Survey on Information Retrieval and Pattern Matching for Compressed Data Size using the SVD Technique on Real Audio Dataset

Poonam Dhumal
Department of Computer Engineering,
Pimpri-Chinchwad College of Engineering,
Pune-411044

S. S. Deshmukh
Department of Computer Engineering,
Pimpri-Chinchwad College of Engineering,
Pune-411044

ABSTRACT

Due to increasing size of text and audio data over internet, various techniques are needed to help with the finding and extraction of very specific information relevant to a user's task. Text mining is a variant on a field called data mining that tries to discover curious patterns from large databases. Singular value decomposition this technique is used for dimensionality reduction of large database on Apache MAHOUT Hadoop framework. In this paper, different existing Information Retrieval, Pattern Matching, Rule Generation Algorithm is reviewed. In addition for extracting questionnaires and curiosity based sentences from large database some different implementation of the algorithms is proposed. Finally the extract user required information from large unstructured database.

Keywords

Apache MAHOUT; hadoop; imformation retrival; pattern matching;rule generation.

1. INTRODUCTION

Text mining is a variant on a field called data mining that helps users to find interesting patterns from large databases. Text mining is an interdisciplinary part of field which draws on data mining , information retrieval, statistics, and machine learning. Text analytics helps analyst's user extract pattern, meaning, and structure hidden in unstructured text data.

1.1 Steps Of Text Mining

Text mining involves application of techniques from different areas like information retrieval, natural language processing (NLP), information extraction, and data mining. By using these type of various techniques are mining the useful information from text.

1.1.1 Information Retrieval (IR)

Information System systems recognize the documents in a collection which match a user's query. The most famous Information Retrieval systems are search engine such as Google, which identify those documents on the World Wide Web that are relevant to a set of given words. In libraries, where the documents are typically not the books themselves but digital records holding information about the books there IR systems are often used[1].

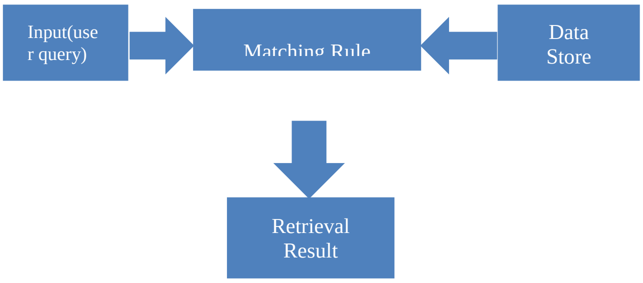

Fig.1.Information Retrieval system

1.1.2 Natural Language Processing (NLP)

Natural Language Processing is the one of the most difficult problem in artificial intelligent. NLP is analysis of natural languages, so computer can understand the human languages. The main role of NLP in text Mining process is to make available the systems in the information extraction phase with language data that they need to perform their task. Often this is done by marking text documents with information like sentence boundaries, part-of-speech tags, parsing results, which can then be read by the information extraction tools. Pattern matching is the one of the important task in NLP[1]. It match the patterns related to user query and easily result get it back to the user.

1.1.3 Data mining

Data mining, the extraction of unseen analytical information from large databases. Data mining is the process of identifying unique patterns in large sets of data. The aim is to find previously unknown, convenient knowledge. Data mining is applied to the details generated by the information extraction phase when used in text mining.

1.1.4. Information Extraction

Information Extraction is the process of spontaneously finding structured data information from an unstructured natural language document which is in the large size. Often this includes defining the general form of the information that are interested in as one or more patterns, which are then used to guide the information extraction process. IE systems depend on the data generated by natural Language Processing systems.

1.2 Singular value Decomposition (SVD)

Singular Value Decomposition is the dimensionality reduction technique. Singular Value Decomposition is a matrix factorization technique which takes a rectangular matrix of text defined as A. Where A is a m x n matrix in which the m rows represents the terms, and the n columns represents the documents The SVD performs its operations on matrices. The

Singular Value Decomposition (SVD) of a rectangular matrix representation A is a decomposition of the form

A = U S VT

The main working to working with SVD of any rectangular matrix A is to consider decomposition matrix AAT and ATA. The columns of U matrix, that is m by m, are AAT, the columns of V, that is n by n, are ATA for calculating the eigenvalues and eigenvectors. The singular values on the diagonal of S, that is m by n, are the positive square roots of the nonzero eigenvalues of both AAT and ATA. Eigenvalue-eigenvector factorization i.g. A = USVT , where UUT=I ,$VV^T=I$, and S= singular values. So finally get reduce matrix.

