Knowledge Discovery in Text Mining using Association Rule Extraction

Manasi Kulkarni
Department of Information Technology
PIIT, New Panvel, India

Sagar Kulkarni
Department of Computer Engineering
PIIT, New Panvel, India

ABSTRACT
Internet and information technology are the platform where huge amount of information is available to use. But searching the exact information for some knowledge is time consuming and results confusion in dealing with it. Retrieving knowledge manually from collection of web documents and database may cause to miss the track for user. Text mining is helpful to user to find accurate information or knowledge discovery and features in the text documents. Thus there is need to develop text mining approach which clearly guides the user about what is important information and what is not, how to deal with important information, how to generate knowledge etc. Knowledge discovery is an increasing field in the research. For a user reading the collection of documents and get some knowledge is time consuming and less effective. There has been a significant improvement in the research related to generating Knowledge Discovery from collection of documents. We propose a method of generating Knowledge Discovery in Text mining using Association Rule Extraction. Using this approach the users are able to find accurate and important knowledge from the collection of web documents which will reduce time for reading all those documents.

General Terms
Association Rule, Text mining

Keywords
Text Mining, Association Rule, knowledge discovery, stemming, term frequency

1. INTRODUCTION
Internet and information technology are the platform where huge amount of information is available to use. Searching the exact information is time consuming and results confusion to deal with it. When user wants to retrieve some information from web, he/she may get both relevant as well as irrelevant data. The information present on internet may be of two types structured or unstructured. . The search for information may be in any language, there is problem that the script contains too many stop words and suffixes due to morphological richness of language. To generate accurate knowledge from collection of web pages, the user needs to find out the relationship between all the keywords present in those web pages. But finding relationship is also becomes complex if the document is not filtered well. The filtration of stop words (as they do not carry meaningful information) and suffix (which may attach with same word but with different forms) is required to find relationship between the keywords. For user filtering that information manually will take much time and may miss the track for what they are searching. While filtering information user may also lose some important information because web information contains large data in which user is actually not interested but it is presented to user when searching something else.Even though the data mining provides tool such as association rule mining, frequent item set mining, pattern mining for effective mining, there are still some issues such as how to handle large number of patterns, how to find relationship between keywords of patterns and how to generate knowledge using keyword relationship. To deal with structured and unstructured information, text mining can be helpful to user to find accurate information or knowledge and features in the text documents [3]. Text mining is an increasing field in the research where one can easily deal with web information to perform different operations on it.

2. LITERATURE SURVEY
Text mining is an increasingly important research field because of the necessity of obtaining knowledge from large number of documents available on the Web. With large amount of information the manual analysis and effective extraction of useful information are not relevant. The author A. K. Ojo, A.B. Adeyemo, 2012, in [1] discusses that the enormous amount of information stored in unstructured texts cannot simply be used for further processing by computers, which typically handle text as simple sequences of character strings. Therefore, specific (pre-) processing methods and algorithms are required in order to extract useful patterns. A well-known rule based stemming algorithm called Porter stemmer is used. In [3] by Vaishali Bhujade and N.J. Janwe has developed system which extracts association rules from web document based on the keyword features. It is also discussed about correlations between features in the text using association rules. The TF-IDF method is used to weight terms to generate association rules. A word is selected as a keyword if it does not appear in a predefined stop-words list. Moreover, association rules are easy to understand and to interpret for an analyst or may be for a normal user. The Giridhar N S, Prema K.V, N .V Subba Reddy in [5] discussed about need for stemming operation before the information retrieval from the documents to increase the effectiveness of retrieval system. How porter stemming algorithm works well is discussed in [6]. According to Ms. Anjali Jivani in [7] the porters stemming algorithm is found to be satisfactory and effective to achieve more relevant information. The results produced while generating association rules using EART system and Apriori-based system are discussed in [10] by Hany Mahgoub. It is proved that there is noticeable difference between execution time of two systems. The EART system in [10], [11] is designed based on keyword features TF-IDF to discover association rules amongst keywords in the documents.

