

**A Novel approach to find the Syntactic similarity between two texts.**

A Dissertation Proposal submitted

By

**Anterpreet kaur**

To

**Department of Computer Science**

In partial fulfilment of the Requirement for the

Award of the Degree

of

**Master of Technology in  
Computer Science and  
Engineering**

**Under the guidance of**

**Ms. Sukhbir kaur**

Ass. Professor,

Computer Science Engineering domain

School of Civil Engineering

**(May 2015)**

## **ABSTRACT**

Syntactic similarity is an important activity in the area of high field of text documents, data mining, natural language processing, information retrieval. Natural language processing (NLP) is the intelligent machine where its ability is to translate the text into natural language such as English and other computer language such as c++. Web mining used for task such as document clustering, community mining etc to performed on web. However to find the similarity between the two documents is the difficult task. So with increasing scope in NLP require technique for dealing with many aspects of language, in particular, syntax, semantics and paradigms.

## **ACKNOWLEDGEMENT**

I would like to take this opportunity to express my deep sense of gratitude to all who helped me directly or indirectly during thesis work.

Firstly, I would like to thank my supervisor Ms Sukhbir kaur for being great mentor best adviser I could ever have. Her advice, encouragement and critics are so innovative ideas, inspiration and cause behind the successful completion of this dissertation. I am highly obliged to all faculty members of computer science and engineering department for their support and encouragement. I would like to express my sincere appreciation and gratitude towards my friends for their encouragement, consistent support and invaluable suggestions at the time I needed the most.

I am grateful to my family for their love, support and prayers

Anterpreet kaur

Reg.no 11312700

## DECLARATION

I hereby declare that the dissertation proposal entitled, A novel approach to find the syntactic similarity between two texts submitted for the M.Tech Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: \_\_\_\_\_

**Anterpreet kaur**

**Reg no: 11312700**

## CERTIFICATE

This is to certify that Anterpreet kaur has completed M.Tech dissertation proposal titled A novel approach to find the syntactic similarity between two texts under my guidance and supervision. To the best of my knowledge, the present work is the result of her original investigation and study. No part of the dissertation proposal has ever been submitted for any other degree or diploma. The dissertation proposal is fit for the submission and the partial fulfilment of the conditions for the award of M.Tech Computer Science & Engg.

Date:

Signature of Advisor

Name: Sukhbir kaur

UID:



## TABLE OF CONTENTS

Sr. No.	Topic	Page
1.	<b>INTRODUCTION</b> .....	<b>1</b>
	1.1 Data mining used in similarity.....	2
	1.2 Text mining.....	3-4
	1.3 Definition of syntactic similarity.....	5
	1.4 Difference between syntax and semantics.....	6
	1.5 Difference between syntactic and semantic similarity.....	7-8
2.	<b>LITERATURE REVIEW</b> .....	<b>9-18</b>
3.	<b>PRESENT WORK</b>	
	3.1 Research Design.....	19-20
	3.2 Research Methodology.....	21-26
	3.3 Objective & Scope of study.....	27
4.	<b>RESULTS &amp; DISCUSSIONS</b>	
	4.1 Data set .....	28-29
	4.2 Results & Discussion.....	30-37
5.	<b>SUMMARY AND CONCLUSION</b> .....	<b>38</b>
6.	<b>REFERENCES</b> .....	<b>39-41</b>

## LIST OF FIGURES

<b>FIGURE NO .</b>	<b>TOPIC</b>	<b>PAGE</b>
1.	Example of text mining.....	3
2.	How actually a similarity between two words is calculated.....	8
3.	Present the outline of developed method.....	10
4.	Present the outline of developed method.....	12
5.	Show the steps of research design.....	20
6.	Present the outline of proposed method.....	25



## LIST OF TABLES

<b>TABLE NO .</b>	<b>TOPIC</b>	<b>PAGE</b>
1.	Find the Most Similar Word to “Healed” .....	24
2.	Show the accuracy for the each question paper.....	28
3.	Graphically representation for each question paper.....	29
4.	Show the accuracy for the ten sample of question paper.....	31
5.	Graphically representation for ten samples of question paper..	32

# CHAPTER1

## INTRODUCTION

---

Syntactic similarity is playing an important activity in the of text documents, data mining, natural language processing, information retrieval. Natural language processing (NLP) is the intelligent machine which has the ability to translate the text into natural language, natural language such as English and other computer language such as c++. However to accurate the similarity between the two texts is the difficult task. So with increasing scope in NLP require technique for dealing with many aspects of language, in particular, syntax, semantics and paradigms. In the field of data mining, syntactic similarity is exploited in application like cleansing data for mining and warehousing, to detect the duplication with in words, mining knowledge from text etc. The problem of measuring similarity between short units has become increasingly important for many tasks. Task such as:

Similarity between two documents.

Similarity between the query and product name.

Similarity between the user's query and given keywords.

Similarity between the question papers.

It's not important that the similarity can only be measured in the two texts. We can also apply the similarity in the two short texts with the help of STASIS and LSA for use in conversational agents [7]. CA is computer programs that interrelate with humans through natural language conversation. "Short texts" are basically 20-25 words long, but it's not compulsory that it accurate the grammatically sentences. The main purpose of Similarity measure is also for the classification and clustering of compositions.

## **1.1 Data mining used in similarity:**

Data mining is a process which is used to examine a big quantity of records and find the hidden data which is important for the business and other organization. Various industries have been espousing data mining to their task-serious business processes to gain more advantages and help in business development. There are some data mining applications in marketing, banking, finance, health care, insurance, transportation and medicine. The main use of Similarity in a data mining context is typically identified as a distance with dimensions representing features of the objects. A little distance demonstrating a similarity in a high degree and a large distance demonstrating similarity in a low degree. Similarity is subjective and is highly reliant on the field and function. For example if the two people are similar because of their first name and the city where they live. Suppose we considered two people similar by their height and want to know how distant apart they presently live from each other. If we want to calculated both of these in centimeters, then the distance between them are find.

## 1.2 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. The results can be analyzed to determine useful knowledge, some of which would only be established by a human reading and analyzing the data. There are more tasks which is used in Text Mining, but are not restricted to Topic removal, Concept removal, Frequency-based Analysis and many more. Some of the tasks could not be satisfied by a human, which makes Text Mining more useful and suitable tool in modern computer science.

Example of text mining:

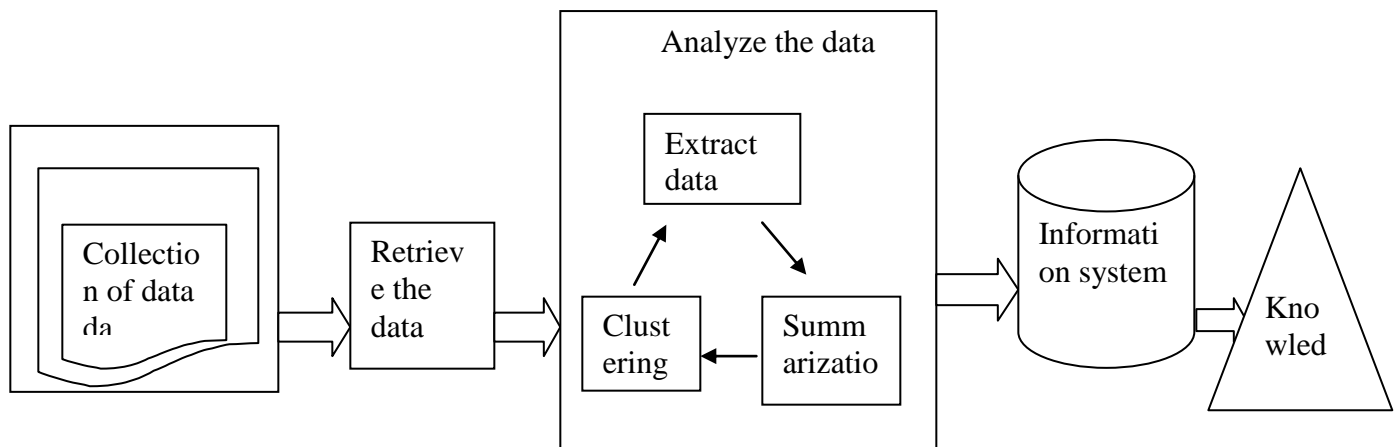


Fig 1 presents the example of text mining

As most the more information about 85% is stored as the text. In text mining, firstly collect the number of data and then the data which is useful that data will be retrieved. Then the data will be extracted, summarizing and been clustered from different resources. There are number of steps that can be used in the pre-processing.

1. **Sentence splitting:** In the sentence splitting, we can split those sentences in which the some common symbols are used such as ( ? " . & ). In any sentence when these symbols are used then the sentence is splitted. Let us assume the sentence:

The employee whose age is above than 27, he or she can only entered in the examination test.

Now this sentence can be split as:

The employee whose age is above than 27,  
he or she can only entered in the examination test.

2. **Tokenization:** It is the method of replacing the records with single classification symbols that maintain all the important information about the data without compromise its security. Tokens are recorded as interpretation in their own explanation set.
3. **POS Tagging:** POS is the Part-of-speech in which the data such as synonyms, lemma and lexemes can also be identified in this stage. POS information is stored as features of the token explanation.
4. **Stop word filtering:** This is used to filtered the words which are used more in the documents. Such as: "the, a, this, how, who, what, am". The stop words are used to ignore this type of words.
5. **POS Filtering:** POS filtering is used to read documents as input and convert the tokens for that file which is based on part of speech tag information.

### 1.3 Definition of syntactic similarity:

Similarity is a concept which has been defined in philosophical and information theory communities. Similarity means that to find the relevant meaning of a given sentence or the verb and find the accuracy between them. The main aim to find similarity is that to find repeated questions in the question paper (a.k.a automatic question paper vetting) and try to reduce these types of problem with the help of NLP and machine learning techniques. Whenever people talk about words, usually they think about the semantic similarity. Semantic similarity means that the synonyms of the given word. Although nobody can know about the syntactic similarity. But sometime it's important to learn about the syntactic similarity of words, i.e. how similar are two texts with respect to their syntactic function or role? Syntactic similarity is the concept in which the similarity will be measured by word to word. But the main issue to find the similarity is that there are some common words which are mostly used in the text such as: "THE, WHAT, A, WHY, IS, ARE" if we can't ignore these types of words then the similarity percentage will be high. So to ignore these types of words we can use the "STOPWORDS". Basically in computer stop words are the words which are cleaned out or removed the common words. Some tools specifically avoid removing these **stop words** to support phrase search. For a good performance to measure the similarity we can use the stop words.

