Developing an Expert IR System from Multidimensional Dataset

Anagha Chaudhari
Assistant Professor,
Department of IT
Pimpri Chinchwad College of Engineering
Pune, India

Amitabh Mudiraj
Sr. Software Engineer

Tweddle Group
Clinton Township, USA

Swati Shinde, Ph.D
Assistant Professor,
Department of IT
Pimpri Chinchwad College of Engineering
Pune, India

ABSTRACT

Now-a-days due to increase in the availability of computing facilities, large amount of data in electronic form is been generated. The data generated is to be analyzed in order to maximize the benefit of intelligent decision making. Text categorization is an important and extensively studied problem in machine learning. The basic phases in the text categorization include preprocessing features like removing stop words from documents and applying TF-IDF is used which results into increase efficiency and deletion of irrelevant data from huge dataset. Application of TF-IDF algorithm on dataset gives weight for each word which summarized by Weight matrix. Preprocessing reduces the size of dataset which ultimately improves the performance of search engine. After that, index is generated from dataset. Index contains term with its occurrence in file and also its location in file. This paper discusses the implication of efficient Information Retrieval system for text-based data using clustering approaches.

General Terms

Data Mining, Information Retrieval.

Keywords

Information retrieval, stop words, TF IDF, text based clustering, fitness functions

1. INTRODUCTION

The ability to search for a user need within a collection of data which is relevant to the users query is called Information Retrieval. In data mining field size of database is becoming huge and complex. To retrieve appropriate information from such a database is time consuming. Hence to minimize time and enhance the capabilities of IR system we are going to apply some data mining approaches to database with clustering so as to facilitate efficient IR system. The different measures for evaluating the performance of information retrieval systems have been proposed. To implement the measures it requires a collection of documents (dataset) and a query. All the common measures described here assume a ground truth notion of relevancy, every document is classified to be either relevant or non-relevant to a particular query. In practice the queries may be ill-posed and there might be different shades of relevancy. The importance of preprocessing is emphasized by the fact that the quantity of training data grows exponentially with the dimension of the input space. It has already been proven that the time spent on preprocessing can take from50% up to 80% of the entire classification process [1], which clearly proves the importance of preprocessing in text classification process. The first step during preprocessing is to remove these Stop words [2], many of the most frequently used words in English are useless in Information Retrieval (IR) and text mining. These words are called 'Stop words'. Stemming techniques are used to find out root/stem of a word. For example, the words, user, users, used, using all can be stemmed to the word 'USE'. In the present work, the Porter Stemmer algorithm [3], which is the most commonly used algorithm TF-IDF is one of the most commonly used term weighting algorithms in today’s information retrieval systems. Two parts of the weighting were proposed by Gerard Salton [4] and Karen Spärck Jones [5] respectively.

TF, the term frequency, also called Local Term Weight, is defined as the number of times a term in question occurs in a document. IDF, the inverse document frequency is also called Global Term Weight and is based on counting the number of documents in the collection being searched that are indexed by the term.

The product of TF and IDF also known as TF-IDF, is used as an indicator of the importance of a term in representing a document[6].The sequential pattern mining in document databases plays an important role, because it allows to identify valid, novel, potentially useful and ultimately understandable patterns[7]. Many multiple-scan approaches are proposed for mining sequential Patterns. Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto has proposed a novel sequential pattern mining method, called Prefix Span [8] (i.e., Prefix-projected Sequential Pattern Mining), which explores prefix projection in sequential pattern mining. Prefix Span mines the complete set of patterns but greatly reduces the efforts of candidate subsequence generation. A typical Apriori like method such as GSP [9] adopts a multiple-pass, candidate generation and test approach in sequential pattern mining.In clustering, the algorithms perform an initial division of the data in the clusters and then move the objects from one cluster to another based on the optimization of a pre-defined criterion or objective function [10]. The frequent representative algorithms that use these techniques are k-means, k-medoids and expectation maximization. The k-means algorithm is the most popular because it is easy to implement and its time complexity is O(n), where n is the number of patterns or records, but it has serious disadvantages it is sensitive to outliers, it is sensitive to the selection of the initial centroids, it requires a prior definition of the number of clusters[11].

