A Framework To Automatically Categorize The Unstructured Text Documents

Anshika Singh*

Bhaskaracharya College of Applied Sciences, Dwaraka, University of Delhi, New Delhi – 110075, Delhi, India;
anshika1807@gmail.com

Abstract

Objectives: This paper presents an application of Text mining to automatically categorize the text documents from a predefined set of class or categories based on the content of the document. Method: In our framework, Association rule mining is used to generate frequently occurring word sets from training data set rather than the word sets in the whole document to derive a feature set to be used as training data. The new documents are then classified by applying classification method on the extracted feature sets. Findings: The results show that instead of using the whole word set extracted from the training document, frequently associated word sets extracted by applying association rules provide much accurate classification of the new document. Applications: The associated word sets not only relatively reduce the size of the training data but also relate to context of the document, which is further used to classify the new document. This approach can be used for efficient storage and retrieval of documents over the web or organizational databases. Hierarchical classification organizes the documents into tree like structure or meaningful hierarchy to form a single path structure which can add new categories or classes when needed. This approach can be combined with hierarchical structure for efficient browsing on search engines and provide a boulevard for future research.

Keywords: Association Rules, Classification, Text Categorization, Text Mining

1. Introduction

Data mining can be described as Knowledge Discovery from data, i.e., significant extraction of valuable information or knowledge from an extensive assortment of data or simply database. Though, data mining itself is an integral part in knowledge discovery as shown in Figure 1.

![Knowledge discovery from data](image1.png)

Figure 1. Knowledge discovery from data.
Data mining can be categorized into the following, shown in Figure 2:

![Figure 2](image)

**Figure 2.** A model for Data mining.

### 1.1 Data Collection
Collecting the data for processing usually stored in database or data warehouses in the form of spreadsheets, surveys, files etc.

### 1.2 Data Processing
comprises of two steps: Feature extraction to abstract out the significant features for a specific application and Data Cleaning and Integration to eradicate erroneous or missing data in order to ensure no inconsistencies presenting data and integrate the extracted data into a unified format which can be utilized for further processing.

### 1.3 Data Analysis or Analytical processing
Efficient analysis of processed data and determining an ideal algorithm to figure out the best solution to the problem. A problem may further be divided into sub problems and represented as building blocks to generate the final outcome or conclusion.

Text mining can be addressed as Intelligent Text Analysis or Knowledge Discovery in Text (KDT) and defined as a knowledge intensive process which involves user interaction with a huge assortment of data to ascertain and evaluate the useful information by identifying and exploring interesting patterns of specific type in its core knowledge discovery process with the help of analytic tools.

Text mining is analogous to data mining, except the data mining presumes that data is already structured; data preprocessing focus on scrubbing and normalizing the data and creating table joins. Text mining extracts information from unstructured or semi-structured data in the documents in the data collection or databases. Here, the preprocessing aims to identify and extract the features for representing natural language text and transform these unstructured representative features into an intermediate format which is explicitly structured.

Text mining is an automated or partially automated processing of text. Text mining can be classified in three major steps: Data Collection, Data Preprocessing and Analysis as illustrated in Figure 3.

![Figure 3](image)

**Figure 3.** A model for Text Mining.

In text mining, the primary concern is document collection which may be formed by the grouping of various textual documents and denoted by Text corpus. A text corpus is simply a raw text, i.e., unstructured. The first aim of applying grammatical parsing and preprocessing operations on text corpus is to select or extract relatively more subsets of features to represent the documents in order to achieve the accurate calibration of volume and semantic level of features to depict the document meaning precisely. The latent features employed to represent the document are characters, words, terms, concepts,
linguistic phrases, key phrases etc. The other objective is to streamline the representative feature sets by identifying the features which are computationally efficient and useful for discovering patterns. The process is sometimes supported by normalization, validation, cross reference of features against constrained vocabularies or dictionaries, ontologies, lexicons or knowledge bases to generate a smaller subset of refined semantically rich features. The core functionality is the analysis of frequency sets, their associations, distributions, and co-occurrence, to discover the nature and relationships between the features as reflected in the Text corpus as a whole. Consequently, a document-term matrix is computed which is more structured representation from either text corpus or text database. Document-term matrix represents bag-of-words mechanism and consists of frequencies of each term for all documents in the corpus which is further used as input for analysis phase. In the analysis phase, core mining operations such as classification, clustering, relationship identification, syntax analysis, document summarization etc. are performed on the document term matrix and visualization tools are used to facilitate the navigation and exploration of concept patterns which use graphical approaches to express complex relationships between data.