## **1.4 Difference between syntax and semantics:**

Syntax is the figurative 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. In computer science generally, the syntax is the set of rules or the system that defines the collection of symbols that are measured to be a correctly structured paper or fragment in that language. This will be relevant both in programming languages where the file represents source code and the file will represent the information. 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 defined what the actually program work or compute. Then those semantics will one to one mapping between how the user interface wants and how actually it work.

## 1.5 Difference between the Syntactic and Semantic similarity:

In today life, some people don't have to know the basic difference between the syntactic and semantic similarity. **Semantic similarity** is the term in which the meaning or the synonyms of the given is same.

Example:

The "servant" cleans the house.

The "maid" cleans the house.

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 given words 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 studying in the college"

"I am studying in the University"

In this example here meaning will not check only the words to words check. In the field of data mining it is difficult to measures the syntactic similarity between the two documents.

One of the major problem that search engine face, in order to satisfy users information needs is "judging" means that whether a piece of (textual) information is relevant to a given information need as specified by a text query. The advantage for using semantic is in the case of frequently asked question system. FAQ is a question answer texting machine which firstly finds the question sentence from the given question's collection and then returns its correct response to the users. It may happen many times that the accurate answer will not be come as output. But the related answer will be displayed means that the meaning of the answer will be same. The work of matching questions to related questions-answer pairs has become major issue in a FAQ-system. The work of matching questions to related questions-answer pairs has become major issue in a frequently asked question system. In the past Zhong Min Juan (7) presented a method to find matching system in the question in FAQ corpus and the users text. With the combining



of statistical and semantic technique, a similarity method is generated, which firstly build semantic knowledge base, namely, co-occurrence word corpus, then used for count term frequency of question answer sentence by using statistic method.

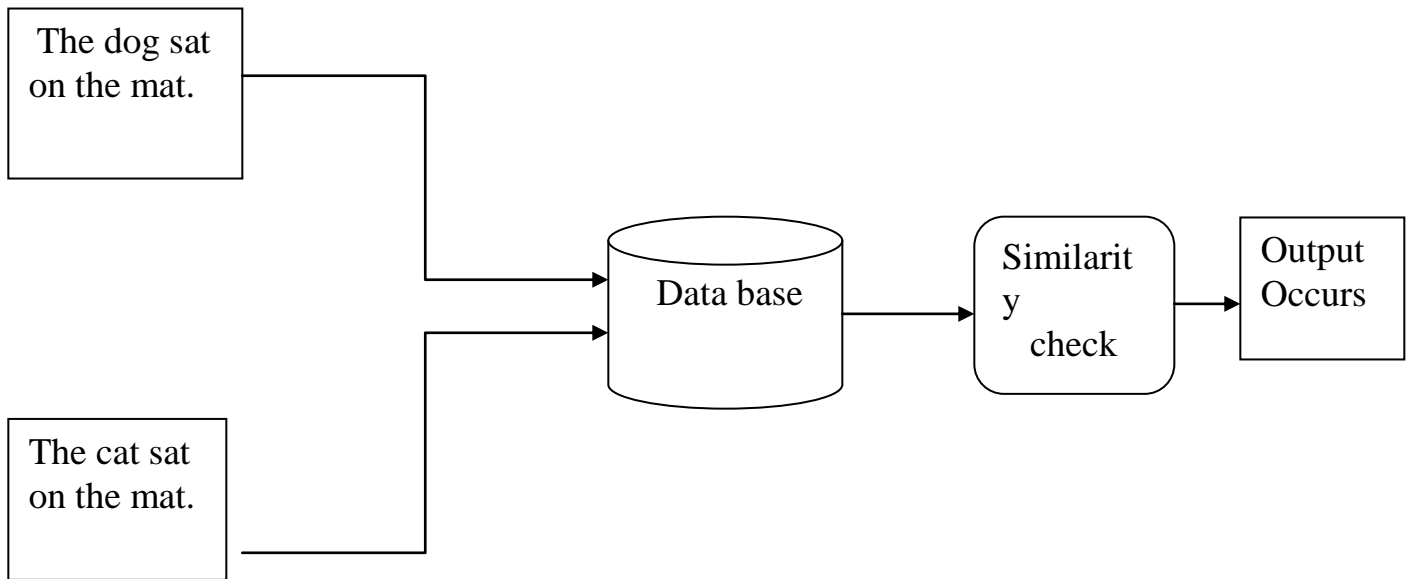


Fig 2 Figure represents how actually a similarity between two words is measured.

Here the user can enter the two texts, and then those words are stored in a database, after that the software which is developed to evaluate the similarity, they check the similarity between two texts and at last the similarity level is displayed as output.

## CHAPTER-2

# LITERATURE SURVEY

---

In [5] Manasa. Ch and V. Ramana presented an approach to measure the similarity between the words. The similarity between words is also known by using the lexical dictionary, lexical dictionary such as word net. But the main problem for using the lexical dictionary is that they are not having the recent information of words in different contexts. For example, the word “Apple”, in the field of computer science has another meaning. It is the name of the company in the hardware as well as software technology. However this word is unnoticed in the lexical dictionaries, they consider it as a fruit. Many new words are created which have their different meaning and relationships with other words, which are not introduced in the lexical dictionaries.

To overcome this disadvantage a new method is present that automatically finds the semantic similarity between words based on the page count and text snippets from web search engine like Google.

Methodology used in [5] are as:

In the case of Page count based co-occurrence, the user can send their input of two words A and B to the search engine and these words are given to page count by the search engine. The four major word co-occurrence measures are: jaccard, overlap, dice and Point-wise Mutual information (PMI) are used in proposed work to find the similarity between words.

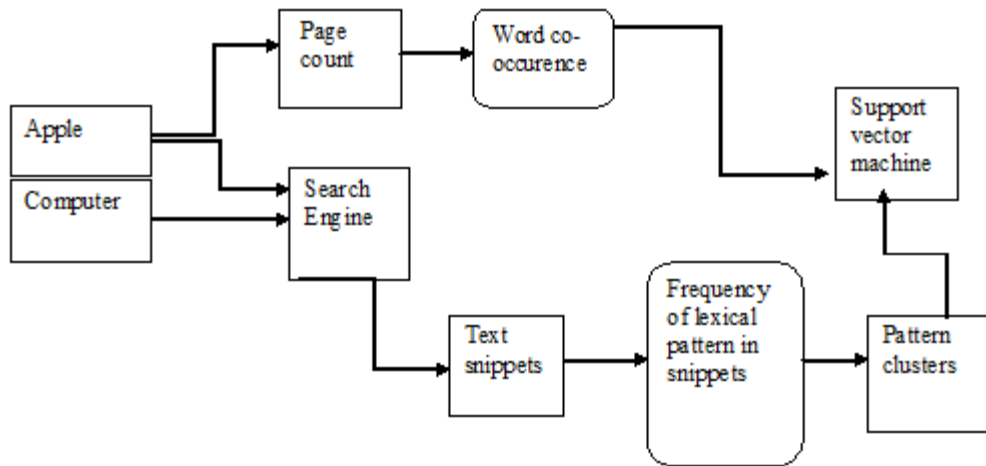


Fig 3: Figure presents the outline of the developed method.

Result: Using the algorithm like pattern clustering and pattern extraction that helps to find various relationships between words. The results are made with synonyms and non synonyms word pair that are collected from the word net synsets.

Limitation:

- Usages of page count method to measure the similarity between words are not an appropriate solution, because it does not suggest the number of times a word which has occur in each page.
- A one expression may show many times in a file and same expression in another file but the page count measure ignores this type of problem.

In [1] R. Menaha and G. Anupriya presented approach which is proposed to measure the similarity between words. To recover the disadvantages of measuring the similarity using page counts and snippets [5] this paper proposed a method to measure the similarity between words. Semantic word distance (SWD) helps to find the accuracy of similar word in each document and normalizes it over all documents. Snippets is a programming term for a small region of re-

usable source code, machine code and text. It helps to provide information regarding the local context of query term.

Methodology used which is used in[1] to find the semantic similarity are:

Pattern extraction: In this method here the user enter the words which they want, wildcard query helps to display these words like R\*M, R\*\*M, R\*\*\*M, M\*R, M\*\*R, R\*\*\*M and then those queries are searched in web search engine. The operator “\*” matches only one word not more than one in web pages.

Result: Google is used as a search engine to remove a web pages for a given word pair. The cluster score of the word pairs are measured and the Support vector machine is skilled to categorize either the given word pair as synonyms or non synonyms word pairs.