3. PROPOSED SYSTEM
To generate Knowledge discovery we are using a collection of web document as input to the system. The proposed system is divided into three phases Text preprocessing phase, Association rule mining phase and Knowledge discovery

phase. In text preprocessing phase we first tokenize the input document in to tokens. Then the tokens are filtered by removing stop word as they do not carry any meaningful information. Normally token contains many suffixes and it is required to remove all the suffixes to achieve better result in knowledge discovery. Then the tokens are indexed using TF-IDF (term frequency – inverse document frequency) values. The Association rule mining phase generates association rules based on weighting scheme TF-IDF.That is depending on the users requirement the high frequency keywords are selected to generate association rules. The last phase is to generate Knowledge discovery using association rules [3]. The system can be useful for all information retrieval schemes such as Text mining, Text summarization, Categorization etc. when one wants to generate knowledge from a collection of web documents.

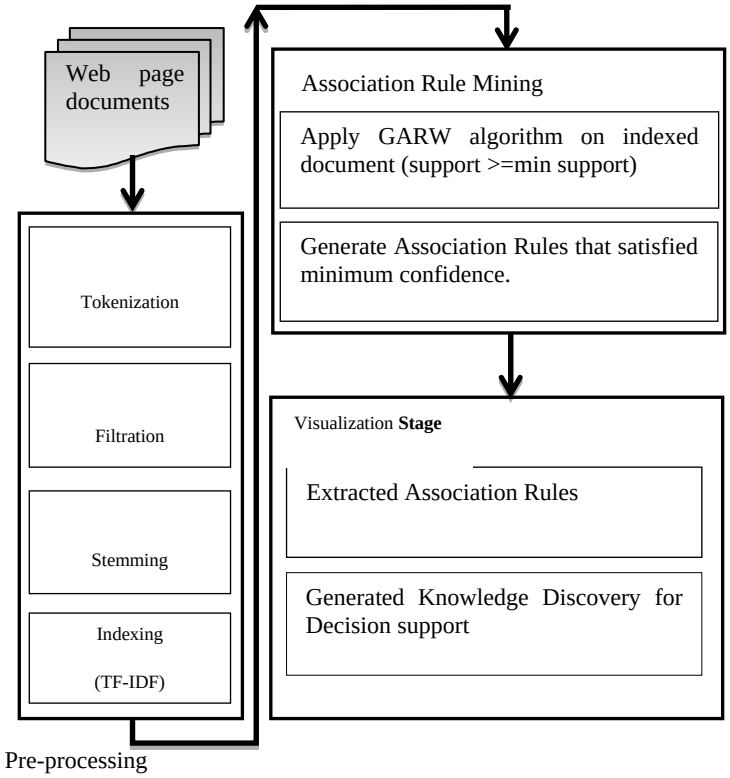


Figure1. Block diagram for Knowledge discovery using Association Rules.

As shown in Figure1the architecture to generate knowledge discovery consists of three stages:

a) Text Pre-processing stage

b) Association rule mining stage

c) Visualization stage.

3.1 Text Pre-processing phase

For mining large document collections, it is necessary to pre-process the input documents and store the information in a data structure, which is more appropriate for further processing than a plain text file. Text pre-processing classically means tokenization and filtration of keywords, stemming, indexing the keywords etc.

Tokenization

Tokenization is the process of splitting the text into words or terms. This phase plays very important role in generating association rule. The main objective is to convert unstructured document in to structured document. Normally web page contains information in unstructured format. This create problem for accurate and relevant text mining. Thus it is required to convert those unstructured information to structured format. We have use MySQL data structure to perform this operation. The textual information is scanned by the system and arranged in to table in row and column wise. Once the web information is arranged in to tabular i.e. structured format, it is easy to perform further operations on it such as filtration stemming etc. The system is designed so as to form only valid tokens by detecting and removing the irrelevant information such as special characters, parentheses, commas, etc. If we allow this irrelevant thing in the tokens then unnecessarily it will consume memory and time to process. Thus it is better to ignore this irrelevant information while reading the web document information.