Perhaps, Text mining process exploits various methodologies and techniques like data mining, information extraction, information retrieval, and corpus based computational linguistics. For instance, Text mining make use of Information Extraction technique for preprocessing a corpus of text documents into a more structured format which can further be used as input to data mining module as illustrated in Figure 4. Information Extraction (IE) is defined as the process which automatically obtains structured data from unstructured data, i.e., natural language document. IE perform the following:

(a) Term Analysis, i.e., identifying the terms in a document, where a term may consist of one or more words,
(b) Named-entity recognition, i.e., identifying the names in document, such as name of people, place, time, quantities and associated units, etc. and
(c) Fact extraction, i.e., identifying and extracting complex facts from documents such as relationships between entities or events. Information Retrieval techniques have extensively used the “bag-of-words” model for tasks such as document matching, ranking, and clustering.

1.4 FP-Growth Algorithm

FP-Growth algorithm is used for generating association rules based on divide and conquer method. This algorithm generates frequent item sets using the combination of data attributes, without any candidate set generation. The algorithm consists of two major steps:
1) Building a compressed data structure showing frequent item set called FP-tree.
2) Divide FP-tree into a set of conditional Database and mines each database separately, thus extract frequent itemsets from FP-tree directly.

**Step 1: Constructing a FP-tree.**

Input: A transaction database D.
Output: A FP-tree

**Procedure:**
1. Create the root of the FP-tree labeled as “NULL”.
2. Scan the transaction database D and finding minimum support (min_support) for each item.
3. Collect the frequent item sets (F_items) and create nodes for each F_item and map the path for each node.
4. Note: If two transactions share the common prefix items then the path for both will overlap and frequency count for that node will be incremented by 1. Otherwise, the paths for two transactions sharing common item but uncommon prefix items will be disjoint and link the nodes of shared item for both the transactions.

![Figure 4. A text mining framework based on Information Extraction.](image)
5. Sort frequent item sets, $F_{\text{items}}$ in decreasing order of their min_support to form a sorted list of frequent item sets, $L_{\text{freq}}$.

   For each transaction $T_{id}$ in $D$, Repeat steps 2 to 4 till all the transactions are mapped to FP-tree path.

   For example, consider the transaction database $D = \{T_1, T_2, T_3, T_4, T_5, T_6, T_7, T_8, T_9, T_{10}\}$, where $T_1: \{a, b\}$, $T_2: \{b, c, d\}$, $T_3: \{a, c, d, e\}$, $T_4: \{a, d, e\}$, $T_5: \{a, b, c\}$, $T_6: \{a, b, c, d\}$, $T_7: \{a\}$, $T_8: \{a, b, c\}$, $T_9: \{a, b, d\}$, $T_{10}: \{b, c, e\}$.

   Firstly, Create the root node of FP-tree as NULL. Then start scanning the transaction database $D$.

   Read $T_1: \{a, b\}$ and create two nodes – $a$ and $b$ with the path $\text{NULL} \rightarrow a \rightarrow b$ and set min_support of $a$ and $b$ to 1.

   Read $T_2: \{b, c, d\}$ and create three nodes- $b$, $c$ and $d$ with the path $\text{NULL} \rightarrow b \rightarrow c \rightarrow d$ and set min_support of $b$, $c$, and $d$ to 1. Though item $b$ is shared by both $T_1$ and $T_2$, the path for $T_1$ and $T_2$ are disjoint as they do not share common prefix. Add the links between $b$ of both transactions.

   Read $T_3: \{a, c, d, e\}$: Since $T_1$ and $T_3$ share common prefix item $a$, the path for both will overlap and min_support for node $a$ will be incremented by 1. Create three nodes $c$, $d$, and $e$ with the path $a \rightarrow c \rightarrow d \rightarrow e$ and set min_support of $c$, $d$, and $e$ to 1. Though items $\{c, d\}$ is shared by both $T_2$ and $T_3$, the path for $T_2$ and $T_3$ are disjoint as they do not share common prefix. Add the links between $c$ of $T_2$ and $T_3$ and $d$ of $T_2$ and $T_3$ respectively.

