International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 08, August-2019

# Punjabi Speech based Searching in a Text Document

Rajneet Kaur
Dept of Computer Science and Engineering
Punjabi University, Patiala

Dr. Williamjeet Singh
Dept of Computer Science and Engineering
Punjabi University, Patiala

***Abstract*- In real time scenario, searching the content in a document takes time and less research has been done with respect to speech searching. So, in this proposed work these barriers of searching have been tried to be eliminated. In proposed work, effort is being made to search a word in a Punjabi text document using a Punjabi speech corpus with high efficiency searching rate. The speech corpus is a collection of 200 unique words which consist of some commonly used words in daily life and the text document (.txt) consist of total 500 randomly selected sentences which are composed of those common words. An algorithm has been developed for the same which will search a given spoken word in the document and will extract that word wherever it is present and further processing can be done. Moreover, a provision has been made where the frequency of that word has been calculated, which can be used for text as well as sentiment analysis.**

***Keywords – Searching, Speech corpus, text document***

## 1. INTRODUCTION

Data Retrieval is the method by that a collection of data is described, stored, and searched for the aim of data discovery as response to a client demand[1]. It is the science of searching for information and retrieving information[2].Today there are usually immense of data keep in our local system, we regularly pay a lot of time on looking data that are needed and many kind of data which is historically gathered and saved is increasing rapidly[3]. The traditional data retrieval searching can't meet the present needs of the users thus it will inevitably result in improvement of searching mechanism[4]. Text searching is an array of characters which contain space or special characters are searched. There has been a lot of research on text searching problem but less in context with speech where you speak a word or line and it is searched [5]. Text searching is essential in several applications and disciplines like signal processing, text analysis, authentication and verification, speech analysis and recognition, data retrieval, DNA sequencing and forensics, information compression and computational sciences[6]. The application used to search and find appropriate documents on chosen topics from a database of texts is text-based application

## 2. LITERATURE SURVEY:

i. This paper introduces data retrieval using Hoot, together with categorisation and system design, compares data retrieval of Hoot and data retrieval of Lucene, the experimental results show that Hoot offers efficient data retrieval and has quicker retrieval speed than Lucene[1].

ii. In this paper, a brand new technique of efficient search of a sequence of words in a very massive document using the ides of hashing and linked list is given. For building the hash table, the words are organized so as the amount of occurrences and positions of every word and positions of its next words are added to the record of every word within the hash table. Any word may be searched employing an easy hash function and therefore the succeeding word can be set using the position of next words since this data is accessible with the record of current word creating the hash table as a linked list of words[6].

iii. This paper proposes intelligent categorisation of text documents by classifying a text document supported its content and also the presentation type of the content of the text document style and implementation of few indexing algorithms and searching algorithms on text documents are conferred with a discussion on the benefits and limitations of every algorithms[2].

iv. Now a day, local system keeps variety of files or documents. At an equivalent time, the amount of varied multimedia system files keep in local system, like image, text files, audio and video files. Recent drawback, user how to realize and find file or document as quickly as possible thus this paper present Desktop Full Text searching. We have a tendency to perform the two searching algorithms on projected search system. Keywords searching and File searching algorithms are used for searching on projected search system. The performances of these two algorithms are searched on six differing kinds of documents (like as pdf, txt, doc, docx, and xml) using local system information[4].

v. A Chinese document retrieval methodology enhanced by concept base is proposed during this paper. The main plan of this methodology is to make a typical Chinese concept base to provide a shared understanding of ideas. This enhanced methodology will take advantage of the concept base once analyzing and classification documents, and once looking documents. The document management system will use this methodology to enhance the retrieval performance[3].

vi. The retrieval efficiency of the presently used systems cannot be considerably improved: "Bag of words" interpretation caused loosing linguistics of texts. We

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 08, August-2019

have a tendency to apply the practical approach to present English text documents within the memory of computers. It allows keeping linguistics relations between words once indexing documents and use normal English sentences as queries to experience under a search engine. The planned retrieval algorithm returns extremely relevant documents. The given example illustrates the advantage of the mentioned approach compared to the traditional key word search[5].

## 3. METHODOLOGY:

In this section, the technique of efficient search of a sequence of words in a large text file using is demonstrated[7]. The flowchart shown below shows all the required steps to search a word in a given text file[8]. The main aim of this paper is to search a word in a text file using speech which makes searching a word more easy and efficient.