Filtration of Keywords

The search for information may be in any language and user is expecting to have relevant output from the search. In information retrieval there is problem that the input script contains too many stop words and suffixes due to morphological richness of language. To generate accurate knowledge from collection of web pages, the user needs to find out the relationship between all the keywords but this task may become inefficient if we do not remove stop words and suffixes from the input documents[3]. Finding relationship is also becomes complex if the document is not filtered well. The filtration of stop words (as they do not carry meaningful information) and suffixes (which may be attached with same word but with different forms) are required to find relationship between the keywords. For a user, filtering the large input information manually will take much time and may miss the track for what they are searching. While

filtering information user may also lose some important information because web is platform where large data is presented to user in which he is actually not interested. That is user usually get some extra or irrelevant information while searching something else. The web document mainly contains information in unstructured format. These web documents are firstly needs to be arranged into structured format. Filtering the web documents are required to remove unimportant keywords from the document such as stop words [1]. Stop words do not carry any meaningful information. Therefore the stop words are discarded and more important or highly relevant words are used for the further processing. We have used a list of stop words which are frequently used in English document script. System checks the keywords to analyze whether it is stop word or not. If stop word is found then it is eliminated from the content.

Stemming

A stemming is a process in which the variant forms of same word are reduced to a common form. It is important to appreciate that we use stemming with the intention of improving the performance of knowledge discovery systems. There are some reasons that why we are using stemming.

Why Stemming

1. Suffixes changes meaning of term even main root word is same.
2. Ambiguous association rules will be generated if suffixes are present.
3. Suffixes makes data complex and occupy extra memory.

Word stemming is an important feature supported by indexing and search systems. Indexing and searching are in turn part of Text Mining applications, Natural Language processing (NLP) systems and Information Retrieval (IR) systems. Stemming is usually done by removing any attached suffixes from index terms before the actual assignment of the term to the index. The stemming technique is applied on the target data set to reduce the size of data set which will increase the effectiveness of IR System. A stemming algorithm is a process of linguistic normalization, in which the variant forms of a word are reduced to a common form [6]. It is important to appreciate that we use stemming with the intention of improving the performance of IR systems. Number of stemming Algorithms, or stemmers, have been developed, which attempt to reduce a word to its stem or root form. Thus, the key terms of document is represented by stems rather than by the original words. Stemming also reduces the dictionary size. There are several types of stemming algorithms which differ in respect to performance and accuracy. With English script the variable part is the ending or suffix. Taking these endings off is called suffix stripping or stemming, and the residual part is called the stem of word or root word. There are different stemming algorithms like Porter Lovins, Paise, Krovtz[5]. Out of this algorithm Porter algorithm found to be best to perform stemming. With existing porter algorithm there is problem that it does not gives correct root word always. The objective is to replace all the matched suffixes from the keywords with replacement character or word.

The Table1 shows some modified or added rules in existing porter algorithm. So by changing the rules we made in the existing porter algorithm so that the efficiency of generating correct stem words will get increased. We have compared the output produced by existing porter and our modified porter with large set of keywords. After comparing, we observed that there is huge difference in root words generated by using both algorithms. The modified porter works well and gives efficiency of 94.53% in producing correct root word.

Table1. Modification in Porter Stemming Rules

Sr. No.	Rule added/Changed	Example	Output by original Porter	Output by modified Porter
1	OUS -> OUS	Obvious	Obviou	Obvious
2	ICAL -> ICAL	Medical	Medice	Medical
3	IAL ->E	Official	Offici	Office
4	ANT->ANT	Servant	Serv	Servant
5	DENT-DENT	Accident	Accid	Accident
6	ATION-> ATE	Corporation	Corpor	Corporate
7	E -> E	Village	Villag	Village
After removal of ‘ing’ character use replacement rule:				
8	S -> SE	Exposing	Expos	Expose
After removal of ‘ed’ character use replacement rule:				
9	R -> RE	Declared	Declar	Declare
10	D -> DE	Provided	Provid	Provide
11	G -> GE	Emerged	Emerg	Emerge
12	V -> VE	Removed	Remov	Remove
13	FI -> FY	Identified	Identifi	Identify

Indexing

The weighting scheme TF-IDF (Term Frequency, Inverse Document Frequency) is used to assign weights to distinguish terms in a document. The term frequency is the count that represents how many times keyword(x) has occurred in the document. The inverse document frequency is the count that represents the total number of documents that contains the keyword(x) at least once. We have used TF-IDF method to get relevant knowledge discovery only for frequently occurred keywords in the input documents. Using this weighting scheme we select user expected higher frequency keywords for generation of association rules. With Apriori algorithm the main disadvantage is that it considers all the keywords for generating association rules without knowing importance of those keywords. Due to this there will be large collection of association rules which will become tedious and time consuming for generating it. Also the knowledge discovery generated by these association rules will not be so meaningful, as it considers all the keywords which are of lower weight.