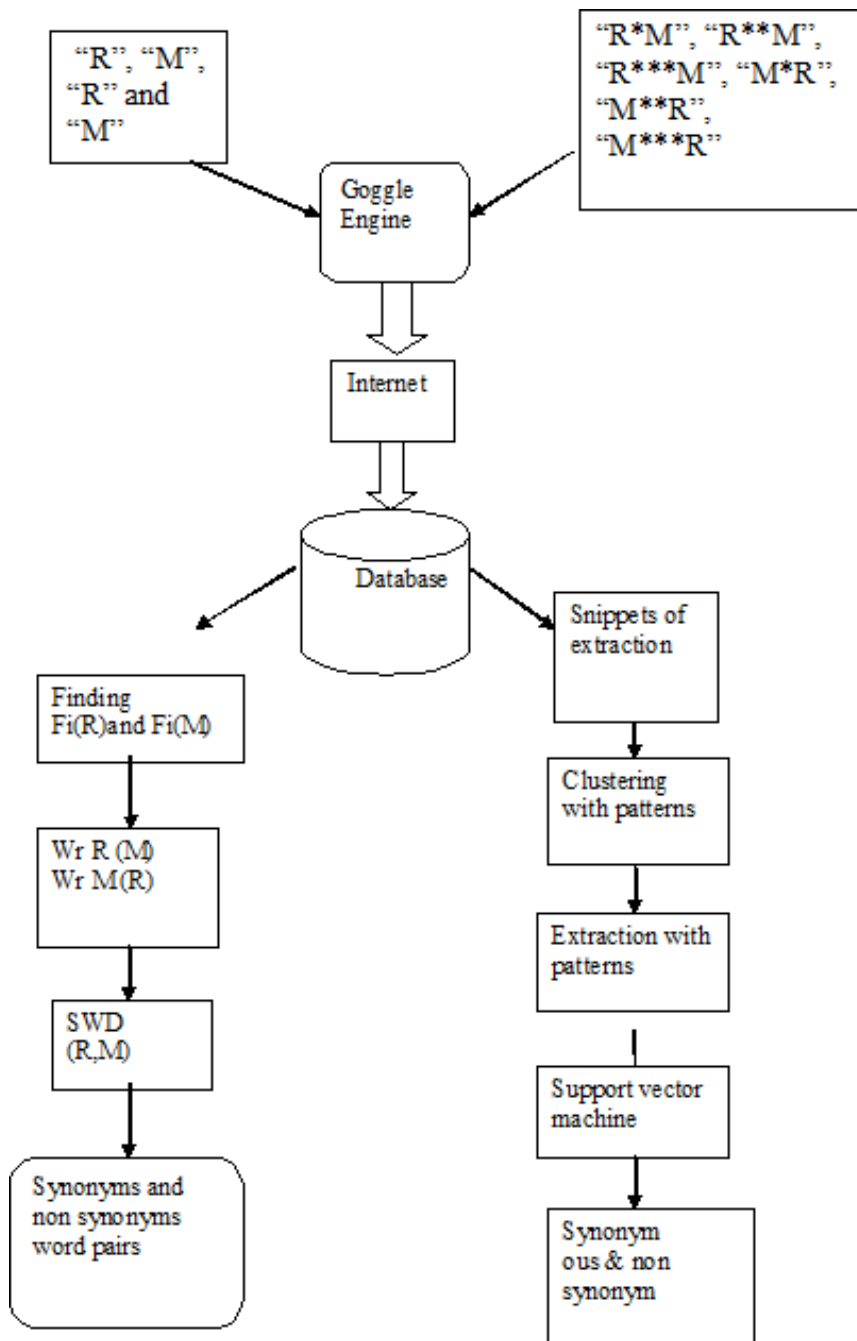


Fig 4: Figure presents the outline of the developed method.

In [6] Vasileios Hatzivassiloglou and Judith L. Klavans , attention on problem to detect whether two small paragraphs contains common information or not. When the large number of text is compared to detect the similarity then the overlap method is enough to

find similarity; but when the unit of texts are small then simple surface matching of words are used. The

main motive is to recover collection of small text units from a collected works of documents so that each text phrases within a given set describes the same action.

Methodology used:

It presents a element which support vector over a pair of textual units, where a feature is either primitive or the composite feature.

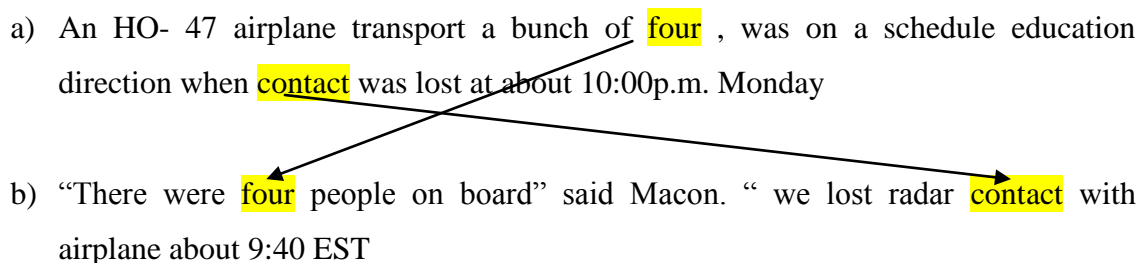
A. Primitive feature: Primitive feature is that which is based on both single lexis and simplex noun phrases. This feature compares a single word from each text document. It also consists of one characteristic. So, in the primitive feature following methods are presented which match between text units.

- Word co- occurrence: In this method it is used for sharing of a single word between text documents.
- Matching noun phrases: In this method they use a LINKIT tool to identify simplex noun phrases and equivalent those that share same head.
- Word Net synonyms: Word net helps to provide common information, placing words in set of synonyms. We match the words which have the same meaning.

B. Composite features: In addition to the primitive features, it presented a new feature which is called as composite feature. Composite features are the combination of primitive features.

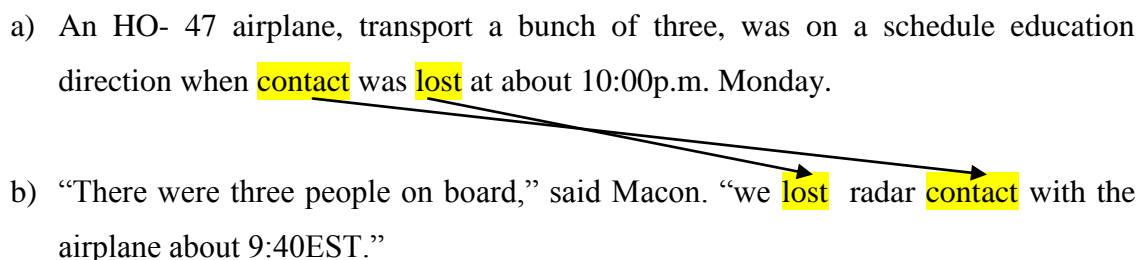
- Ordering: In the ordering technique, suppose there are two elements A & B. So these two elements have the same order in both textual units. The below example1 shows the ordering technique. In this example the word “two” in both of the texts have same order. In both the text it occurs in first order. And the word “contact” in both of the text is in the second order.

### Example 1

- a) An HO- 47 airplane transport a bunch of **four** , was on a schedule education direction when **contact** was lost at about 10:00p.m. Monday
- b) “There were **four** people on board” said Macon. “ we lost radar **contact** with airplane about 9:40 EST
- 

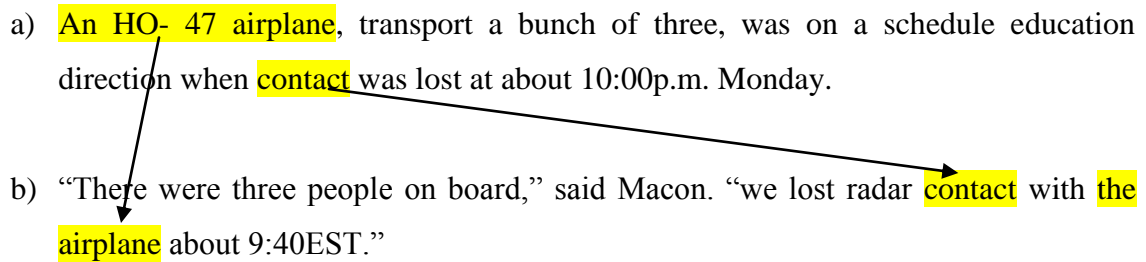
- Distance: In the distance method, distance of both texts will be checked. Example 2 shows the distance technique. The given example, in first text the word “contact and lost” has a distance one. In the second text the word “lost and contact” has a distance one. The distance of both the text has same.

### Example 2:

- a) An HO- 47 airplane, transport a bunch of three, was on a schedule education direction when **contact** was **lost** at about 10:00p.m. Monday.
- b) “There were three people on board,” said Macon. “we **lost** radar **contact** with the airplane about 9:40EST.”
- 

- Primitive: In the primitive feature here we check the words in both the text have the relative match to each other or not. Relative match means that if we change the synonyms of a given word, then the meaning of sentence is same. Example 3 shows the example of primitive.

### Example 3:

- a) An HO- 47 airplane, transport a bunch of three, was on a schedule education direction when contact was lost at about 10:00p.m. Monday.
- b) “There were three people on board,” said Macon. “we lost radar contact with the airplane about 9:40EST.”
- 

In [4] Yi Liu, Qiang Liu present a new technique to evaluate the similarity of sentences based on feature set. This method is used to define the key features in similarity definition and then combine their contribution to obtain the sentence similarity.

Methodology used in [4] are as:

Feature similarity is further divided into three parts:

- Surface feature similarity
- Structure feature similarity
- Semantic feature similarity

In the surface feature similarity, Jaccard similarity coefficient or word overlap is used. These methods doesn't work properly sometimes. For example consider these two sentences are as follows:

S1= It is the part of my life.

S2= It is the part of my life.

Here in the given sentence in both texts, all the words are same. So the two texts are exactly same.

However in some cases these methods will not work well, when the meaning of the sentence is same but position of some words are different. For example, we have two sentences as follows:

S3: Music is the part of my life.

S4: The part of my life is music.



Here the two sentence also contain the same word but at the different location. So here they proposed a method to compute a surface feature, using both word overlap and word order.

In [9] Xu Liang and Dongjiao Wang find a problem in Vector support machine based Sentence Similarity Algorithm. In generally it is based on Sentence Similarity Algorithm which mainly identify the geometric information of words in questions like arithmetic, numeric and geometric, but doesn't take the word importance in the other field and the semantic information of words. To see the disadvantages they propose further an enhancement in Sentence Similarity algorithm which is based on vector support machine, concerning impression as the basic linguistic unit of sentences.

For improvement in the VSM they firstly decided to Abstracting the concept, after that they try to give them a professional weight.

In [2008] vector based Juan M. Huerta paper present a new approach to find a semantic similarity. They decided to present a novel measure of the semantic linear equality between two sentences by means of a modified Latent Semantic Indexing (LSI) approach which is based on the indiscriminate particular Value Decomposition. Basically LSI is a process which is used by Google and other important search engine. With the help of semantic weight, they describe a new way, BLEU to include discriminatively finding the similarity. Mostly, the weights tell us how much involvement to discrimination the feature make available and is always equal or larger than zero. In the vector based approach they basically use the a) n-gram features, (b) discriminatively skilled weights in the categorization matrix vectors, and normalized amount counts for the utterance vector and (c) cosine distance between topic matrix vectors and expression vector for the development in similarity measure.

