American Journal of Intelligent Systems 2017, 7(3): 54-59
DOI: 10.5923/j.ajis.20170703.03

Converting and Deploying an Unstructured Data using Pattern Matching

Anujna M.*, Ushadevi A.

Department of Computer Science and Engineering, St. Joseph Engineering College, Mangaluru, India

Abstract Text mining is also known as knowledge discovery from textual databases; its job is to derive a high level knowledge from the text. The process of obtaining useful information from the records is also known a text mining. The system uses many data mining approaches to extract the patterns from the text documents. The challenge is using those updated patterns and implementing an algorithm for pattern discovery is still an open research issue. The paper focuses on using the pattern matching technique for regular expression to find the relevant data from the text/word file. The text file containing large number free-text is used to fetch all the discovered words or characters from the documents. The system is helpful for the users to search the relevant document, and converts all the unstructured data into structured form.

Keywords Text Mining, Document Clustering, Pattern Discovery, Pattern Matching, Pre-processing

1. Introduction

Text mining is a discovery of knowledge from textual databases; its task is to extract a effective knowledge from the text. The process of deriving useful information from the records is also known as text mining. There are several methods of data mining techniques used to select the patterns from the text files. The dataset used in the data source could be in the form of an unstructured or semi structured text documents.

The text mining process contains: The basic operations such as extrication and identification of the text pattern. Document consists of a discontinuous or a discrete textual data. i.e. by collecting the real universe documents like business report, email, and news story etc. whereas the categorization of documents based on texts is called document clustering. The document can be picked out from any of unstructured files or documents. The format of unstructured documents is of free style text and it does not carry any precised format. The structure of the text is inadequate compared to the structured format. These types of unstructured document are mostly found in research areas and the online documents type like the web pages.

Pattern discovery is a text mining technique; its main function is to select the document text of different pattern. The different pattern styles may consists of a word or a group of words. The main objective of the paper is to identify the patterns and placing it over an unstructured data. Also to use those deployed data to get in a proper document.

Text mining is a process, in which structured information is collected from an unstructured text, and it is also used to extract and discover high quality knowledge automatically hidden in texts. Unstructured data refers to information does not hold any built-input text files. This data is usually in a paragraph or passage. The data might contain file with several information. It is an overall description of selecting an input from a file. The data in a file does not contain any database. Unstructured data can be word-based or non-word based texts.

2. Literature Review

The first paper focuses on the challenges faced in different kinds of text mining techniques are used in the extraction of specific information from the required documents. The commonly used text mining techniques in the data mining is the approach of term based. This paper concentrates on the new ideas which use different concept to discover the pattern efficiently to get relevant data in [1]. An automatic text development approach was implemented is based on the categorization concept. The categorization approach made use for the document text extraction was implemented using the real world information collected from various databases across the world. The outputs are demonstrated for showing the performance and the quality of the categorization approach in [2]. A data is extracted from the information document which is called as Information Extraction (IE) and is based on the rule model. The algorithm used is based on the unsupervised strategy, also made use of the inductive learning approach for the text mining. To identify the natural languages the morphological features of the documents are

* Corresponding author:
anujnamanu3@gmail.com (Anujna M.)
Published online at http://journal.sapub.org/ajis

used in the processing of the documents. Based on the information extracted during the experiments the patterns are constructed in [3]. A query compliable algorithm for unstructured text data are text bound for the patterns. On the special and regular dictionary expressions, the text queries take a huge amount of time. This technique works for the token detection of the sequenced pattern. This algorithm has successfully extracted and detected the information from the bundles of several dictionaries. This improves the storage for the numerous words in the dictionary in [4]. A paper presents a result outline for finding out the unstructured data in the student forums. The text analytical extraction tool is used for finding the unstructured data and this tool predicts the performance of the student in the academics from the documents. By using the key graph the results from the analysis is build so the professors in the college can access the results and improvise the capability of their teaching and experiments and with different advanced learning ideas to improve the skill and performance of the students in [5].

3. Proposed Methodology

In text mining, the effective pattern discovery technique is a system used to fetch the discovered patterns to identify the relevant data from the document. i.e. By using the word or text files, the system can use those retrieved data to find the relevant results obtained in the database form. The document in the form of word or text files is given as an input to the system. In this process, the paper will use different pattern discovery techniques.