Term Frequency

The number of times a term occurs in collection of document is called its 'term frequency'. The Term Frequency is defined as:

$$(tf)_{i,j} = \sum_{i=j=1}^{n} (Nt_i, d_j)$$

Where Nt_i, d_j denotes the number the term t_i occurs in document d_j.

Inverse Document Frequency

The number of documents in collection where the considered term occurs at least once is called as inverse document frequency [4]. The Inverse Document Frequency is defined as:

$$(idf)_{i,j} = \sum_{i=j=1}^{n} \log\left(\frac{|C|}{Nt_i}\right)$$

Where Nt_i denotes the number of documents in collection C in which t_i occurs at least once & $|C|$ denotes the total number of documents in collection.

The aim of this step is to identify and filter the keywords that may not be of interest in the context of the whole document collection because they do not occur frequently enough. Depending on the user's requirement, the top 'N' tokens are taken as the final set of keywords to be used in the knowledge discovery phase.

3.2 Association Rule Mining phase

Association rule is of the IF-THEN structure, but it can predict attribute combination, and they are not intended to be used together as a set. For each rule IF antecedent THEN consequent we count its support and confidence matches with user specified values[10]. Support is the probability that a randomly selected instance will fulfill both the antecedent and consequent, and confidence is the conditional probability that a randomly selected instance will fulfill the consequent given that the instance fulfill the antecedent [3]. In our developed system we have generated only those association rules which satisfy criteria such as support, confidence, and TF-IDF value of keywords. The algorithm called as GARW (Generating Association Rule using Weighting Scheme) is found to be better than that of conventional Apriori algorithm. The GARW algorithm works in similar way of Apriori but with some additional steps to resolve problem of Apriori and to generate relevant association rules [3].

GARW Algorithm:

1. Let 'N' denote the number of top keywords that satisfy the threshold weight value.
2. Store the top 'N' keywords in indexed file along with their frequencies (TF value) in all documents.
3. Use minimum support and confidence values given by user for association rule generation.
4. Scan the indexed file and find all keywords that satisfy the threshold minimum support. These keywords are called large frequent 1-keywordset 'L'.
5. In k ≥ 2, the candidate keywords Ck of size k are generated from large frequent keyword sets.
6. Scan index file and compute frequency of candidate keyword sets Ck that generated in step 5.
7. Compare the frequencies of candidate keyword sets with minimum support.
8. Large frequent k-keyword sets, which satisfy the minimum support, is found from step 6.
9. For each frequent keyword set, find all the association rules that satisfy the threshold minimum confidence.

The GARW algorithm does not make multiple scanning on the original documents. It scans large frequent keyword sets to generate association rules which satisfy the user given support and confidence.

4. KNOWLEDGE DISCOVERY PHASE

We have developed a system which enables the user to generate knowledge from a collection of web page documents. The Knowledge discovery using Association Rule extraction can be very useful for the newspaper sectors where one can discover the knowledge from a collection of documents without reading all the documents manually. The analysis of information manually is time consuming and not feasible. This is because, for a user it very hard and tedious to remember all important information throughout the collection and get some knowledge from it [12]. This may cause in making some decision using discovered knowledge. Also there is very less guarantee that the knowledge discovery generated manually contains enough details to support for decision making.

We have used association rules generated form web documents related to diseases such as Cholera and Dengue. Now to generate knowledge discovery from the large number of association rules, we used training system which contains details about diseases, locations, victims, reasons of spreading diseases and impacts of that diseases etc[3]. All the association rules are trained using this system to know the relationship of rule contains (Antecedent and Precedent). When user gives query to discover knowledge, then the query is compared throughout the association rules to extract the important keywords from it. Thus using extracted keywords from association rules depending on the query supplied, it will be easy for user to get some knowledge from it and to support decisions making.