   Repeat the procedure till all the transactions in $D$ are mapped to FP-tree path. Constructed FP-tree is given in Figure 5.

Step 2: Generating Frequent item sets:

   The FP-growth algorithm follows Bottom-up approach (i.e., from leaves towards the root) to generate the frequent item sets from FP-tree.

1. Firstly, extract prefix path subtrees ending in an item set to extract frequent itemsets by decomposing the problem recursively and obtain the conditional FP-tree. While building Conditional FP-tree, we consider the transactions which contain only a particular item set and removing the rest from transaction database. For example. Prefix path subtree with path that contain node e is illustrated in Figure 6.

2. Then merge the results.

![Image of FP-tree constructed for transaction database D.](null)

**Figure 5.** FP-tree constructed for transaction database D.

![Image of Prefix path subtrees with path that contain node e.](null)

**Figure 6.** Prefix path subtrees with path that contain node e.

**Algorithm for generating frequent itemsets with FP-tree by pattern fragment growth:**

In order to mine frequent itemsets from FP-tree, we call the method: $\text{FP – growth} (\text{FP – tree}, \text{NULL})$ which is implemented as follows:

- **Input:** FP-tree constructed in step 1 and minimum support threshold.
- **Output:** Complete set of frequent patterns

\[
\text{FP – growth} (\text{Tree}, P1)
\]

- If tree contains a single prefix path $P$ then for each combination set of nodes in $P$ generate a pattern $P1 \cup P2$ with min_support= minimum support of nodes in $P2$;
else for each item \( a_i \) in multipath subtree \( Q \),
then generate a pattern \( P2 = a_i \cup P1 \) with
\[
\text{min\_support} = a_i.\text{min\_support}
\]
construct \( P2 \)'s conditional pattern base and conditional FP-tree say \( \text{Tree}_{P2} \);

if \( \text{Tree}_{P2} = \emptyset \)
then call \( \text{FP\_growth}(\text{Tree}_{P2}, P2) \)

Frequent item sets for transaction database \( D \):

\begin{align*}
\text{Prefix} & \quad \text{Frequent itemsets} \\
e & \quad \{e\}, \{d, e\}, \{a, d, e\}, \{c, e\}, \{a, e\} \\
d & \quad \{d\}, \{c, d\}, \{b, c, d\}, \{a, c, d\}, \{b, d\}, \{a, b, d\}, \{a, d\} \\
c & \quad \{c\}, \{b, c\}, \{a, b, c\}, \{a, c\} \\
b & \quad \{b\}, \{a, b\} \\
a & \quad \{a\}
\end{align*}

1.1 Naïve Bayesian Classification for Text

The naïve-bayes classifier for calculating the probability (PNB) that a document having attribute \( a_i \) belongs to class \( j(C_j) \) is given as:

\[
\text{PNB} = P(C_j) \Pi P(ai | C_j)
\]

\( P(C_j) \) is estimated based on the fraction of each target class in the training data. \( P(ai | C_j) \) is calculated after m-estimate approach in order to avoid zero probability value using the following equation:

\[
\frac{n_k + 1}{n + \text{vocabulary}}
\]

where \( n \): Total no of frequent word sets in all training examples whose class value is \( C_j \),
\( n_k \): no. of occurrences of the frequent word set found amongst \( n \),

\( \text{vocabulary} \): Total no. of distinct word sets found within all the training data,

\( P_c \): Conditional Probability

\[
\text{Training\_Naive\_Bayes\_Classifier}(\text{class, Document}) \{
\text{Vocabulary} \leftarrow \text{Extract\_vocabulary(Document)}
\text{N} \leftarrow \text{Count\_documents(Document)}
\text{For each class} \in \text{Category}
\text{do} \quad \text{N}_c \leftarrow \text{Count\_documents\_in\_class}(\text{Document, Class})
\text{Prior}[c] \leftarrow \frac{\text{N}_c}{\text{N}}
\}
\]

\[
\text{Testing\_Naive\_Bayes\_Classifier}(\text{Category, Vocabulary, Prior, Document}) \{
\text{Weight} \leftarrow \text{Extract\_Tokens\_From\_Document(Vocabulary, Document)}
\text{For each class} \in \text{Category}
\text{do} \quad \text{score}[\text{class}] \leftarrow \log \text{prior}[\text{class}]
\text{For each term} \in \text{Weight}
\text{do} \quad \text{score}[\text{class}] += \log P_c[\text{term}][\text{class}]
\text{return} \quad \arg \max_{\text{class} \in \text{Category}} \text{score}[\text{class}]
\}
\]