2. SYSTEM ARCHITECTURE

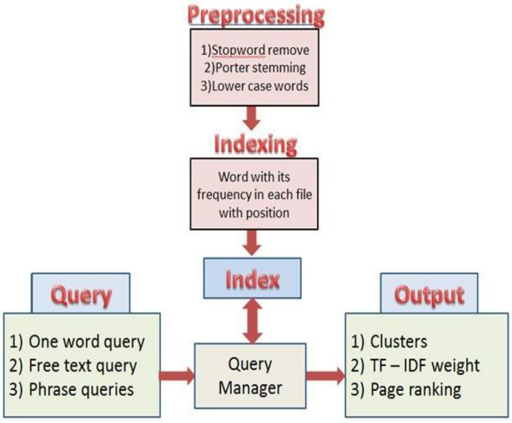


Fig 1: System Architecture

1. Module 1 : Preprocessing
2. Module 2 : Generating Inverted Index File
3. Module 3 : Clustering
4. Module 4 : Fitness Function
5. Module 5 : Information Retrieval

1. Module 1 : Preprocessing

Given data may be incomplete, inconsistent & having some errors and also some irrelevant data. For example there are many stop words in a given text file at any given instance, these words increase the database size and also slows the further processing of data mining techniques [1]. The preprocessing techniques used in this experiment are stop word removal and term frequency inverse document frequency (TF IDF).

A] Stop-Word Removal

Many of the most frequently used words in English are useless in Information Retrieval (IR) and text mining. These words are called 'Stop words'. Stop-words, which are language-specific functional words, are frequent words that carry no information (i.e., pronouns, prepositions, conjunctions). In English language, there are about 400-500 Stop words. Examples of such words include 'the', 'of', 'and', 'to'. The first step during preprocessing is to remove these Stop words, which has proven as very important. The present work uses the SMART stop word list.

B] TF-IDF

Tf-idf stands for term frequency-inverse document frequency, and the tf-idf weight is a weight often used in information retrieval and text mining [16]. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus.

2. Module 2 : Generating Inverted Index File

Inverted index is a word-oriented mechanism for indexing a text collection to speed up the searching task The inverted index structure is composed of two elements: the vocabulary and the occurrences [17]. The vocabulary is the set of all different words in the text. For For each word in the vocabulary, the index stores the documents which contain that word (inverted index).

3. Module 3: Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters) [12]. Clustering is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions. The most popular clustering technique is k-means clustering. K-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations.

4. Module 4 : Fitness Function

Fitness function is a performance measure or reward function,which evaluates how each solution, is good. The fitness function plays a very important role in obtaining the best solutions within a large search space. Lead to a better retrieval performance.

Without any prior knowledge of those relationships [13]. The k-means algorithm is one of the simplest clustering techniques and it is commonly used in data mining, biometrics and related fields [15].

5. Module 5 : Information Retrieval

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information need [4].

Precision is the proportion of retrieved documents that are actually relevant and Recall is the proportion of relevant documents that are actually retrieved.

3. MATHEMATICAL MODEL

A] Procedure for Preprocessing

Algorithm: Preprocessing

```
Input: Dataset (Contains n no. of text files)

Step 1: for each document in Dataset remove stop-
        words.

Step 2: for each word in the dataset calculate TF_IDF
        and store the result in a weight  matrix.

Step 3: for each element in weight matrix set the
        threshold 'c' and calculate document
        Frequency (DF)
Step 4: for each term If DF< c then remove the term
        along with its weight from weight matrix.
```

Output: Preprocessed dataset.

Term Frequency: It measures how frequently a term occurs in a document [1].

$$TF(t) = \frac{(\text{Number of times term t appears in a document})}{(\text{Total number of terms in the document})}$$

Document Frequency: DF of term is defined by the number of documents in which a term appears.

Inverse Document Frequency: It measures how important a term is [1]. We need to weigh down the frequent terms while scale up the rare ones, by computing the following:

$$\text{IDF}(t) = \log e \frac{(\text{Total number of documents})}{(\text{Number of documents with term t in it})}$$

B] Inverted Index File

Given a query, we use the index to return the list of documents relevant for this query. The inverted index contains mappings from terms (words) to the documents that those terms appear in. Each vocabulary term is a key in the index whose value is its postings list.