5. RESULTS AND DISCUSSION

We have performed the experiments on newspaper documents selected from archival collection of Times of India. The input documents are selected for diseases such as Cholera and Dengue. When we applied our developed system to these documents then the produced knowledge discovery is found to be relevant for the user to make further decisions.

Tokenization:

This phase is basically used to convert unstructured documents in to structured format. Web documents are normally unstructured or semi structured. Thus for text mining it is first step to convert those documents in to structured format. The Figure2 shows the output after tokenization process.

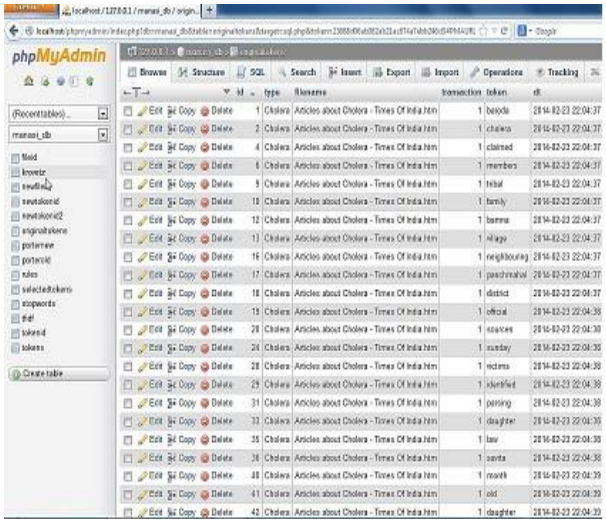


Figure2. Tokenization of input web document

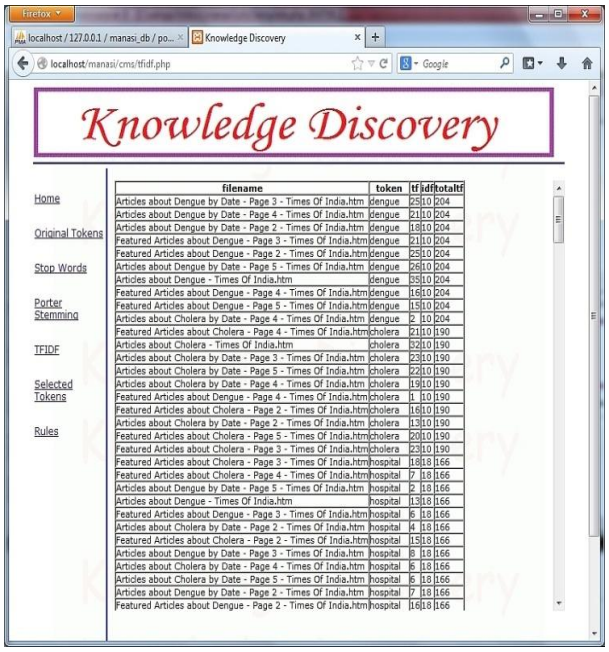
Filtration:

The filtration of stop words is required to perform as they do not carry any meaningful informant. Also processing stop words is time consuming. To have better keyword relationship we decided to remove frequently occurring stop words. The stop words used are taken from internet.

Stemming is performed using our modified porter algorithm.

Indexing:

The TF-IDF method is used to calculate the term frequency of keywords. The keywords are then arranged in descending order. The value of keyword is total-TF and computed by adding all the term frequencies of that keyword. Figure3 shows some of the TF-IDF values of the keywords.