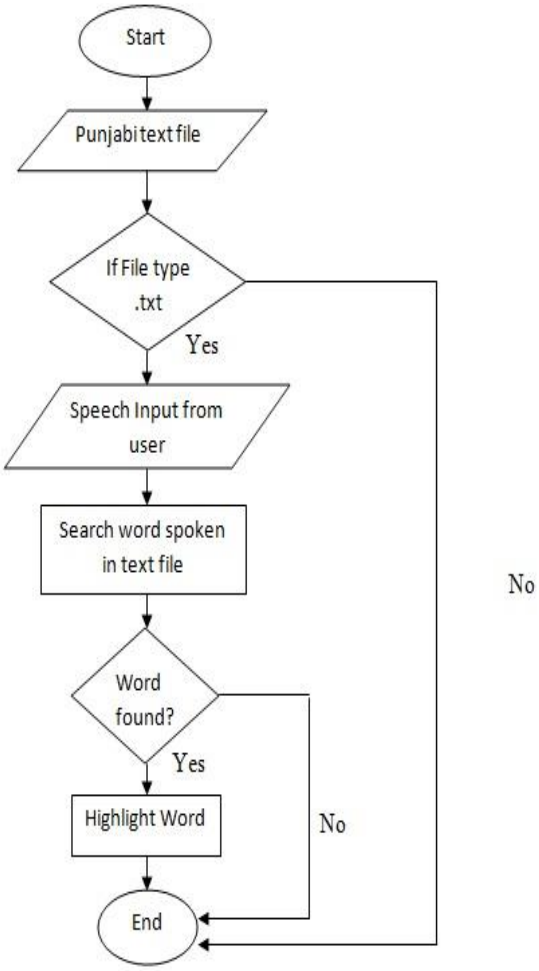


Figure: 1 Flowchart

Firstly, a Punjabi text document is prepared in which proper sentences are stored which includes common words from categories like months, weekdays, fruits, vegetables etc. The file type of text document is .txt which includes multiple pages[9]. Then, Punjabi speech corpus is collected and maintained which include words from those categories. The speech corpus consists of 200 unique words that are present in the text document. This speech corpus will help in searching a word in a text document when spoken which makes the searching fast and efficient.

## 4. ALGORITHM:

**Step 1.** Input a text document (TDi) with Punjabi sentences
**Step 2.** If document file type is .txt (TDt) go to step 3
Else go to step 9
**Step 3.** Speech input (SIu) from user as a word
**Step 4.** Finding word in a speech corpus of 200 words.
**Step 5.** Matching word in a text document uploaded at the back end.
**Step 6.** If word matches (SIu = TDi) go to step 7
Else go to step 9
**Step 7.** Highlight the word where it is present in the document.
**Step 8.** Searching done and Criteria matched
**Step 9.** Stop searching

## 5. RESULTS AND ANALYSIS:

The framework has been prepared in this section tells us how the searching of a word can be done using a Punjabi speech corpus from a text document file. For efficient search some scenarios are made with respect to corpus size of the unique words and text document which consist of sentences as shown in Table 1.

| | Corpus size (in words) | Text document size (in sentences) | File type of text document |
|---|---|---|---|
| Scenario 1 | 50 | 100 | .txt |
| Scenario 2 | 100 | 200 | .txt |
| Scenario 3 | 150 | 300 | .txt |
| Scenario 4 | 175 | 400 | .txt |
| Scenario 5 | 200 | 500 | .txt |

Table 1: Scenarios

As shown in above figure, if scenario 1 is taken into consideration firstly 50 words speech corpus can be taken and the text document file size will be 100. Searching will take place in these 100 sentences which is a small and searching a word will be fast. Similarly, if we consider scenario 2, file size of text document will be 200 and speech corpus will be 100 set of words and same with other scenarios.

Searching a text in a document can take advantage of speech to make the search fast and efficient. As of now total 500 sentences in a text file are taken to search a word in it and it will help people to search a word in document by only speaking that word, they do not have to write a word to be searched.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 08, August-2019

## 6. DISCUSSIONS AND FUTURE SCOPE:

Data retrieval with speech recognition research is to address recognition problem computationally by building systems that map from an acoustic signal to a string of words is a significant field of study[10]. As speech is a primary mode of communication so it help people to interact with systems easily. Limited researches had been done in this field and several flaws were left unattended. In this paper, an approach is being made to investigate

- Data retrieval in a text document using speech corpus.
- Data retrieval done only in a text document file type .txt which gives good results.

## FUTURE SCOPE:

In future, the present work may be extended on the following lines:

1. The speech corpus size can be extended to more words.
2. Larger text document in sentences can be build for searching.
3. Searching using corpus can be done in file types like .pdf , .ppt , .xml , .docx` .
4. Searching using speech corpus can be done in other languages like English, Hindi, Urdu etc.
5. Frequency of words can be measured in further work.
6. Time taken to search a word with speech can be calculated.

## REFERENCES

[1] S. J. Patil and D. K. Budhwant, “Efficient Information Retrieval Using Indexing,” vol. 6, no. 2, pp. 106–109, 2017.
[2] D. Minnie, “Intelligent Search Engine Algorithms on Indexing and Searching of Text Documents using Text Representation,” pp. 121–125, 2011.
[3] J. Su, W. Weng, and Z. Wang, “A Chinese Document Retrieval Method Enhanced by Concept Base,” pp. 200–203, 2009.
[4] S. Lakhara and N. Mishra, “Design and Implementation of Desktop Full-Text Searching System,” *2017 Int. Conf. Intell. Sustain. Syst.*, no. Iciss, pp. 480–485, 2017.
[5] V. Klyuev and V. Oleshchuk, “Semantic Retrieval of Text Documents,” pp. 189–193, 2007.
[6] M. N. Kabir, Y. M. Alginahi, and O. Tayan, “Efficient Search of a sequence of words in a Large Text File,” pp. 0–5, 2014.
[7] B. W. Gawali, “A Review on Speech Recognition Technique,” vol. 10, no. 3, 2010.
[8] R. Kaur, “Speech based Retrieval System for Punjabi Language,” *2018 Int. Conf. Smart Syst. Inven. Technol.*, no. Icssit, pp. 498–502, 2018.
[9] B. Singh, N. Kapur, and P. Kaur, “Speech Recognition with Hidden Markov Model : A Review,” vol. 2, no. 3, 2012.
[10] R. Sultana and R. Palit, “A Survey on Bengali Speech – to – Text Recognition Techniques,” pp. 26–29, 2014.