In [2005] they works on the word co-occurrence. Basically the word co-occurrence analysis is generally used in various forms of research regarding the domains of analyzing the content, text mining, construction of thesauri etc. In general, its main work is to find similarities of meaning between word pairs or similarities of meaning within word patterns. In word co-occurrence two matrixes are used for the better improvement rectangular matrix

and square matrix. At the start of their brief expedition they attention on the following assertions:

- a) Two (or more) words that be likely to occur in related linguistic contexts (i.e.to have related co-occurrence patterns), be likely to positioned nearer together in semantic space.
- b) Two (or more) words that be likely to occur in related linguistic contexts (i.e.to have related co-occurrence patterns) tend to resemble each other in meaning.

In [2006] they decided to presented a work on sentence similarity which is based on semantic nets and corpus statistics. New applications of natural language processing current a need for an significant technique to calculate the similarity between very short documents or sentences. Firstly, semantic similarity resultant is taken from a lexical knowledge base and a corpus. The lexical knowledge base models general human knowledge about words in a natural language; this knowledge is usually established across a wide range of language application areas. A main work of corpus is to replicate the concrete usage of language and words. Thus, our semantic similarity not only store common human knowledge, but it is also able to correct an application area using a corpus exact to that application. Secondly, the method which is proposed considers the collision of word order on sentence meaning. The resultant word order similarity measures the number of different words as well as the number of word pairs in a dissimilar order. Then the generally sentence similarity is then defined as a combination of semantic similarity and word order similarity. To estimate the best result for similarity algorithm, they collected a set of sentence pairs from a variety of books and from article.

Then later on [2007] KANG CHEN, XIAO-Z HONG FAN, JIE LIU present a new approach to calculate semantic similarity in Chinese question sentence. To calculate a similarity in Chinese question a new method is performed which is divided into two steps:

- 1) First step is to remove the Question Semantic representation from the question,
- 2) The second step is to calculate the question Semantic similarity based on the Question Semantic representation

They calculate the question semantic similarity on the basis of the QSR. Basically QSR is the question similarity representation. The formalized appearance of the question semantic information is called QSR i.e. Question Semantic Representation.

Some complex questions or the simple questions which includes a variety of simple queries which can be divide into some simple questions and has only one equivalent QSR but the problem is that the one QSR can be expressed by a number of dissimilar kinds of questions. For simple questions, they compare QSR on the basis of QSM matching and for complex or irregular and typical questions, such as “yas 我 ǔ 病 œ 毒 了, 乚 怎 么 杀™?” they take the related strategies which are based on meaning of some keywords in questions and hypothesize on QSR according to their possibility.

In this paper the methodology they used are very unique. Firstly they eliminate the polite words such as ǎ œ ± ǎ ǎ 杀 and so on. There is no use of these types of polite words in the QSR extraction. So they collect a these type of words to ignore them in a daily life. So the remove of the polite words are filtered in the first step. For the next step they use the segmentation, for the higher priority they use ICTCLAS system which was developed in VC, but their question system program is developed in java. Then in the third step they use the recognition for the semantic chunk. Under the order of syntax tree according to SC's composition rules, semantic chunk recognition is realized by bottom up chart analysis algorithm.

In [2] LIN LI, XIA HU, BI-YUN HU, JUN WANG presented a work on Measuring Sentence Similarity from different aspects. This paper proposes a new way to resolve sentence similarities from different resources. It may happen that information which people got can obtain from a sentence, which is objects the sentence describes, properties of those substance and behaviors of that substance. They defined a four methods[2] to find the sentence similarity.

Those four methods have their own properties. First, two assume that sentences are respectively chunked with verb as well as noun phrases. Secondly, for each word, all nouns in noun phrases are chosen as the objects particular in the sentence, all adjectives and adverbs in noun phrases as the objects properties. Then, the four similarities are considered, based on a semantic vector method.

## CHAPTER-3

### PRESENT WORK

---

#### **3.1 Research design:**

The work of research is carried out in number of stages starting from „Problem identification“ to literature review about the state of technology specific to “A novel approach to find a syntactic similarity between two short texts.” Most the time is spent in identifying and selecting the problem and in literature review. Here to find the accuracy of the repeated words in the two texts, I decided to use the net beans tool for measure the similarity.

***PHASE1: TOPIC IDENTIFICATION & SELECTION:*** In this phase firstly we read the different type of paper to identify the topic. What the topic exactly is? Is it suitable? In that topic, is there work possible or not? After the topic identified and the selection of the topic, the next phase is literature survey.

***PHASE2: LITERATURE SURVEY:*** In this phase we can read that paper properly in which we can work. This phase is very difficult task and takes too much time. After the literature survey, the next work is to find the problem.

***PHASE3: SELECTION OF PROBLEM:*** This phase is very interesting. In this phase we can select the problem in which we can further work. After Problem definition we can choose the technique or the method which we can run. It will also take too much time to select the algorithm.

***PHASE4: CODING:*** In this phase the main work is on logic, how we can implement the work. In this we can also phase a difficulty to implement the result. But we can gain more knowledge to implement the result.

If the coding is completed, then the result will be occurred.

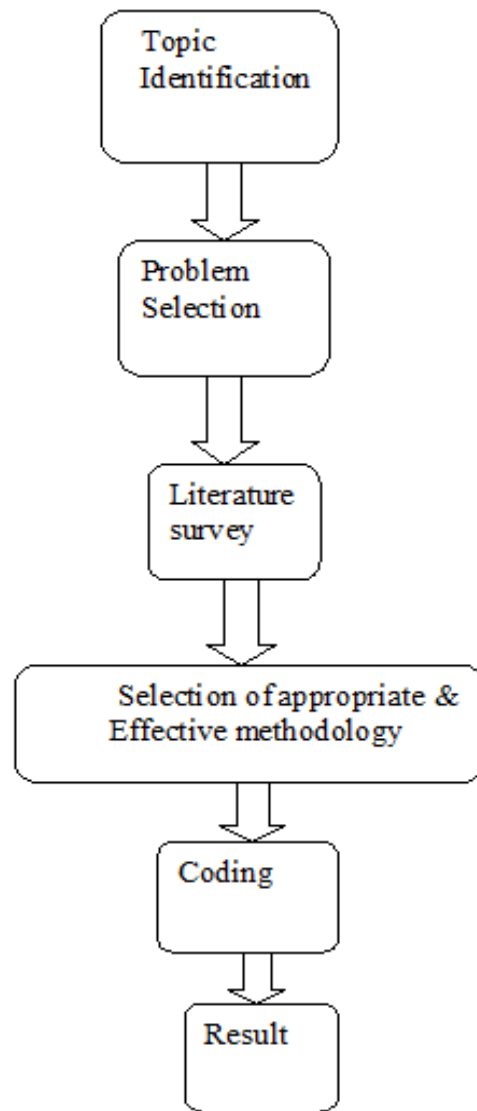


Fig 5 show the steps of research design

With the help of these steps, I will show how to done the thesis work.

### 3.2 Research Methodology:

As I assume that there is no briefly research in the syntactic similarity. So I have decided to make an improvement in the syntactic similarity between two papers. There are various algorithms which are help to find the similarity between words, Algorithms such as Edit distance, longest common substring, bi-gram algorithm and Soundex algorithm. But in these algorithms there is some problem to find syntactic similarity between words. Those approaches don't work on the some conditions. The Soundex Algorithm is a similarity algorithm, which simply defined that given two strings are similar or not. However, it would not describe any similarity between 'FRENCH' and 'REPUBLIC OF FRANCH', because they don't start with the same letters they started with different letters.

On the other hand the Edit Distance algorithm would distinguish some better result than the Soundex algorithm between the two strings, but would rate 'FRANCE' and 'FRENCH' (with a distance of 6) to be more similar than 'FRENCH' and 'REPUBLIC OF FRENCH'. And at last The Longest Common Substring would give 'FRENCH' and 'REPUBLIC OF FRENCH' having a good rating of similarity (a common substring of length 6). However, it is undesirable that according to new approach, the string 'FRENCH REPUBLIC' is equally similar to the two strings 'REPUBLIC OF FRANCE' and 'REPUBLIC OF CUBA'.

Having to seen the drawbacks of the existing algorithms, I proposed new string similarity metric that doesn't matter on the ordering method. In addition, I decided to present a new approach which not only considers the single longest common substring, but also other common substrings too. If the two strings are pronounced same then the similarity of that string are usually high, but there is difference in both of that strings, so it doesn't mean that there is not similarity between that words. Firstly I decided to check that how many adjacent characters are contained in both the strings. The purpose is that by allowing for *adjacent* characters, I take explanation not only of the characters, but also of the character ordering in the original string, since each character pair contains a little information about the original ordering.

Let me clear this statement by taking the algorithm:

- 1) Firstly take the two strings which we decided to find the similarity between them.

Example:

SYNTACTIC

SEMANTIC

- 2) Then map them both to their upper case characters and then decided to split them up into their character pairs. Example of such statement is that:

SYNTACTIC: {SY, YN, NT, TA, AC, CT, TI, IC}

SEMENTIC: {SE, EM, MA, EN, NT, TI, IC}

- 3) Then I check out which character pairs are in both strings. So in the given example, the intersection is {TI, IC}.
- 4) At last, I would like to explain the way of finding the similarity as a mathematically which reflects the size of the intersection comparative to the sizes of the given strings.

$$\text{Similarity (s1, s2)} = \frac{2 \times |\text{characters}(c1) \cap \text{characters}(c2)|}{|\text{characters}(c1)| + |\text{characters}(c2)|}$$

This new algorithm is also work in the following on the following requirements:

**A true indication of lexical similarity:** This means that two string or the words which have the small differences should be accepted as similar. It means that a considerable two string which have common characteristics should point to a gave a high level of similarity between the strings.

**It's not possible to changes of word order:** The given two strings which contain the same words in the given documents, but they are in a different order, should be renowned as being similar. On the other hand, the given two documents should be renowned as

dissimilar, if one string is just a same anagram of the characters contained in the other document.

**Language Independence** - This algorithm should also work on many different languages not easily only in English, and gave a better result to find the similarity between two documents.