The newly proposed scheme consists of following five modules:

1. Pre-processing Module

In this module, the specified document undergoes two major processes:

a. Removal of the stop word

In the stop word process, the document is scanned thoroughly to find all stop words and these stop words are eliminated from the text or word documents to get a free stop words. eg:"he","she","for","is","and" etc.

b. Stemming of the text

In the stemming process, the filtered document from the previous step will be broken down to their stem or the root.

Example used for this are words like designing is changed to design, hopefulness is changed to hopeful. The task of the stemming process is to remove the tail part of the word in text files. Some of the generally used words are: "ing","ed","es","al","y","ful","ize" etc.

2. Regular Expression Module

Here to identify the text patterns of an unstructured document the regular expression technique is used. The pattern technique is also useful for extracting all the structured data from the files.

3. Deploying Pattern Module

In this module, the numbers of patterns are discovered from the text documents uploaded by the user. It is used to deploy all the text files by using regular expression as a pattern.

4. Evolution of Inner Pattern Module

Here the noisy data and shuffling is identified by removing the noisy patterns. It consists of:

c. Removal of noisy data

It includes some of the false data in documents which by mistake have been considered true data. It shows that there are some of the erroneous data which may be present in the original document collection. These patterns are generally referred cause problems or do something wrong. In the proposed system, noisy pattern removes the unwanted symbols such as semicolon, brackets etc.

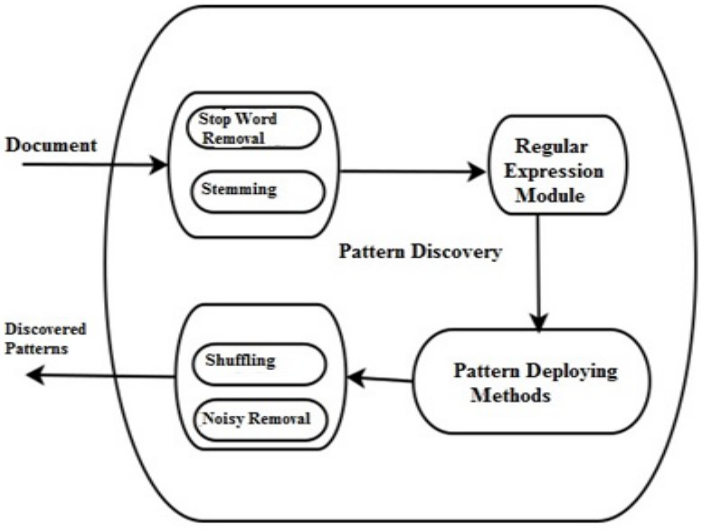

Figure 1. Basic Architecture of Pattern Discovery

d. Shuffling

The shuffling technique is used to remove the noisy patterns by selecting the unnecessary patterns in text files. In the proposed system the data received from the documents are arranged in alphabetical order. The shuffling process is used because the given inputs may be in different document format.

Structured Data Construction: The data will be continuous in an unstructured format. Once the discovery of pattern is done, the system constructs the data into structured form i.e. column wise. The final outcome of the system gives the tabular form structured data i.e., excel sheet.

5. Pattern Matching Module

The system uses matching module to match the patterns from the extracted words and finds the relevant documents as an output. The pattern matching technique is derived for different parameters. It includes username, email, date of birth and mobile number. By using regular expression technique the pattern is matched and discovered.

4. Implementation

The system is designed for windows application by creating forms for different modules using visual studio as front end and SQL database in the back end. In the paper, the two algorithms are used as following:

a) Porter D Algorithm for Stemming Module:

For the stemming process the text document is minimized for the chosen words to the stem level. Overall the word is minimized to its lowest form. Porter Stemming algorithm is a technique which removes all the suffixes from the words in English. Porter algorithm is used to remove all extracted words and filter into root or base stem.

Algorithm: A simple rule-based algorithm for stemming.

Input: Document set S.

Output: Stem free document R.

Procedure: The chosen word is broken down to their stem level or the basic root level.

1. Removes all the plurals and suffixes from the words.
2. Move letter y to i and check if there is any other vowel in the stem.
3. Examining the second letter to last one: -ization, -ational, etc.
4. Holds all the suffixes like, -full, -ness etc.
5. Removes –ant, -ence, etc.
6. Eliminates a final –e

b) Pseudo code For Regular Expression Algorithm using Pattern Matching:

By using a regular expression an input text is used to match a pattern. The .NET framework is used for regular expression mechanism that permits pattern matching. A pattern composed of one or more character literals, operators or constructs.