filename	token	tf	idf	totaltf
Articles about Dengue by Date - Page 3 - Times Of India.htm	dengue	25	10	204
Articles about Dengue by Date - Page 4 - Times Of India.htm	dengue	21	10	204
Articles about Dengue by Date - Page 2 - Times Of India.htm	dengue	18	10	204
Featured Articles about Dengue - Page 3 - Times Of India.htm	dengue	21	10	204
Featured Articles about Dengue - Page 2 - Times Of India.htm	dengue	25	10	204
Articles about Dengue by Date - Page 5 - Times Of India.htm	dengue	26	10	204
Articles about Dengue - Times Of India.htm	dengue	35	10	204
Featured Articles about Dengue - Page 4 - Times Of India.htm	dengue	16	10	204
Featured Articles about Dengue - Page 5 - Times Of India.htm	dengue	15	10	204
Articles about Cholera by Date - Page 4 - Times Of India.htm	dengue	2	10	204
Featured Articles about Cholera - Page 4 - Times Of India.htm	cholera	21	10	190
Articles about Cholera - Times Of India.htm	cholera	32	10	190
Articles about Cholera by Date - Page 3 - Times Of India.htm	cholera	23	10	190
Articles about Cholera by Date - Page 5 - Times Of India.htm	cholera	22	10	190
Articles about Cholera by Date - Page 4 - Times Of India.htm	cholera	19	10	190
Featured Articles about Dengue - Page 4 - Times Of India.htm	cholera	1	10	190
Featured Articles about Cholera - Page 2 - Times Of India.htm	cholera	16	10	190
Articles about Cholera by Date - Page 2 - Times Of India.htm	cholera	13	10	190
Featured Articles about Cholera - Page 5 - Times Of India.htm	cholera	20	10	190
Featured Articles about Cholera - Page 3 - Times Of India.htm	cholera	23	10	190
Featured Articles about Cholera - Page 3 - Times Of India.htm	hospital	18	18	166
Featured Articles about Cholera - Page 4 - Times Of India.htm	hospital	7	18	166
Articles about Dengue by Date - Page 5 - Times Of India.htm	hospital	2	18	166
Articles about Dengue - Times Of India.htm	hospital	13	18	166
Featured Articles about Dengue - Page 3 - Times Of India.htm	hospital	6	18	166
Articles about Cholera by Date - Page 2 - Times Of India.htm	hospital	4	18	166
Featured Articles about Cholera - Page 2 - Times Of India.htm	hospital	15	18	166
Articles about Dengue by Date - Page 3 - Times Of India.htm	hospital	8	18	166
Articles about Cholera by Date - Page 4 - Times Of India.htm	hospital	6	18	166
Articles about Cholera by Date - Page 5 - Times Of India.htm	hospital	6	18	166
Articles about Dengue by Date - Page 2 - Times Of India.htm	hospital	7	18	166
Featured Articles about Dengue - Page 2 - Times Of India.htm	hospital	16	18	166

Figure3. TF-IDF values of keywords after Indexing

Comparison of different Stemming Algorithms:

	Original Tokens	Root Word	PorterNew	Changed	PoterOld	Changed	Lovin	Changed	Paise	Changed	Krovetz	Changed
2	baroda	baroda	baroda	TRUE	baroda	TRUE	barod	FALSE	barod	FALSE	baroda	TRUE
3	cholera	cholera	cholera	TRUE	cholera	TRUE	choler	FALSE	choler	FALSE	cholera	TRUE
4	claimed	claim	claim	TRUE	claim	TRUE	claim	TRUE	claim	TRUE	claim	TRUE
5	family	family	family	TRUE	famili	FALSE	fam	FALSE	famy	FALSE	famili	FALSE
6	members	member	member	TRUE	member	TRUE	member	TRUE	memb	FALSE	member	TRUE
7	neighbouring	neighbour	neighbour	TRUE	neighbour	TRUE	neighbour	TRUE	neighbo	FALSE	neighbouring	FALSE
8	district	district	district	TRUE	district	TRUE	district	TRUE	district	TRUE	district	TRUE
9	official	office	office	TRUE	offici	FALSE	offic	FALSE	off	FALSE	offici	FALSE
10	sources	source	source	TRUE	sourc	FALSE	sourc	FALSE	sourc	FALSE	sourc	FALSE
11	:	:	:	:	:	:	:	:	:	:	:	:
12	:	:	:	:	:	:	:	:	:	:	:	:
13	:	:	:	:	:	:	:	:	:	:	:	:
14												
8620	private	private	private	TRUE	privat	FALSE	priv	FALSE	priv	FALSE	privat	FALSE
8621	suffering	suffer	suffere	FALSE	suffer	TRUE	suffer	TRUE	suff	FALSE	suffering	FALSE
8622	positive	positive	positive	TRUE	posit	FALSE	posit	FALSE	posit	FALSE	positiv	FALSE
8623	infected	infect	infect	TRUE	infect	TRUE	infect	TRUE	infect	TRUE	infect	TRUE
8624	dengue	dengue	dengue	TRUE	dengu	FALSE	dengu	FALSE	dengu	FALSE	dengu	FALSE
8625												
8626	Total Correct words Found:			8139		5601		3615		3531		5546
8627												
8628	Percentage			94.43		64.98		41.94		40.97		64.35
8629												
8630	Total Incorrect words Found:			480		3018		5004		5088		3073