C] K-means flowchart for clustering

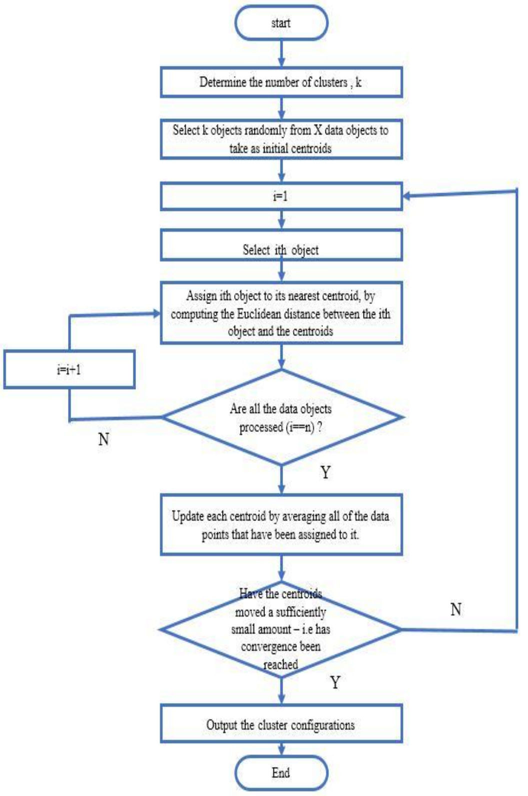

start
Determine the number of clusters , k
Select k objects randomly from X data objects to take as initial centroids
i=1
Select ith object
Assign ith object to its nearest centroid, by computing the Euclidean distance between the ith object and the centroids
i=i+1
Are all the data objects processed (i==n) ?
N
Y
Update each centroid by averaging all of the data points that have been assigned to it.
Have the centroids moved a sufficiently small amount – i.e has convergence been reached
N
Y
Output the cluster configurations
End

Fig.2: K-means Flowchart

D] Fitness functions

Similarity function is a performance measure or reward function, which evaluates how each solution, is good. They are used to improve the efficiency of the Information Retrieval system [14]. Here we have used 2 similarity functions:

Cosine Fitness Function:

It is used to measure the angle between query and document. Lower the angle higher the degree of similarity and higher the angle lower the degree of similarity between query and document [14].

$$\frac{\sum_{i=1}^{t} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{t} x_i^2 \cdot \sum_{i=1}^{t} y_i^2}}$$

where x is query vector and y is a document vector.

Jaccard Coefficient

The Jaccard coefficient, measures similarity as the intersection divided by the union of the objects. For text document, the Jaccard coefficient compares the sum weight of shared terms to the sum weight of terms that are present in either of the two documents but are not the shared terms. The formal definition is:

J= (A intersection B) / (A∪B)

where A and B represents the documents and the value of coefficient ranges between 0 and 1.

E]IR Performance Measure

Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage[18].

$$Precision = \frac{|\{Relevant\ Docs\} \cap \{Retrieved\ Docs\}|}{|\{Retrieved\ Documents\}|}$$

Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage [18].

$$Recall = \frac{|\{Relevant\ Docs\} \cap \{Retrieved\ Docs\}|}{|\{Relevant\ Documents\}|}$$

4. EXPERIMENTATION AND RESULTS

Dataset used for this experiment was created from 2 reference books namely Management Information System and Information System .Text files are created for each topic from book. Total 137 documents were created. Table 1 shows result after applying Stop Word Removal preprocessing technique to Initial dataset. There are total 6 2,978 words in dataset. Total 26,676 Stop-Words were removed which result into reduction of dataset size in 26.58%. The size of dataset is showed in Table 2 [18].

Table 1. Stop word Removal

Total words in dataset	Removed Stopwords from dataset
62,978	26,676

Table 2. Dataset Size

Initial Dataset size	Dataset size after removing stopwords
415,403 bytes	304,984 bytes