2. Proposed Framework

Text categorization or Document categorization can be defined as a method of assigning a target class or category to the document from a predefined set of class or categories based on the content of the document. In our proposed framework, we automatically classify the research papers into a set of categories using Text mining. In our implementation, we used the tool Rapidminer which provides an integrated environment for text mining, as illustrated in Figure 7. We initially start with a collection of research papers of various disciplines to form a text corpus which is simply an assortment of natural-language documents. Each document is considered as a transaction in text corpus and set of items represent frequently occurring words in the transaction. The next step is the application of text mining techniques in order to obtain the Document term matrix for further analysis which starts with parsing and preprocessing the text corpus. Once the Term-Document matrix has been obtained, we discover the frequent item sets from each document by applying FP-growth algorithm. Lastly, we apply Naïve Bayes classification on the derived features sets to categorize the new documents.
2.1 Data Collection
Initially, we collected 34 research papers of three disciplines: Computer Science (CS), Electrical Engineering (EE) and Mechanical Engineering (ME) from the World Wide Web to form a text corpus.

2.2 Data Preprocessing
The preprocessing includes grammatical parsing, data cleaning, stemming, part-of-speech tagging of the text corpus in order to achieve the accurate calibration of volume and semantic level of features to depict the document meaning precisely. This process involves eliminating commas, punctuations, periods, stop words like “the,” “all” etc. and extracting the latent features such as words, terms, concepts, key phrases etc. Stemming refers to extraction of the <root> of the word based on dictionary or set of rules given by linguists. For example, the word “process” is extracted from the word “processing”. Part-of-speech is used to mark up the words equivalent to a particular part of speech. For example, identifying the words as nouns, verbs, adjectives, adverbs, etc. based on definition or context. The next step is creating the Term document matrix, $T$ by counting the number of occurrences of each term in each document $d_j$ where $d_j$ is a multidimensional vector of keywords. Each row $t_i$ of $T$ denotes occurrence of the term in each document $D_j$. The document $d_j$ is represented as

$$
d_j = \begin{bmatrix}
    f_{1j} \\
    f_{2j} \\
    \vdots \\
    f_{nj}
\end{bmatrix}
$$

where $f_{ij}$ represent no. of occurrences of $i^{th}$ term in document $j$. This type of representing complete set of these $n$ vectors for all documents is referred as Vector space model is shown in Figure 8. The Term-Document matrix is as follows:

$$
D = \begin{bmatrix}
    f_{11} & f_{12} & \cdots & f_{1n_d} \\
    d_{11} & d_{12} & \cdots & d_{1n_d} \\
    \vdots & \vdots & \ddots & \vdots \\
    d_{n_t1} & d_{n_t2} & \cdots & d_{n_tn_d}
\end{bmatrix}
$$

There are several methods to calculate the weights of the words or terms. One of the popular methods is Term Frequency Inverse Document Frequency (TF-IDF) where TF represent the frequency of word or term in a document and IDF denote the measure of the rarity of a word or term within the complete set of documents. TF and IDF are represented as:

$$IF_i = \frac{n_i}{\sum_{i=1}^{k} n_k} \quad \text{and} \quad IDF_i = \frac{\log N}{df_i}$$

where $n_i = \text{no. of occurrences of the considered terms}$ $n_k = \text{no. of occurrences of all terms in the document}$ $N = \text{no. of occurrences of the considered terms}$ $df_i = \text{no. of documents that contain term } i$.

TF-IDF method measures the statistical strength of the given term in reference to the query.
TF – IDF = \text{TF}_i \times \text{IDF}_i

### 2.3 Analysis

Firstly, we generate an extensive feature set from the association rules obtained from FP-growth algorithm. Association Rules and Rules tree generated from application of FP-growth Algorithm is shown in Figure 9 and Figure 10 respectively. The feature set \( L_i \) representing frequent word sets is generated from the term document matrix \( T \) is generated using input as minimum support(min_support).