But according to the new approach the similarity between two given strings s1 and s2 is twice the number of character pairs that are common to both strings is divided by the sum of the number of character pairs in the two strings. Note that the formula rates completely dissimilar strings with a similarity value of 0, since the size of the letter-pair intersection in the numerator of the fraction will be zero. On the other hand, if you compare a (non-empty) string to itself , then the similarity is 1. For our comparison of 'SYNTACTIC' and 'SEMANTIC', the metric is computed as follows:

Given that the values of the metric always lie between 0 and 1, it is also very natural to express these values as percentages. For example, the similarity between 'SYNTACTIC' and 'SEMENTIC' is 27%. From now on, I will express similarity values as percentages, rounded to the nearest whole number.

$$\begin{aligned} \text{Similarity} \quad (Syntactic, Semantic) &= \frac{2 \times |\{TI, IC\}|}{|\{SY, YN, NT, TA, AC, CT, TI, IC\}| + |\{SE, EM, MA, AN, NT, TI, IC\}|} \\ &= \frac{2 \times 2}{8 + 7} \\ &= \mathbf{0.27} \end{aligned}$$

Suppose we don't want to know how similar two strings are? But want to know which of the string is more similar to the given string. Suppose the given string is "SEALED" and check that which of the strings is most similar to given string?



### RESULT RANK

WORD	SIMILARITY
Dealed	80%
Healthy	36%
Heard	22%
Herald	20%
Hold	0%

Table 1: Find the Most Similar Word to 'Healed'.

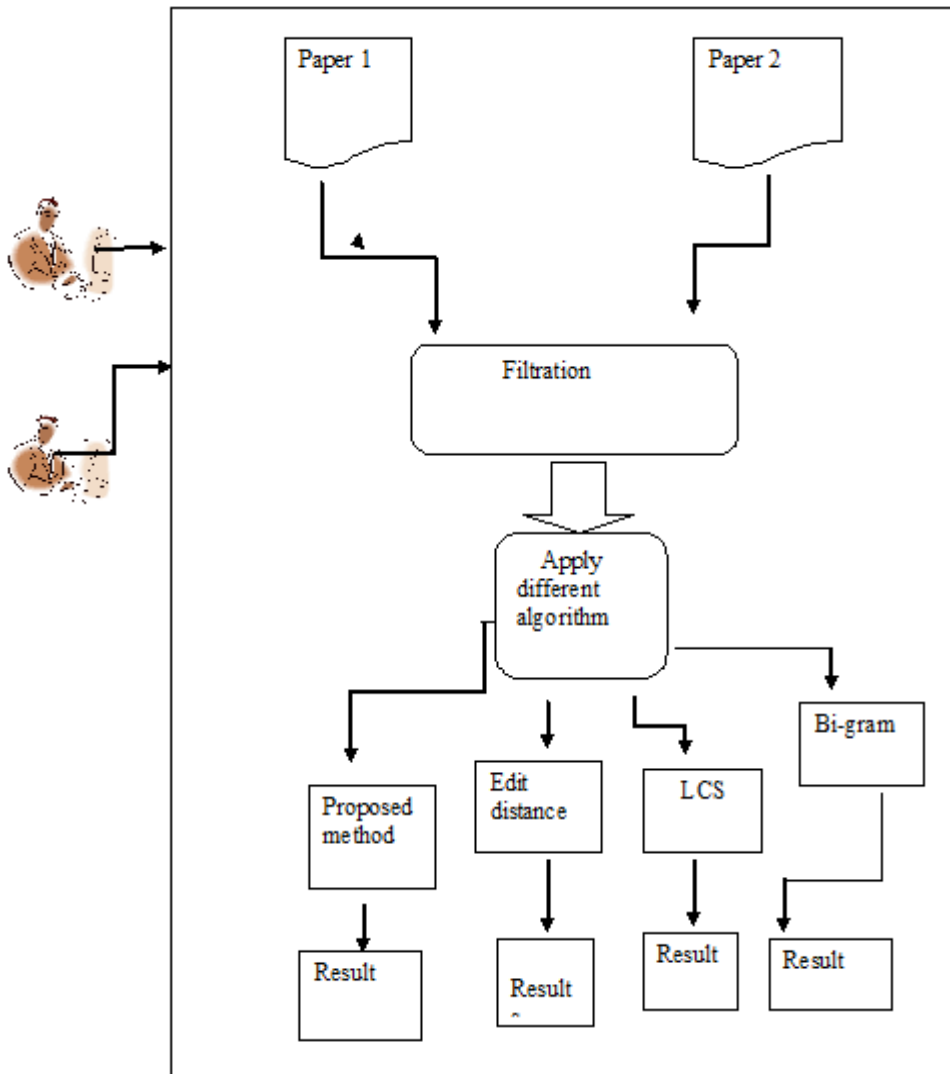


Fig 6: Proposed methodology

The very step of my research from which I will able to implement my result are as:

1. Firstly I Collect the different papers from which I will show the result which i implement.

2. Then I collect that two papers from which the users will see the similarity.
3. Next step is to filtrate the data, in this step “Stop words are use to filtrate that words which is very common like “ A, an, The, for, is” etc
4. Then applying the previous algorithm to show that either there is enhancement in the new approach or not.
5. To see the results, it’s clear that the proposed method which is used for to check the similarity between the two questions paper it’s actually work.
6. The result of the proposed methodology is better than to the previous algorithms.

### **3.3 Scope of study**

The scope of the research is to find the similar words which are present in the question paper& in the question paper the main focus is to increase the accuracy of the repeated words in two documents. Some related work has been done in past to find the similarity between two documents with the help of some techniques. It happened many times in the question paper that repeated questions occur. The purpose of this research is to detect the syntactic similarity between questions in the question papers.

### **3.4 Objective of study**

1. To study various methods of syntactic and semantic similarity.
2. To calculate syntactic similarity by using the stop words.
3. To calculate the syntactic similarity by using the Edit distance, LCS and Bi-gram.
4. Comparison of the previous algorithm by using the proposed method.
5. To calculate the similar words that can be occurred in the documents.
6. To calculate the overall similarity between the documents.

## CHAPTER-4

# RESULTS & DISCUSSIONS

---

### 4.1 Data set:

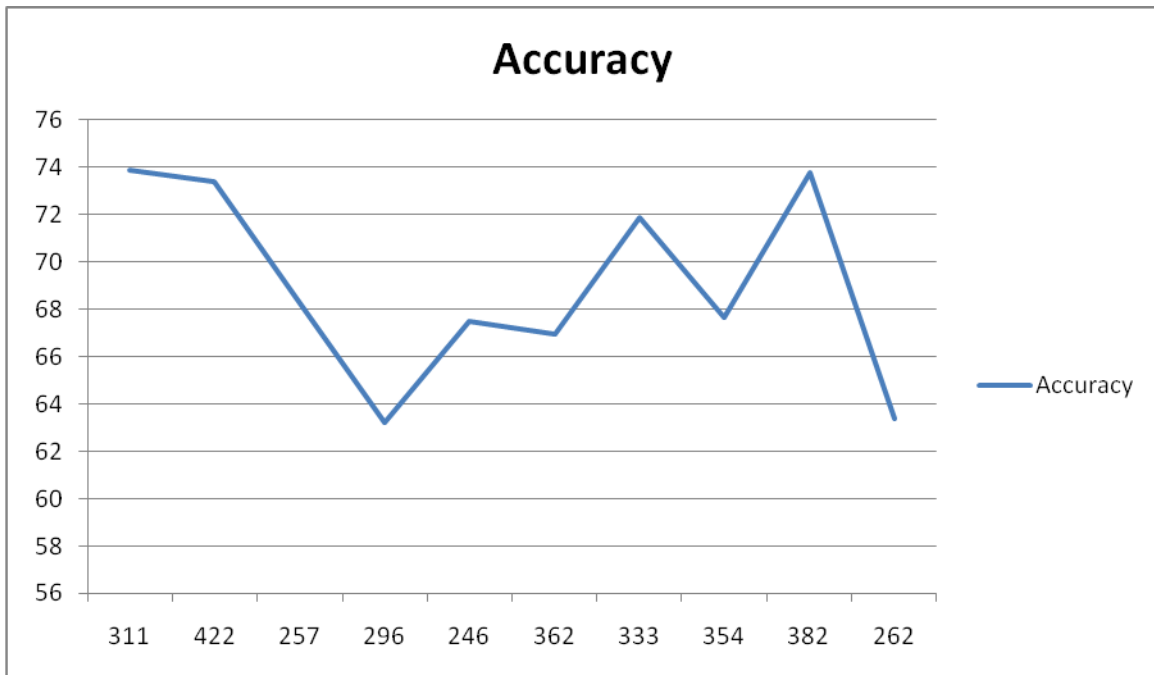
For experimental result, collect a set of questions from different resources. In the data set, length of questions is from 10 to 15, which helps to measure the similarity between the questions. Here questions are chosen with minimum and maximum length size because focus in this research is on to measure the similarity between questions. A user has randomly chosen the questions and the accuracy of similarity is stored on database. Following table1 describes the data sample which has been used. Data set has been taken in limited field which includes different kinds of questions related to computer science.