Pseudo code: Pattern matching algorithm to find regular expression anywhere in the text.

Input: Regular Expression R, Text T.

Output: Set of discovered pattern found in a Text T.

```
Match [R, T]
   if R[0] is equal to '^'        // '^'Represent start of text
do
     MatchHere[R, T]
     return 1
     until all text is searched
    return 0
end if
end while
MatchHere[R, T]
    if R[0] is equal to NULL
       return 1
   end if
   if R[1] is equal to *           // '*' represent all characters
     return MatchStar[R[0],R[0]+2,T]
   end if
  if T is not equal to NULL and R[0] equal to '.' Or R[0] is equal to T
     return MatchHere(R+1, T+1)
  end if
  return 0
MatchStar(c, R,T)                 //c is a literal
   do
     if (MatchHere[R,T])
         return 1
   until all the text is searched and text ++ is equal or c is equal to'.'
   end if
   end while
   return 0
```

5. Experiments and Results

In a proposed system, the currently collected dataset is done by collecting all the data of the Belthangadi Hospital Staff members by using required parameters such as staff member's username, email id, date of birth and mobile number. The collected data is real and is handled by using a text file. In text file a large number of data's are maintained containing 50 username, email id, date of birth and mobile number.

The system speed is evaluated by calculating all execution time of proposed methods. The proposed system uses the Stop words and stemming using portal algorithm. Regular expression algorithm for pattern matching and search time to find how many times system takes to search users data in an excel sheet. The existing system uses the discovered patterns and matches it on the dataset and retrieves all the relevant documents.

The below table shows the performance of each categories are mentioned.

Table 1. Execution time of the different instances

Execution Time (ns)	Performance		
	Start Time (ns)	End Time (ns)	Results (ns)
Reading Text Files	473	866	393
Stop Word Removal	286	397	111
Porter Algorithm	314	607	293
Pattern Matching using RE	327	913	586
Search Time	549	551	2.0

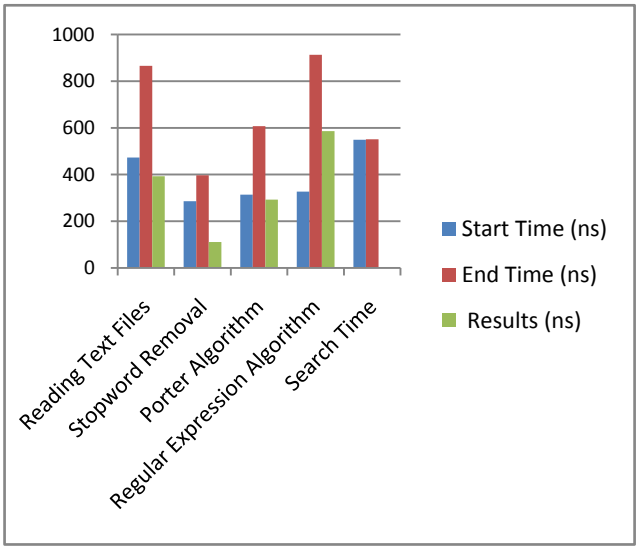

Figure 2. Graph of the proposed system

Now comparing the performance between the porter algorithm using stemming and the regular expression using pattern matching in a proposed system. The execution time between two algorithms is calculated based on their speed. This can be easily done by plotting the graph between two instances as shown:

Table 2. Comparing the performance of two algorithms

Speed (ns)	Iteration 1	Iteration 2	Iteration 3
Porter using stemming	315	303	302
Regular expression using text mine	158	042	122

The graph shown in the above is the comparison between the porter algorithm and regular expression algorithm. The performance is done for 3 iterations to find out which gives the better result.