**Association Rules**

**Association Rules**

\[
\begin{align*}
\text{[based]} & \rightarrow \text{[using]} \quad \text{(confidence: 0.010)} \\
\text{[based]} & \rightarrow \text{[introduction]} \quad \text{(confidence: 0.010)} \\
\text{[work]} & \rightarrow \text{[using]} \quad \text{(confidence: 0.012)} \\
\text{[example]} & \rightarrow \text{[using]} \quad \text{(confidence: 0.012)} \\
\text{[case]} & \rightarrow \text{[using]} \quad \text{(confidence: 0.012)} \\
\text{[study]} & \rightarrow \text{[introduction]} \quad \text{(confidence: 0.012)} \\
\text{[example]} & \rightarrow \text{[introduction]} \quad \text{(confidence: 0.012)} \\
\text{[system]} & \rightarrow \text{[time]} \quad \text{(confidence: 0.012)} \\
\text{[case]} & \rightarrow \text{[time]} \quad \text{(confidence: 0.012)} \\
\text{[using, given]} & \rightarrow \text{[introduction]} \quad \text{(confidence: 0.012)} \\
\text{[using, current]} & \rightarrow \text{[introduction]} \quad \text{(confidence: 0.012)} \\
\text{[work]} & \rightarrow \text{[using, introduction]} \quad \text{(confidence: 0.012)} \\
\text{[introduction, work]} & \rightarrow \text{[using]} \quad \text{(confidence: 0.012)} \\
\text{[using, current]} & \rightarrow \text{[time]} \quad \text{(confidence: 0.012)} \\
\text{[system]} & \rightarrow \text{[using, time]} \quad \text{(confidence: 0.012)} \\
\text{[using]} & \rightarrow \text{[introduction]} \quad \text{(confidence: 0.015)}
\end{align*}
\]

**Figure 9.** Association Rules Generation.

The we proceed with regular expression search for finding the matching feature set or word sets in \( L_i \) with that of words in the new document and calculating the probability values of each feature set \( L_i \) containing frequent word sets using Naïve Bayes method. Lastly, we apply the Naïve Bayes classification algorithm and calculate the probability of each class using conditional probability of the matched features sets or word sets and the probability values for each target class in order categorize the new document based on the target class having highest probability values.

### 3. Implementation and Results

To illustrate our framework, we collected 34 research papers of three disciplines: Computer Science (CS), Electrical Engineering and Mechanical Engineering from the World Wide Web to form a text corpus. Considering each document as a transaction and minimum support is two. We preprocess the text corpus to ensure that the data doesn’t contain any unnecessary word or filling the missing gaps in the data. Then the extracted data is transformed into a term-document matrix of training set which serves as an input for further application of FP-growth algorithm in the analysis phase. As a result, we get the frequent word sets and their number of occurrences in each predefined classes called Computer Science (CS), Electrical Engineering (EE) and Mechanical Engineering (ME). We find the probability of the frequent word sets using Naïve Bayesian classifier. After preprocessing the new document, we search for matching word sets in the feature sets obtained from training data with that of frequent word set of new document. We collected conditional probability of the matched word set for each target class and calculated the probability values for each target class using Naïve Bayesian Method. Unlike the word set of the whole document as training data, we used the frequent word sets created from FP-growth algorithm which...
employs association rules for much accurate classification of the new document.

We selected a random text file named “testme.txt” as input to our framework as shown in Figure 11 and the output shows the document belongs to the category: Electrical Engineering (EE). The results are illustrated in the Figure 12.

![Figure 11. New Text Document.](image1)

![Figure 12. Prediction of category for a new document as EE.](image2)

### 4. Conclusion

In our approach, we considered documents as transactions and the set of frequently occurring words as a set of items in the transaction. We used FP-growth to generate frequent occurring word sets from training data set rather than the word sets in the whole document to derive a feature set to be used as training data. The new documents are classified by applying Naïve Bayes method on these feature sets. The results show that our approach is an effective way to categorize the text document.

### 5. References