SAMPLE	NO OF WORDS	NO OF QUESTIONS		ACCURACY of find similarity= $\frac{2 \times  \text{character}(c1) \cap \text{character}(c2) }{ \text{character}(c1) + \text{character}(c2) }$
		Paper1	Paper2	
S1	311	15	15	73.86%
S2	422	15	15	73.36%
S3	257	15	15	68.22%
S4	296	15	15	63.2%
S5	246	15	15	67.5%
S6	362	15	15	66.91%
S7	333	15	15	71.86%
S8	354	15	15	67.62%
S9	382	15	15	73.74%
S10	262	15	15	63.36%

Average accuracy= 70%

**Table- 2 show the accuracy for the each question paper.**

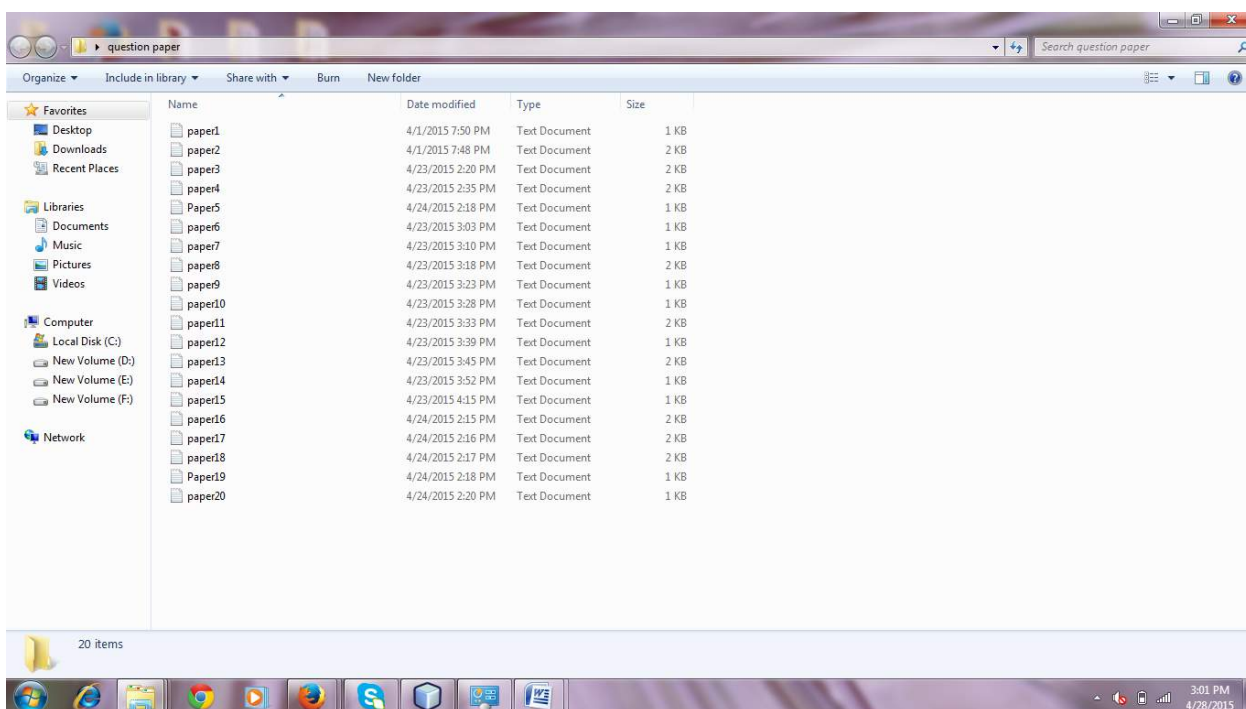
Now in the below graph, it represent the graphically representation of the data set.



The overall accuracy of my project is 69%. In the given dataset I take 10 sample of the question paper and in 1 sample two set of questions das been taken. Then with my proposed method I show the accuracy between the two set of question paper.

## 4.2 Results and discussions:

In order to check the accuracy and simplicity and to evaluate the performance of proposed system ten samples of questions sets are used which are presented in table1. Following Table2 shows the results of proposed method comparative to the different algorithms LCS, Edit distance and Bi-gram algorithms.. The evaluation results shows that the similarity based on proposed method has the better performance than the existing algorithms. For the comparative analysis the same data set is used on proposed algorithm and a table is created which is shown in table-3 and further graph is been plotted which shows the considerable amount of improvements in accuracy. But before results I show the dataset of the various question paper, so that with the use of them, I show the results of to measure the syntactic similarity between two question paper.

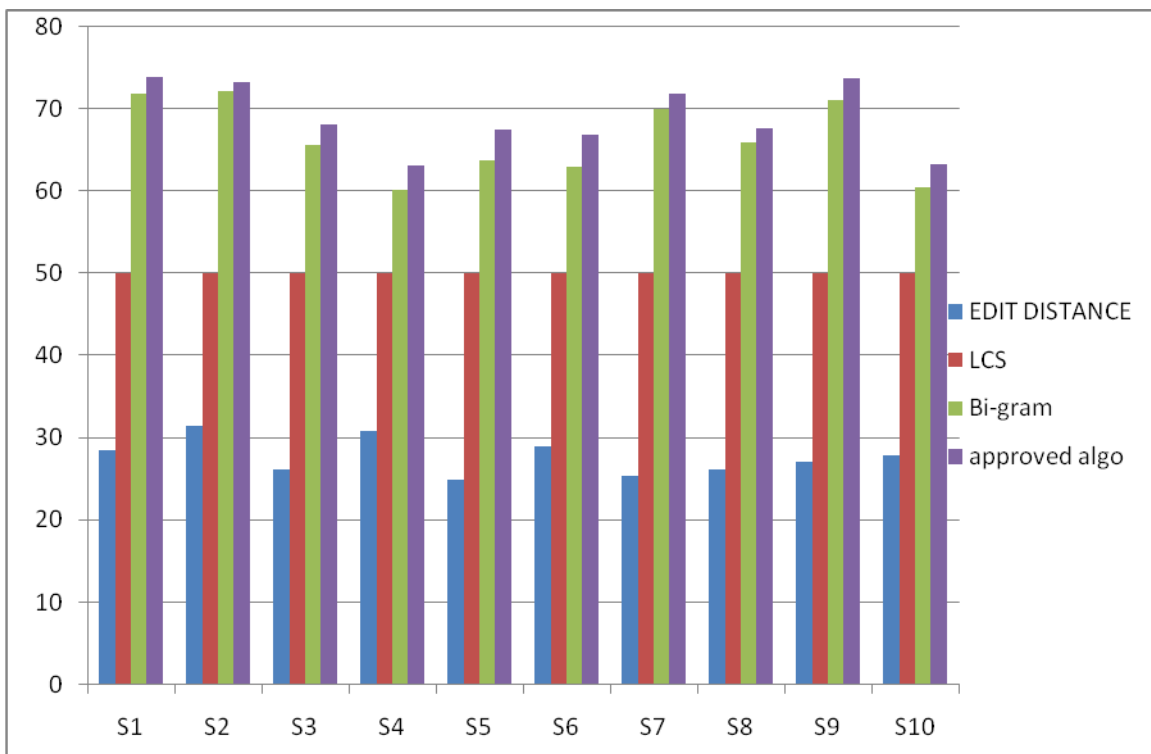


It is the data set which is used for to measure the syntactic similarity between two question papers.

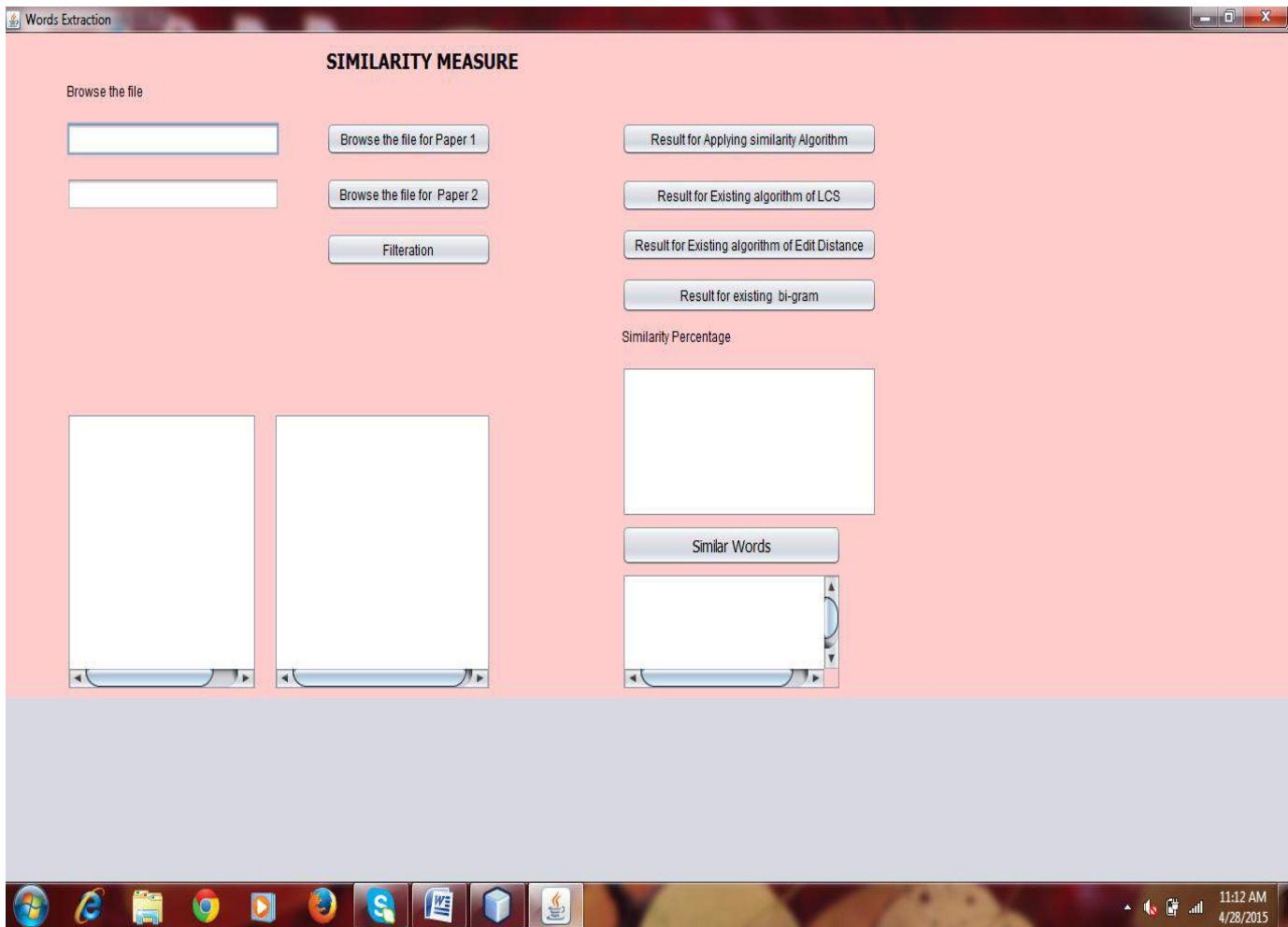
<i>Sample</i>	<i>Edit distance</i>	<i>LCS</i>	<i>Bi-gram</i>	<i>Proposed method</i>
S1	28.6 %	50.0 %	71.84 %	73.86%
S2	31.53 %	50.0 %	72.22 %	73.36%
S3	26.14 %	50.0 %	65.72 %	68.22%
S4	30.83 %	50.0 %	60.13 %	63.2%
S5	24.96 %	50.0 %	63.76 %	67.5%
S6	28.97 %	50.0%	63.02 %	66.91%
S7	25.45 %	50.0%	70.01 %	71.86%
S8	26.26 %	50.0%	65.93 %	67.62%
S9	27.08 %	50.0%	71.12 %	73.74%
S10	27.86 %	50.0%	60.48 %	63.36%

In the above table comparison has been done for the same ten samples. Table values indicate the considerable improvement in proposed method.