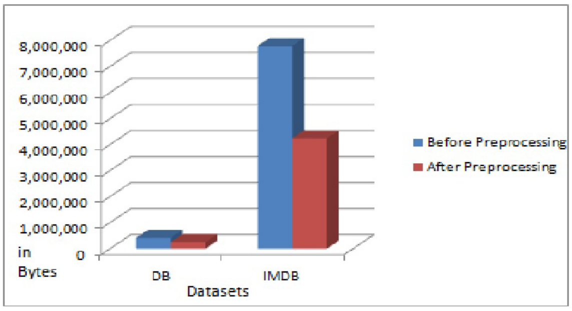
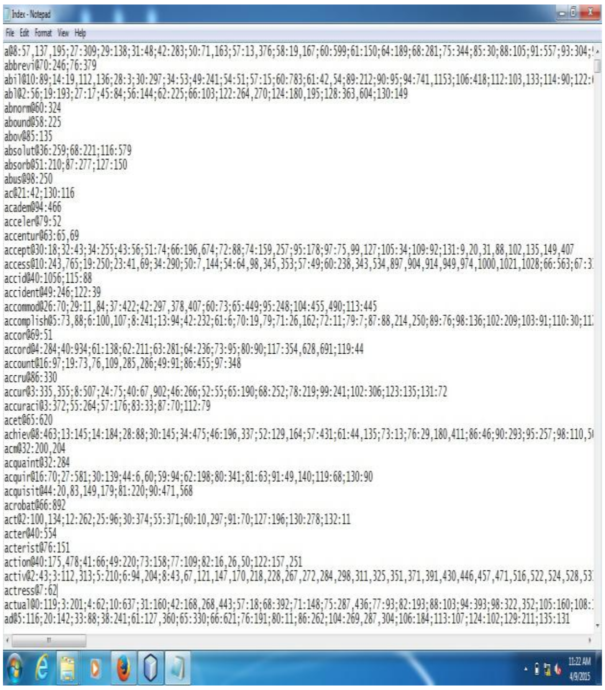


Fig.3: Preprocessing Result

The above graph in Figure 3 indicates approximately 40% reduction in the size of dataset.

The preprocessed dataset is achieved through term by document matrix. The figure given below represents the generated Inverted Index File which contains term with its occurrence and also its location in file. This index file can be further used for developing clusters along with fitness functions and it facilitates easy information retrieval.



```
a@8:57,137,195;27:309;29:138;31:48;42:283;50:71,163;57:13,376;58:19,167;60:599;61:150;64:189;68:281;75:344;85:30;88:105;91:557;93:304;
abbrevi@70:246;76:379
abil@10:89;14:19,112,136;28:3;30:297;34:53;49:241;54:51;57:15;60:783;61:42,54;89:212;90:95;94:741,1153;106:418;112:103,133;114:90;122:
abl@2:56;19:193;27:17;45:84;56:144;62:225;66:103;122:264,270;124:180,195;128:363,604;130:149
abnorm@60:324
abound@58:225
abov@85:135
absolut@36:259;68:221;116:579
absorb@51:210;87:277;127:150
abus@98:250
ac@21:42;130:116
academ@94:466
acceler@79:52
accentur@63:65,69
accept@30:18;32:43;34:255;43:56;51:74;66:196,674;72:88;74:159,257;95:178;97:75,99,127;105:34;109:92;131:9,20,31,88,102,135,149,407
access@10:243,765;19:250;23:41,69;34:290;50:7,144;54:64,98,345,353;57:49;60:238,343,534,897,904,914,949,974,1000,1021,1028;66:563;67:3
accid@40:1056;115:88
accident@49:246;122:39
accommod@26:70;29:11,84;37:422;42:297,378,407;60:73;65:449;95:248;104:455,490;113:445
accomplish@5:73,88;6:100,107;8:241;13:94;42:232;61:6;70:19,79;71:26,162;72:11;79:7;87:88,214,250;89:76;98:136;102:209;103:91;110:30;11
accor@69:51
accord@4:284;40:934;61:138;62:211;63:281;64:236;73:95;80:90;117:354,628,691;119:44
account@16:97;19:73,76,109,285,286;49:91;86:455;97:348
accru@86:330
accur@3:335,355;8:507;24:75;40:67,902;46:266;52:55;65:190;68:252;78:219;99:241;102:306;123:135;131:72
accuraci@3:372;55:264;57:176;83:33;87:70;112:79
acet@65:620
achiev@8:463;13:145;14:184;28:88;30:145;34:475;46:196,337;52:129,164;57:431;61:44,135;73:13;76:29,180,411;86:46;90:293;95:257;98:110,5
acm@32:200,204
acquaint@32:284
acquir@16:70;27:581;30:139;44:6,60;59:94;62:198;80:341;81:63;91:49,140;119:68;130:90
acquisit@44:20,83,149,179;81:220;90:471,568
acrobat@66:892
act@2:100,134;12:262;25:96;30:374;55:371;60:10,297;91:70;127:196;130:278;132:11
acter@40:554
acterist@76:151
action@40:175,478;41:66;49:220;73:158;77:109;82:16,26,50;122:157,251
activ@2:43;3:112,313;5:210;6:94,204;8:43,67,121,147,170,218,228,267,272,284,298,311,325,351,371,391,430,446,457,471,516,522,524,528,53
actress@7:62
actual@0:119;3:201;4:62;10:637;31:160;42:168,268,443;57:18;68:392;71:148;75:287,436;77:93;82:193;88:103;94:393;98:322,352;105:160;108:
ad@5:116;20:142;33:88;38:241;61:127,360;65:330;66:621;76:191;80:11;86:262;104:269,287,304;106:184;113:107;124:102;129:211;135:131
```

