,"*Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Republic of Korea, 8-14 July 2012,pp.145-150.

[26] Hiranya Jayathilaka, Alexander Pucher, Chandra Krintz, and Rich Wolski," Using Syntactic and Semantic Similarity of Web APIS to Estimate PortingEfferot,"*International journal of service computing*, Dec 2014, Vol.4, pp. 1-14.

[27] Ludmila Cherkasova, Kava Eshghi and Charles B. Morrey III, Joseph Tucek, Alistair Veitch," Applying Syntactic Similarity Algorithms for Enterprise Information Management," KDD 09 Paris France,Vol.4,June-July 2009,pp. 625-630.

[28] Akhtar Rasool, Amrita Tiwari, Gunjan Singla, and Nilay Khare," String Matching methodologies: A Comparative Analysis,"*International journal of computer science and technology,*vol 3,2012, pp. 3394-3397.

[29] Mr. Rahul B. Diwate and Prof. Satish J. Alaspurkar," Study of Different Algorithms for Pattern Matching," *International journal of advance research in computer science and software engineering,* Vol.3,March 2013,pp.615-620.

[30] Kenji Sagae and Andrew S. Gordon,"Clustering Words by Syntactic Similarity Improves Dependency Parsing of Predicate-Argument Structures*,"Proceedings of the 11th International Conference on Parsing Technologies (IWPT),*Paris,Octover 2012*,* pp. 192–201.

[31] Su-Youn Yoon and Suma Bhat," Assessment of ESL Learners, Syntactic Competence Based on Similarity Measures,"*Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,Jeju Island, Korea,* 12–14 July 2012,pp.600-608.

[32] Diarmuid ´O S´eaghdha and Anna Korhonen, "Probabilistic models of similarity in syntactic context,"*Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing inEdinburgh, Scotland, UK*, July 27–31, 2011,pp.1047-1057.

[33] Yashar Mehdad, Alessandro Moschitti and Fabio Massimo Zanzotto," Syntactic/Semantic Structures for Textual Entailment Recognition,"*Human Language Technologies: The 2013 Annual Conference of the North American Chapter of the ACL*, *Los Angeles, California*, June 2013, pp.1020-1028.

[34] Rafael Ferreira," A New Sentence Similarity Method based on a Three-Layer Sentence Representation," *IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT),* 2014.

[35] D. Das, Schneider, D.Chen, and N.A.Smith," Probalistics frame-semantic parsing, "in Human Language Technologies,*" The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*,2010, pp.948-956.

[36] Jehad Q. Odeh," New and Efficient Recursive-based String Matching Algorithm (RSMA-FLFC),"*International Journal of Computer Applications*,Vol 86,no.15,January 2014,pp.0975 – 8887.

[37] R. Nigel Horspool, "Practice Fast Searching in String*," Journal of Software Practice and Experience,* vol.10, pp. 501-506.

[38] Jiwoon Jeon, W. Bruce Croft and Joon Ho Lee,"Finding Similar Questions in Large Question and Answer Archives,*"*Centre for Intelligent Information Retrieval, Computer Science Department University of Massachusetts, Amherst, MA 010032005.

[39] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and. L. Mercer," The mathematics of statistical machine Translation: parameter estimation," Compute. Linguist.1993, Vol.2, pp.263–311.

[40] R. D. Burke, K. J. Hammond, V. A. Kulyukin, S. L. Lytinen, N. Tomuro, and S. Schoenberg., "Question answering from Frequently Asked Question files," Experiences with the FAQ finder system. Technical report, 1997.

[41] E. Sneiders, "Automated question answering using question templates that cover the conceptual model of the database, "*In Proceedings of the 6th International Conference on Applications of Natural Language to Information Systems-Revised Papers*,2002, pp. 235–239.

[42] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal, "Bridging the lexical chasm: statistical Approaches to answer-finding,"*international conference of artificial intelligence,* 2010, Vol.1, pp. 192–199.

[43] Nimisha Singla, Deepak Garg, "String Matching Algorithms and their Applicability in various Applications,"*International Journal of Soft Computing and Engineering (IJSCE),* January 2012, Vol.1, Issue-6,pp.2231-2307.

[44] Wanli Ouyang, Stefano Mattoccia and Wai-Kuen Cham," Performance Evaluation of Full Search Equivalent Pattern Matching Algorithms," *IEEE Transaction on Pattern Analysis and Machine Intelligence*,January 2012 Vol. 34, no.1.

[45] Zeeshan Ahmed Khan and R.K Pateriya,"Multiple Pattern String Matching Methodologies," *International Journal of Scientific and Research Publications*, July 2012,Vol. 2,pp.2250-3153.

[46] Jingbo Yuan, Jisen Zheng and Shunli Ding ,"An Improved Pattern Matching Algorithm,"*Third International Symposium on Intelligent Information Technology and Security Informatics conference*, 2010,pp.599-603.

[47] http://en.wikipedia.org/wiki/String_Matching.

APPENDIX A

Publication

1. Neha Kumari, Sukhbir Kaur, “Online Assessment of Similarity between Sentences in Question Analogous System” International journal of Advanced Research in Computer Science and Software Engineering, May 2016,vol. 6, pp.828-831.
2. Neha Kumari, Sukhbir Kaur, “A Novel Approach of Syntactic Similarity of Question Analogous System” International Journal of Computer Science and Information Technology Aug 2016 Vol. 7 (4) pp. 2140-2144.