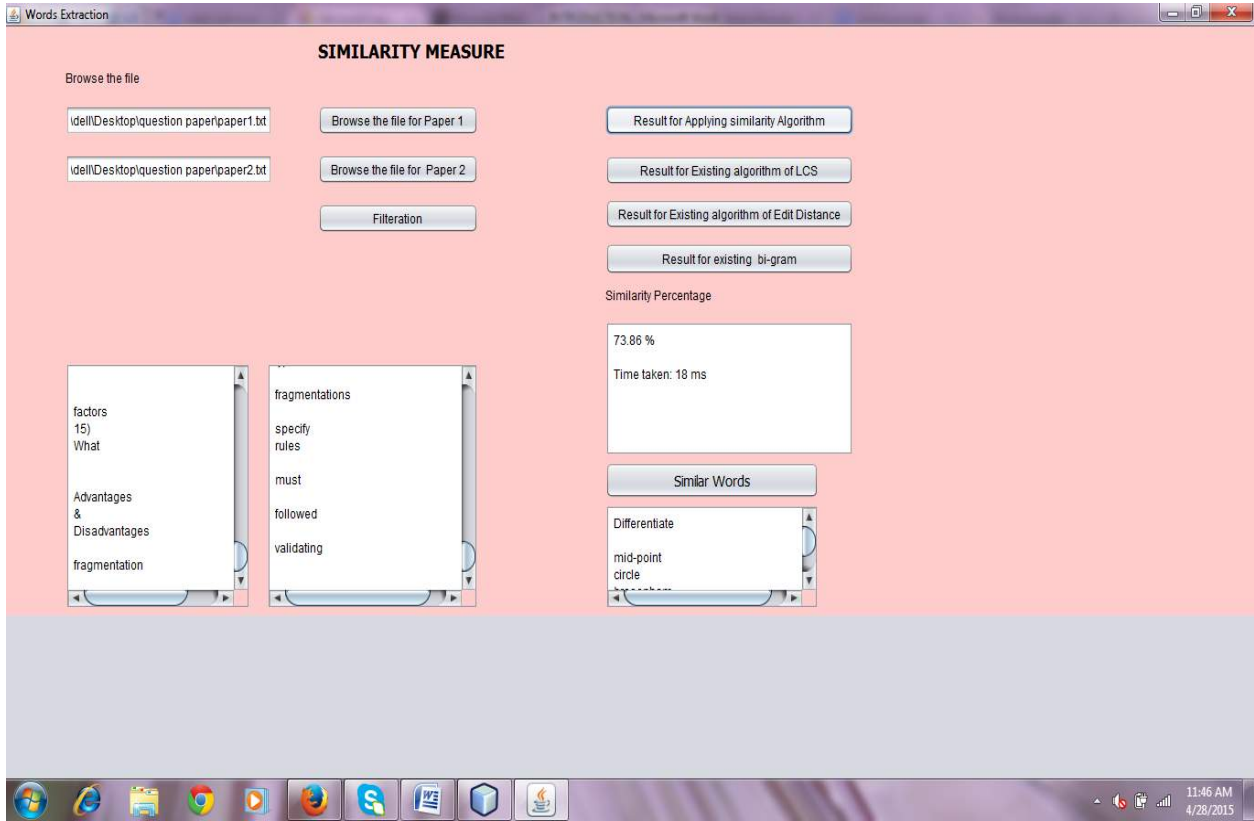


This graph has show the result of the existing algorithm and the proposed method in which I work.



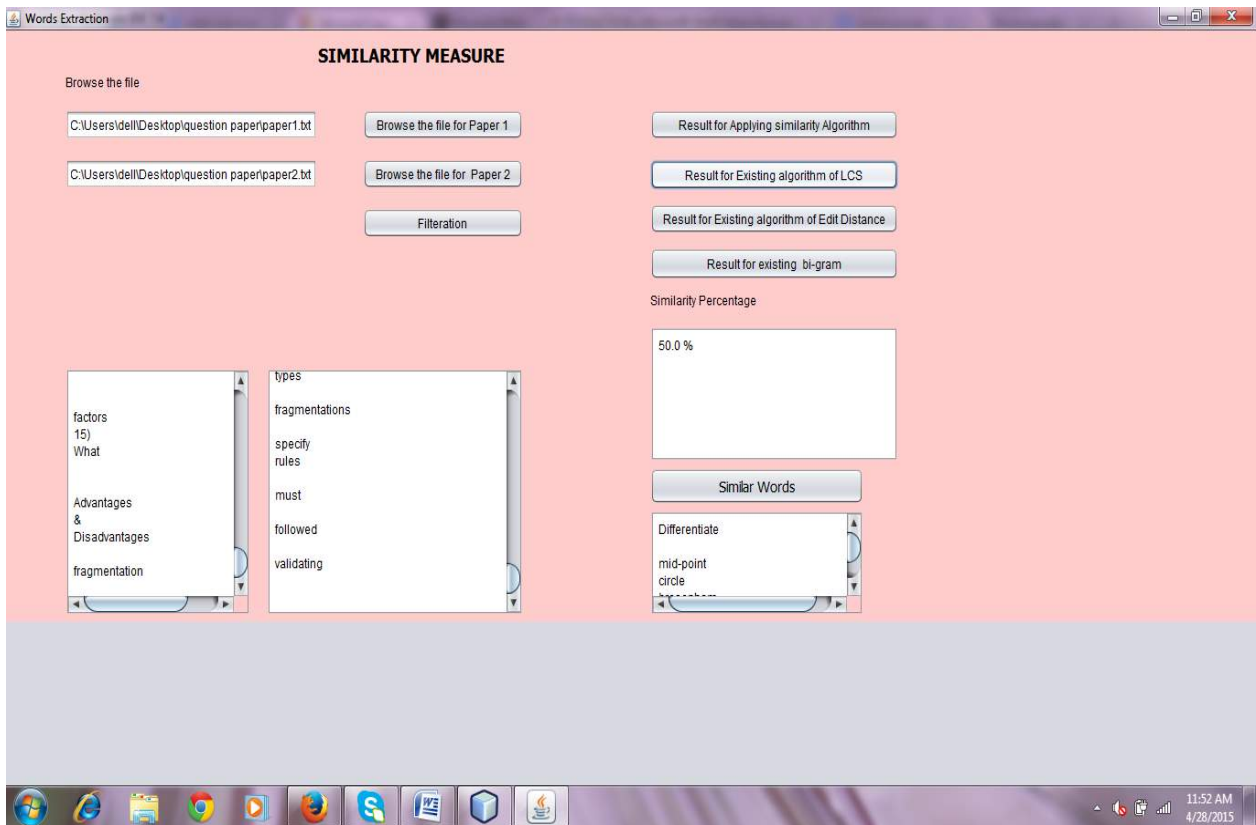
As I decided to work on the syntactic similarity between the two text, to find the similarity between the two question paper. It is the interface which I developed to find the similarity accuracy. It is used to find the similarity between the two question papers. There is also a various methods which is used to find the accuracy, but the new string similarity metric algorithm shows the better result comparative to edit distance, LCS and Bi-gram algorithm.

## Applying the New string similarity metric algorithm



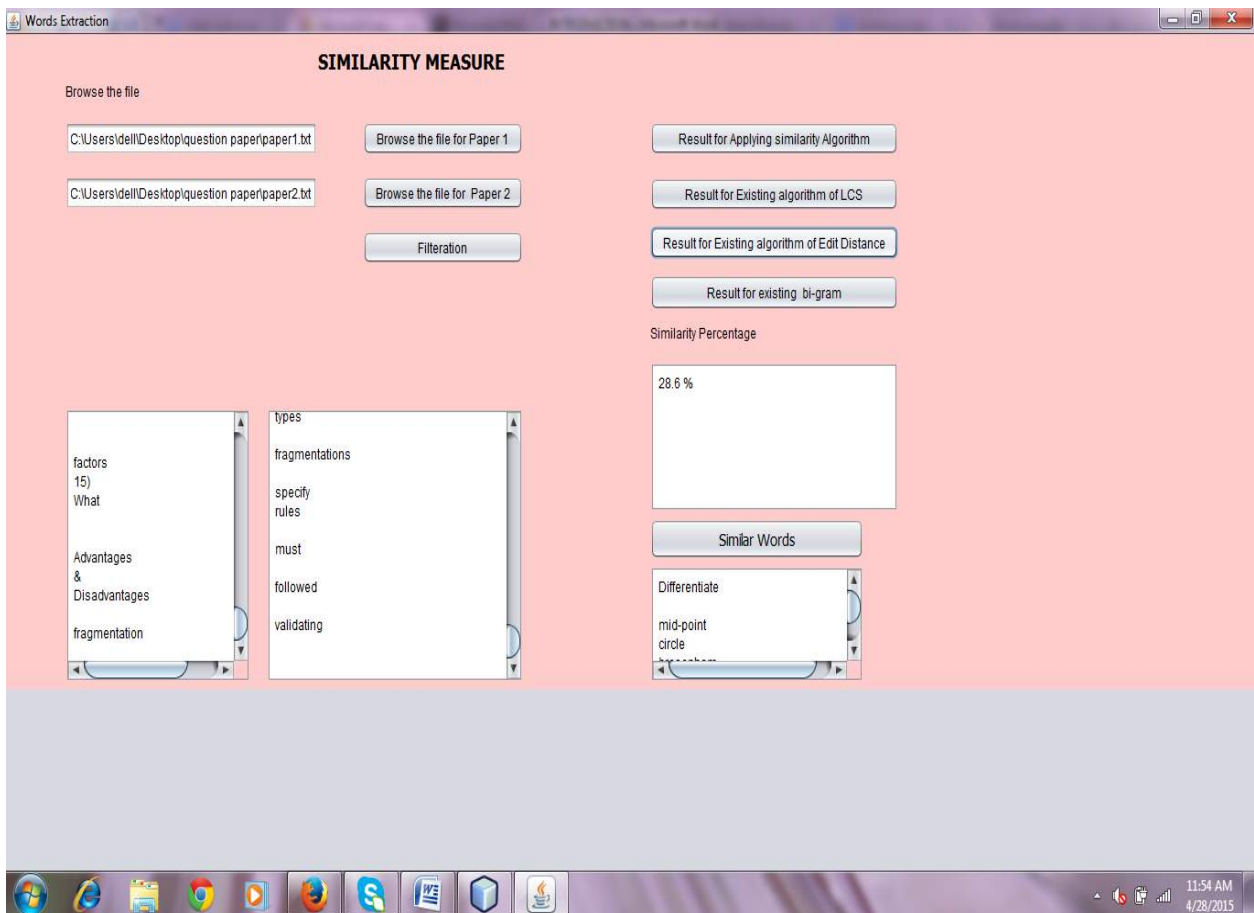
With applying the new string similarity metric algorithm the result for question paper1 and question paper 2 is 73.86%.