Fig.4: Inverted Index File

5. CONCLUSION

The implementation of stop word removal through preprocessing algorithm has been successfully completed for sample dataset of 137 documents, dataset size reduced by 40%. The output of generated inverted index file is also presented in paper which indicates easy and efficient retrieval of text documents as per user query from preprocessed dataset. There are several tasks for future work, among them: 1) Clustering for effective information retrieval along with fitness functions, 2) Page Ranking.

6. REFERENCES

[1] V. Srividhya, R. Anitha , “ Evaluating Preprocessing Techniques in Text Categorization “,ISSN 0974-0767,International Journal of Computer Science and Application Issue 2010

[2] Xue, X. and Zhou, Z. (2009) ” Distributional Features for TextCategorization “, IEEE Transactions on Knowledge and Data Engineering,Vol. 21, No. 3, Pp. 428-442.

[3] Porter, M. (1980) “An algorithm for suffix stripping, Program “, Vol. 14, No. 3, Pp. 130–137.

[4] Salton, G., “Automatic information organization and retrieval”, McGraw-Hill, New York. 1968

[5] Spärck Jones, K., “A statistical interpretation of term specificity and its application in retrieval”, Journal of Documentation, vol. pp. 28, 11–21, 1972.

[6] Tian Xia, Yanmei Chai “An Improvement to TF-IDF: Term Distribution based Term Weight Algorithm”, JOURNAL OF SOFTWARE, VOL. 6, NO. 3, MARCH 2011

[7] René Arnulfo García-Hernández , J. Fco. Martínez-Trinidad and J. Ariel Carrasco-Ochoa, Finding Maximal “Sequential patterns in Text Document Collections and Single Documents” Informatica 34 (2010) 93–101 93

[8] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, “PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth”.

[9] R. Srikant and R. Agrawal. “Mining sequential patterns: Generalizations and performance improvements”. In Proc. 5th Int. Conf. Extending Database Technology (EDBT’96), pages 3–17, Avignon, France, Mar. 1996.

[10] Carlos Cobos, Henry Muñoz-Collazos, Richar Urbano-Muñoz, Martha Mendoza, Elizabeth Leónc, Enrique Herrera-Viedma “Clustering of web search results based on the cuckoo search algorithm and balanced bayesian information criterion” ELSEVIER Publication, 2014 Elsevier Inc. All rights reserved ,21 May9 2014.

[11] X.-S. Yang, “Nature-Inspired Metaheuristic Algorithms” (2008) 128.

[12] Rui Tang, Simon Fong, Xin-She Yang, Suash Deb,” Integrating nature-inspired optimization algorithms to k-means clustering”, 978-1-4673-2430-4/12/$31.00 ©2012 IEEE.

[13] Carlos Cobos, Henry Muñoz-Collazos, Richar Urbano-Muñoz, Martha Mendoza, Elizabeth Leónc, Enrique Herrera-Viedma “Clustering Of Web Search Results Based On The Cuckoo Search Algorithm And Balanced Bayesian Information Criterion ” ELSEVIER Publication, 2014 Elsevier Inc. All rights reserved ,21 May 2014.

[14] Manoj Chahal,Jaswinder Singh "Effective Information Retrieval Using Similarity Function: Horngand Yeh Coefficient",Volume 3, Issue 8, August 2013.

[15] Agnihotri, D.; Verma, K.; Tripathi, P., "Pattern and Cluster Mining on Text Data," Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on, vol., no., pp.428,432, 7-9 April 2014

[16] Patil, L.H.; Atique, M., "A novel approach for feature selection method TF-IDF in document clustering," Advance Computing Conference (IACC), 2013

[17] http://www.ardendertat.com/2011/05/30/how-to-implement-a-search-enginepart-1-create-index/

[18] Anagha Chaudhari, Amitabh Mudiraj, Yogesh Jagdale, Pravin Phjadtare, Raviraj Mohite, Rohan Petare, Pranil Kudale, “Preprocessing of High Dimensional Dataset for Developing Expert IR System”, ICCUBEA-2015, March 2015.