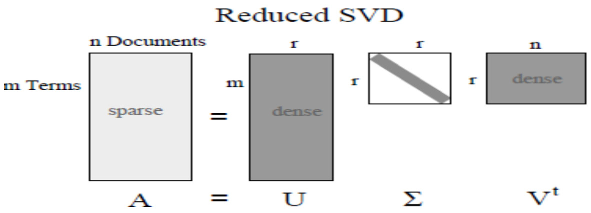

Fig 2. SVD Representation

Apache Mahout

Apache Mahout framework is one of the Apache Hadoop framework projects, which is a collection of libraries, application, and implementations for scalable machine learning functions. For collective intelligence system (CI), Mahout contains three types of algorithms: recommender system, clustering, and classification data mining algorithms[2].

Our implementation is based on Apache Mahout, which has implemented an Singular value Decomposition algorithm in HADOOP. It also has the Map Reduce implementation of the SVD algorithm. Parallel it reduce the size of data and increase the performance of the system.

2. RELATED WORK

In literature, Mrs. Sayantani Ghosh, Mr. Sudipta Roy, and Prof. Samir K. Bandyopadhyay[1] proposed When a user gives a set of words as input query for a search of particular information, Google execute the search on the existing documents available on the World Wide Web to find a match for the required information as per the user's query. While Data mining is usually concerned with the detection of patterns in numeric data, very often important information is stored in the form of text document. A first goal in data mining is feature extraction, i.e., the identification of the terms and concepts most frequently used in the input text documents; a second goal typically is to discover any associations between features. Hence, a first step to text mining typically consists of "coding" the information in the input text; as a second step many methods such as to determine relations between features Association applied Rules algorithms.

Che-Rung Lee, Ya-Fang Chang [2], proposed a new collaborative filtering algorithm for the application of recommendation systems. The stochastic SVD to enhance the accuracy and the performance of the original Apache Mahout implementation. In this paper HADOOP Map Reduce implementation, a number of techniques are applied for performance optimization, including computation generalization, file input output(IO) reduction, truncation of small elements, tiled algorithm, and the tournament selection[2]. Experimental results showed there algorithm and implementation is about 3 times more accurate and 2.5 times faster than the original algorithm/implementation for the testing data of ten millionsRecords[2].

Suresh Fatehpuria, Ankur Goyal [3],proposed a very novel idea for finding a pattern in a given text by pre-processing the text[3]. The first pre-processing phase of Jumping Algorithm helps the matching phase in taking the jumps in given sentence. The asymptotic analysis shows that the preprocessing phase that takes O (n+∑) time for putting the indexes in two dimensional arrays may be a little bit costly but the matching phase is amazingly cheaper in comparison of other existing algorithms[3]. For small texts document, the matching phase is almost constant. This less number of matching is what the key idea behind the algorithm. Since the algorithm shows its excellent behavior in terms of complexity, it can be adopted in string matching's practical applications[3].

3. PROPOSED WORK

In this paper propose the solution to extracting the most curiosity and questionnaires based sentences from large size of text document. To developed three modules to perform the task of unique pattern retrieval. The first module audio to text convertor which has main tasks converting any real time audio to text document. The second module is the implementation of SVD algorithm, which used the dimensionality reduction. The third module is the implementation of Pattern matching and Rule generation algorithm, which used to extract the unique patterns.

3.1 Audio to text

For any real time audio to text conversion using no of tools, Google APIs, or through programming, so in our proposed work convert one real time audio to text document through java programming. Finally get one large size of unstructured format text document.

Fig 3. Speech to Text conversion model

3.2 SVD Implementation

Singular Value Decomposition is the one of the matrix factorization method. SVD is reduces the dimensionality of the matrix. When our database is large in size, so SVD can reduce the irrelevant things and compress the size of the matrix.

3.3 Proposed Pattern Matching Algorithm(Jumping Algorithm)

The pattern matching algorithm propose in this paper consists of two phases:

l. Pre-processing of given text: For pre-processingthe text, a two dimensional array i.g. arr[][] is taken. The arrayshould have ∑ numbers of rows , where ∑ is the size of thetotal number of alphabet (number of identical characters in the text), no matter

how many columns are there[3]. For example row 0 is set to a,1 is set to b and so on.

Matching of questionnaires words pattern in the given large text document: When the pre-processing phase of the text document is done, the pattern is then targeted[3]. For that match the WH type patterns example what, how, when, where, why, when, whom. First of all, the first character of the questionnaires pattern is checked. After getting the first character of the pattern, the index of that character's first occurrence is known from the two dimensional array by searching the index in corresponding row[3]. Once the index is retrieved, the matching process is started from the right next index of the retrieved one text. This process ensures skipping all the words those does not start from the first character of the questionnaires pattern. For instance, if the alphabet size I is let say 26, then all the words those start from at least 25 different characters will be skipped[3]. During the matching process, if the mismatch occurs, the next occurrence of the pattern is retrieved from the two dimensional array and again the matching process is started at the newly retrieved index[3]. By using jumping Algorithm match the all Interrogative words in given documents.