## Applying the LCS algorithm



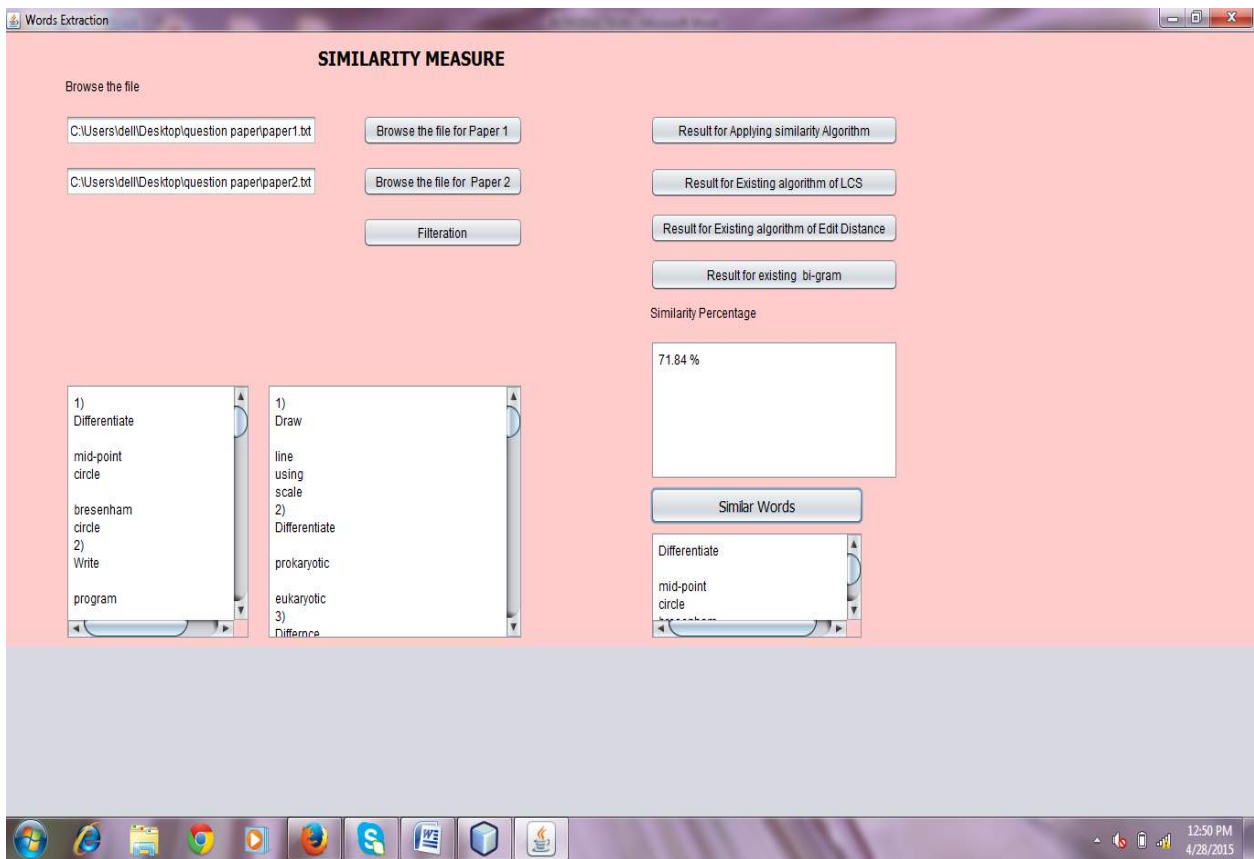
With applying the longest common substring algorithm the result for question paper1 and question paper 2 is 50%.

## Applying the Edit Distance algorithm



With applying the Edit Distance algorithm the result for question paper1 and question paper 2 is 28.6%.

## With applying the bi-gram algorithm



With applying the Bi-gram algorithm the result for question paper1 and question paper 2 is 71.84%

From all the results its clear that the new approach which I present in the paper, the accuracy of that is high comparative to the another algorithms.

## CHAPTER-5

# CONCLUSION & FUTURE SCOPE

---

In this research, NLP is used to improve the accuracy of repeated questions in the question papers. So that with the used of developed methodology our system will easily find the repeated words which are present in the question paper.

### **5.1 Conclusion:**

It may happen many times, in a question paper similar question can be occurred or it may also be happened that the questions are related to each other. So to ignore this type of problem, we proposed a method in which the developed system may try to find those questions which are similar to question paper, so that the possibility of relevant questions are decreased in the future time. From all the literature review it is clear that there is not much work on the syntactic similarity between two short segments, so improvement is done to measures the similarity between questions in two question papers (aka automated question vetting).The future work is on to improve the approaches to measure the syntactic similarity between two short texts. In the data mining field there is more work which is based on the semantic similarity between short texts. So I decided to work on the syntactic similarity between two texts. The accuracy of repeated words in the two question paper is 70%.

### **5.2 Future scope:**

In the future, it's very important to increase the accuracy to find the syntactic similarity between two texts. But for the further improvement it's better to study on the semantic similarity within the syntactic base approaches.

## REFERENCES

---

- [1] D. R. Cutting, J. O. Pedersen, D. Karger, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In Proceedings SIGIR '92, pages 318-329, 1992.
- [2] James O'Shea, Zuhair Bandar, Keeley Crockett, and David McLean, "A Comparative Study of Two Short Text Semantic Similarity Measures," unpublished.
- [3] Zhong Min Juan, "An effective similarity measurement for FAQ Question Answering system," *International conference on electrical and control Engineering, 2010*.
- [4] Federica Mandreoli, Riccardo Martoglia and Paolo Tiberio, "A Syntactic Approach for Searching Similarities within Sentences," unpublished.
- [5] Abolfazl Keighobadi Lamjiri, Leila Kosseim and Thiruvengadam Radhakrishnan, "Comparing the Contribution of Syntactic and Semantic Features in Closed versus Open Domain Question Answering," *International Conference on Semantic Computing*.
- [6] Xinxin Zhao, Tiedan Zhu and Yushu Liu, "Document Classification in Different Granularity," *Computer Engineering, Vol.32 No.20, p.183-184, Oct 2006(In Chinese)*
- [7] Vasileios Hatzivassiloglou, Judith L. Klavans and Eleazar Eskin, "Detecting text similarity over short passages: Exploring linguistic feature combinations via Machine learning," unpublished.
- [8] Xu Liang and Dongjiao Wang, "Improved Sentence Similarity Algorithm Based on VSM and Its Application in Question Answering System," *International conference on electrical Engineering, 2010*.



- [9] Ko, Y., Park, J., and Seo, J., “Improving text categorization using the importance of sentences”, *Information Processing and Management Vol. 40, No.1, pp. 65–79, 2004.*
- [10] Liu, X., Zhou, Y. and Zheng, R., “Measuring Semantic Similarity in Wordnet”, *Proceeding of ICMLC2007 Conference, Hongkong, 2007.*
- [11] LIN LI, XIA HU and BI-YUN HU, “Measuring Sentence Similarity From Different Aspects,” *Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009*
- [12] Manasa.Ch and V. Ramana, “Measuring semantic similarity between words using page counts and snippets,” *Manasa ch et al, International journal of computer science & communication network, vol 2(4), 553-558*
- [13] Takale, S.A. and Nandgaonkar, S.A (2010) ‘Measuring semantic similarity between words using web documents’, *IJASCA-International Journal of Advanced Computer Science and Applications, Vol 1, No.4, pp.78-82.*
- [14] Zhiqiang, L., Werimin, S., and Zhenhua, Y. (2009), ‘Measuring Semantic Similarity between Words Using Wikipedia’, *WISM -International Conference on Web Information Systems and Mining, pp: 251-255.*
- [15] LIN LI, XIA HU and BI-YUN HU, “Measuring Sentence Similarity From Different Aspects,” *Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009*
- [16] R. Menaha and G. Anupriya, “Semantic similarity between words using SWD and snippets,” *International conference on current trends in advanced computing, 2013.*
- [17] Yi Liu and Qiang Liu, “Sentence similarity computation based on feature set,” *13<sup>th</sup> International conferences on computer support cooperative work in design.*

[18] Hongni Dong, Jiang Wu and Xiaohui Zhao, “Study on the calculation of text similarity based on key-sentence,” *2010 International Conference on E-Business and E-Government*.

[19] Wenpeng Lu, Jinyong Cheng and Qingbo Yang, “ Question answering system based on web,” *5<sup>th</sup> Internatonal conference on intelligent computational technology and automation, 2012*.

[20] D. Bollegala, Y. Matsuo, and M. Ishizuka. Disambiguating personal names on the web using automatically extracted key phrases. In *Proc. of the 17th European Conference on Artificial Intelligence*, pages 553{557, 2006.