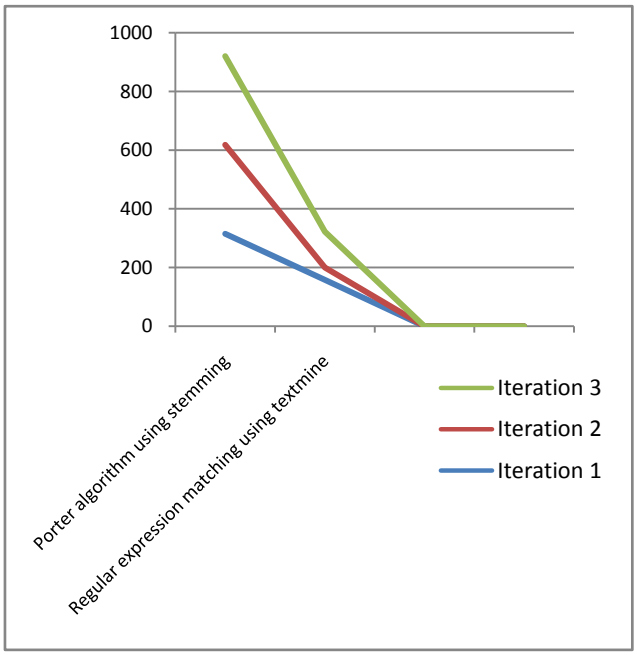

Figure 3. Comparison graph of two algorithms

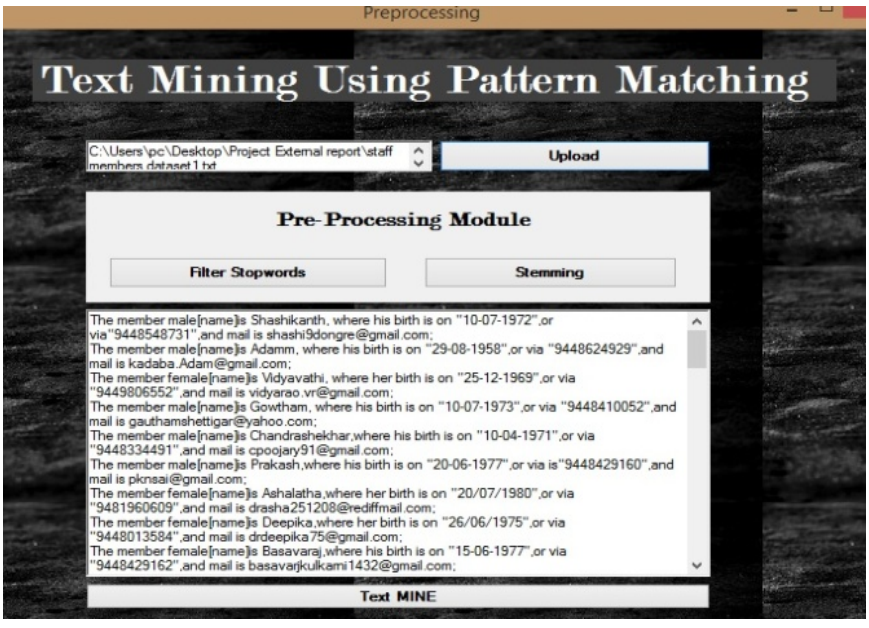

Figure 4. Uploading and Reading a text file

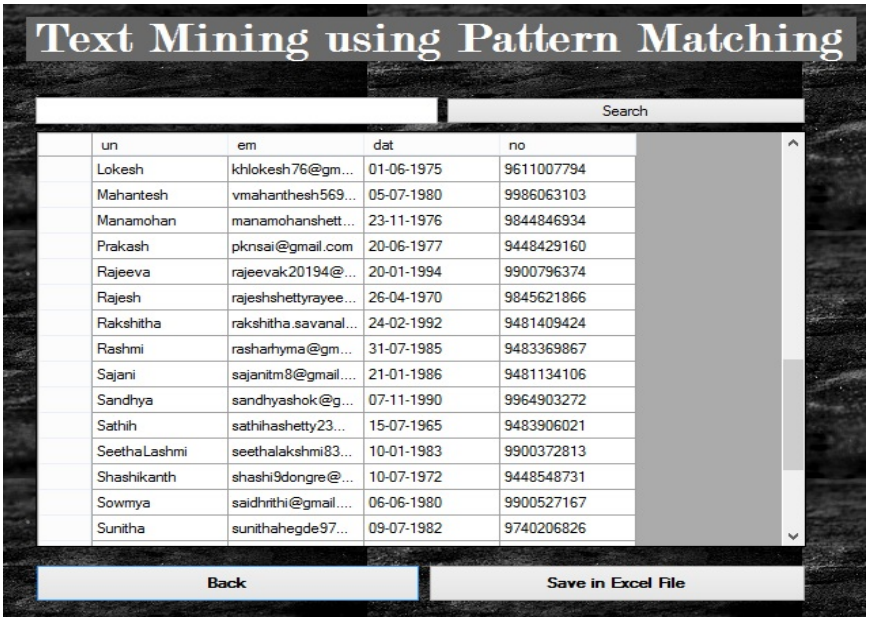

Figure 5. Display of datatypes in tabular form

	A	B	C	D	E	F	G
1	un	em	dat	no			
2	Adamm	kadaba.Adam@gmail.com	29-08-1958	9448624929			
3	Ajai	ajaykallegak@gmail.com	1/5/1988	9480155681			
4	Ajayy	ajaykumar22@gmail.com	22-03-1986	9865633678			
5	Anil	anil.04kumar@yahoo.com	21-04-1985	9535440241			
6	Ashalatha	drasha251208@rediffmail.com	20/07/1980	9481960609			
7	Azmiya	azmiya2009@gmail.com	20-07-1991	9611916892			
8	Basavaraj	basavarjkulkarni1432@gmail.com	15-06-1977	9448429162			
9	Basavaraja	basuiete@gmail.com	24-12-1980	9986073956			
10	Bharathi	bharathishetty2@gmail.com	22-07-1980	9731160124			
11	Chaithra	chaitrakarkala93@yahoo.com	26-11-1993	8971520867			
12	Chandrakala	rkdevasya683@gmail.com	7/4/1985	9480534381			
13	Chandraksha	chandraksha89@yahoo.com	18-10-1989	9480497874			
14	Chandrashekhar	cpoojary91@gmail.com	10/4/1971	9448334491			
15	Deepika	drdeepika75@gmail.com	26/06/1975	9448013584			
16	Devappa	gowdadevappa91@gmail.com	28-07-1991	9880253327			
17	Dharmapal	dharmasuliya@gmail.com	17-01-1980	9743105156			
18	Dinesh	gowdadinesh647@gmail.com	8/3/1989	8762384079			
19	Divya	divyasreeni0@gmail.com	11/6/1986	7090812637			
20	Geetha	geethaumesh8817@gmail.com	23-11-1982	9591948817			
21	Gowtham	gauthamshettigar@yahoo.com	10/7/1973	9448410052			
22	Harini	loshisahani@gmail.com	24-01-1980	8277849127			
23	Indumathi	indumathigowda65@gmail.com	14-11-1972	9632656459			
24	Jayanthi	babachinnu@gmail.com	31-05-1972	9481755483			
25	Josephh	joysvim@gmail.com	7/10/1980	9964903232			

Sheet1 Sheet2 Sheet3

Figure 6. Display of datatypes in an excel sheet

6. Conclusions