3.4 Rule Generation

Using Jumping Algorithm getting unique patterns. By using that patterns generate the rules for extraction of whole questionnaires sentence. Rule-based pattern extractor and a Semi-Supervised NER approach to automatically generate extraction pattern from a limited corpus and label the pre-defined entities in a collection of accident documents [4]. Link Grammar parser and Stanford Part- of-Speech tagger are used in the pattern extractor to identify named entity and construct extraction pattern [4]. The extraction pattern then feed to Semi Supervised NER to categorize the entities into some predefined categories [4]. A cascade approach is generally used where some basic questionnaires are first identified and are latter combined into more complex ways. That time generate the rules like:

E.g. Wh<A-Z>*<? >

patterns for definition questions

- Question: What is A?

1. <A; is/are; [a/an/the]; X>
2. <A; comma; [a/an/the]; X; [comma/period]>
3. <A; [comma]; or; X; [comma]>
4. <A; dash; X; [dash]>
5. <A; parenthesis; X; parenthesis>
6. <A; comma; [also] called; X [comma]>
7. <A; is called; X>

Like this generate the rules for unique patterns.

3.5 Proposed Architecture

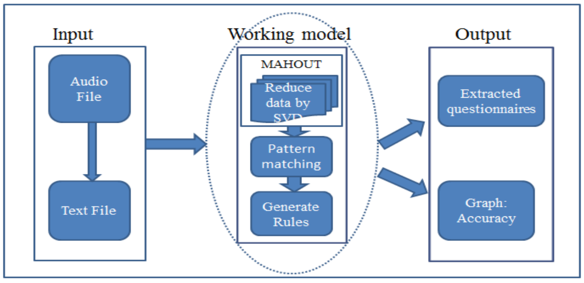

Fig 4. Propose Architecture Diagram

Above propose architecture shows the actual flow of the system. This propose system working through three modules, finally it gets the most curiosity based and questionnaires based sentences extracted from given database.

3.6 Use Case Diagram For Propose System

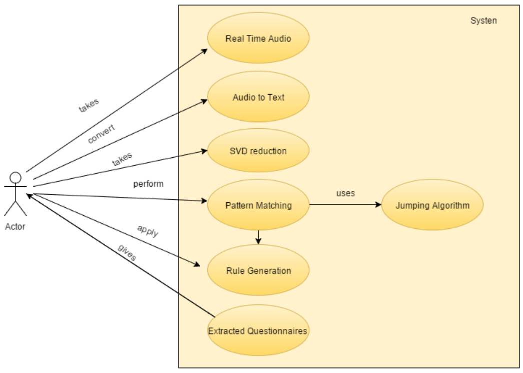

Fig 5. Use Case Diagram

Use Case diagram shows the overall flow of the Propose system. Also it shows the relation between actor/user and the system. when user gives some input as a speech then system gives extracted curious sentences from given large text document.

4. CONCLUSION

At last conclude that, Text mining is also known as Text Data Mining or Knowledge-Discovery in Text (KDT) [5], SVD can reduces the size of the database so easily extract information from compress database. SVD work parallel on apache Mahout to reduce the implementation time, and increase the performance of the system. This paper presents a new idea for finding a pattern in a given text by pre-processing the text. The pre-processing phase of Jumping Algorithm helps the matching phase in taking the jumps [3]. When unique pattern are collect then easily generate the rules for whole curious sentence. Finally get the unique patterns of questionnaires base sentences which is reduce in size of database.

5. REFERENCES

[1] Mrs. Sayantani Ghosh, Mr. Sudipta Roy, and Prof. Samir K. Bandyopadhyay, 2012 A tutorial review on Text Mining Algorithms.

[2] Che-Rung Lee, Ya-Fang Chang, 2013 Enhancing Accuracy and Performance of Collaborative Filtering Algorithm by Stochastic SVD and Its MapReduce Implementation

[3] Suresh Fatehpuria, Ankur Goyal, 2014 A Very Unique, Fast and Efficient Approach for Pattern Matching (The Jumping Algorithm).

[4] Yunita Sari, Mohd Fadzil Hassan, Norshuhani Zamin, 2010 Rule-based Pattern Extractor and Named Entity Recognition: A Hybrid Approach.

[5] Haralampos Karanikas and Babis Theodoulidis Manchester, 2001 Knowledge Discovery in Text and Text Mining Software", Centre for Research in.