Figure4. Comparison of different Stemming Algorithms

Here we have compared few stemming algorithms with their output. The algorithms such as existing Porter, Modified Porter, Lovins, Paise/Husk and Krovetz algorithm are taken for comparison. The result is as shown in Figure4. The comparison graph of stemming algorithms is shown in Figure5. The comparison shows that the output produced by modified porter algorithm is better than other algorithms. Thus we have used this modified porter stemming in our experiments.

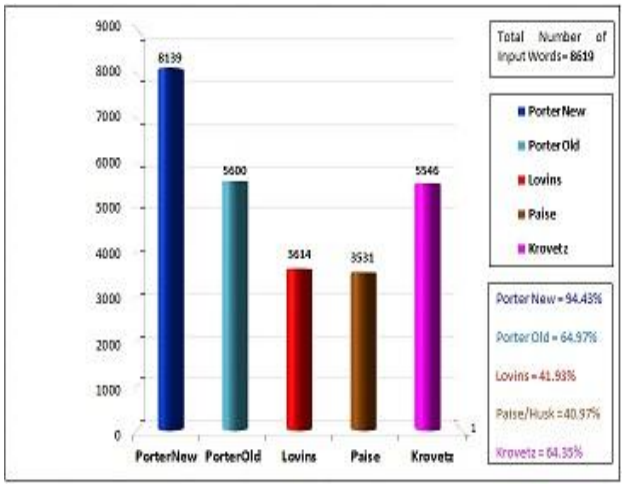


Figure5. Graphical representation of stemming result

Association Rule Generation:

To generator meaningful association rules and to have better keyword relationship it is required that to select top 'N' keywords from Indexed file which has higher weight (term frequency) value. Here we have selected top 400 keywords for association rule generation with support is 15 and confidence is 25. The Figure6 shows the number of association rules generated for given input values.

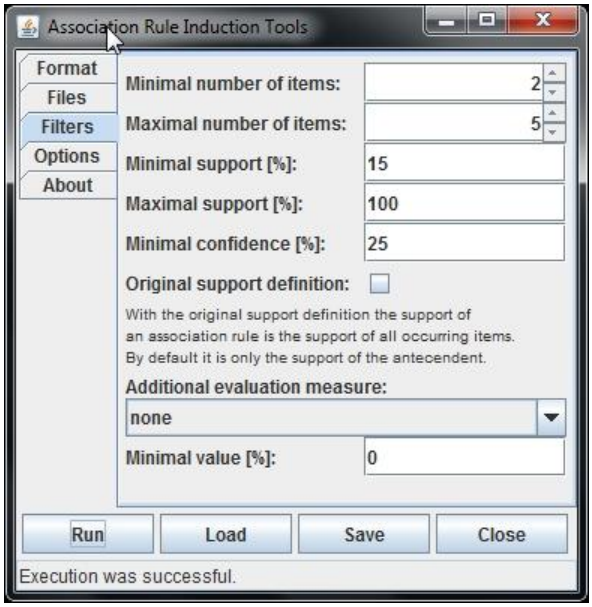


Figure6. Association Rule Generation Tool

Knowledge Discovery using Association Rules:

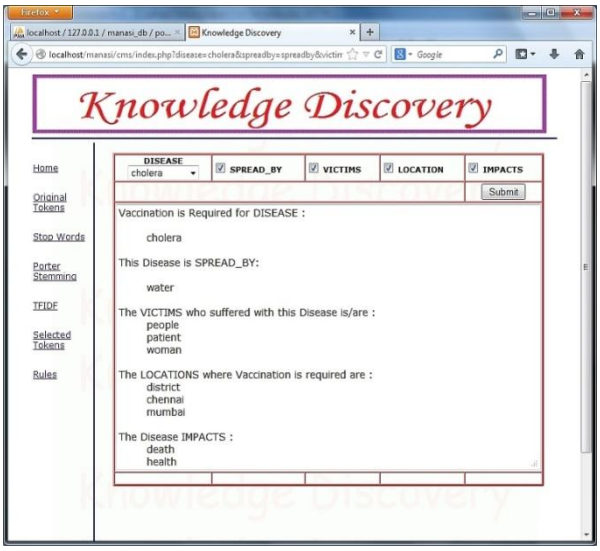


Figure7. Knowledge Discovery using association rule extraction

6. CONCLUSION

The proposed system has presented a text mining technique for extracting association rules from collection of web documents. The system accepts input as web documents which are mainly in unstructured format and transforms them into the structured form. Using association rules one can recognize the relationship between keyword with other keywords and generates knowledge discovery of different item-sets from unstructured document. This work can be used for many information retrieval applications such as Decision support system, expertise location, Web usage mining, intrusion detection, market basket analysis, text Summarization, information retrieval based on ontology etc. Generating knowledge discovery is an important application of association rules. It gives accurate information about keyword or pattern relationships and their occurrences.

7. REFERENCES

[1] A. K. Ojo, A.B. Adeyemo, March-2012 "Framework for Knowledge Discovery from Journal Articles Using Text Mining Techniques", African Journal of Computing & ICT, Vol 5. No.2, ISSN: 2006-1781, Page 35-44.

[2] Vishwadeepak Singh Baghela, Dr. S.P.Tripathi, May 2012, "Text mining Approaches To Extract Interesting Association Rules from Text Document", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 3, ISSN (Online): 1694-0814, page: 545-552

[3] Vaishali Bhujade, N.J. Janwe, 2011, "Knowledge Discovery in Text Mining Technique Using Association Rules Extraction", International Conference on Computational Intelligence and Communication Systems, 978-0-7695-4587-5, IEEE DOI-10.1109/CICN.2011.104, page: 498-502

[4] Vaishali Bhujade, N. J. Janwe, Chhaya Meshram, July-Aug 2011 "Discriminative Features Selection in Text Mining Using TF-IDF Scheme", International Journal of Computer Trends and Technology, ISSN: 2231-2803, page: 196-198

[5] Giridhar N S, Prema K.V, N .V Subba Reddy, JAN-JUN-2011, "A Prospective Study of Stemming Algorithms for Web Text Mining", GANPAT UNIVERSITY JOURNAL OF ENGINEERING & TECHNOLOGY (GNUJET), VOL.-1, ISSUE-1, page: 28-34

[6] N. Sandhya, Y. Sri Lalitha, V.Sowmya, Dr. K. Anuradha, Dr. A. Govardhan, September 2011 "Analysis of Stemming Algorithm for Text Clustering", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 1, ISSN (Online): 1694-0814, page: 352-359

[7] Anjali Ganesh Jivani, NOV-DEC 2011 "A Comparative Study of Stemming Algorithms", International Journal in Computer Technology (IJCTA), Appl., Vol 2 (6), ISSN:2229-6093, page: 1930-1938

[8] Hemlata Sahu, Shalini Shrma, Seema Gondhalakar, 2011 " A Brief Overview on Data Mining Survey", International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 1, Issue 3, ISSN 2249-6343, page: 114-121

[9] Atika Mustafa, Ali Akbar, and Ahmer Sultan, April 2009 "Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization", International Journal of Multimedia and Ubiquitous Engineering, Vol. 4, No. 2, page: 183-188

[10] Hany Mahgoub, Dietmar Rösner, Nabil Ismail and Fawzy Torkey, 2008 "A Text Mining Technique Using Association Rules Extraction", International Journal of Information and Mathematical Sciences 4:1, page: 21-28

[11] Hany Mahgoub, 2008 "Mining Association Rules from Unstructured Documents" World Academy of Science, Engineering and Technology 20, page: 938-943

[12] Fatudimu I.T, Musa A.G, Ayo C.K, Sofoluwe A. B, 2008 "Knowledge Discovery in Online Repositories: A Text Mining Approach", European Journal of Scientific Research ISSN 1450-216X Vol.22 No.2, page: 241-250