The purpose of the project is to study a pattern for each unstructured record. The unstructured data cannot be converted into database directly. By applying all the text mining rules and procedures and coding the unstructured data can be converted into database. Thus, it will be helpful for the users to search the relevant or proper words from the text files.

Future Scope

An algorithm for pattern discovery can be implemented. Categorizing newspaper content for text mining based on the user interest.

REFERENCES

[1] Shivani D Gupta, and B.P. Vasgi, “Implementation of pattern discovery to retrieve relevant document using text mining”, IEEE-2015.

[2] Wai Lam, M. Ruiz and P. Srinivasan, “Automatic text categorization and its application to text retrieval”, IEEE-December 1999.

[3] A. Christy and P. Thambidurai, “Combining information extraction for text mining by using morphological patterns and knowledge discovery”, IEEE-2007.

[4] Raphael Polig, Kubilay Atasu, and Christoph Hagleitner "Token-based dictionary pattern matching for text analytics" IEEE-2013.

[5] Gary K. W. Wong and Simon Y. K. Li, "Academic performance prediction using chance discovery from online discussion forums", IEEE-2016.

[6] V. Aswini, S. K. Lavanya, "Pattern discovery for text mining", IEEE-2014.

[7] Chia-ChuChiang, John Talburt, Ningning Wu, Elizaberth Pierce, Chris Heien, Ebony Gulley, JaMia Moore, "A case study in partial parsing unstructured data", IEEE-2008.

[8] Vaishali Bhujade and N. J. Janwe, "Knowledge discovery in text minimising technique using association rules extraction", IEEE-2011.

[9] Said Gadri and Abdelouahab Moussaoui, "Information Retrieval: A new multilingual stemmer based on a statistical approach", IEEE-2012.

[10] Brindha, Dr. K. Prabha, Dr. S. Sukumaran, "Pattern document weight discovery for text classification method", IEEE- 2014.

[11] Hany Mahgoub, "Mining Association rules from unstructured documents", IJCA- 2006.