Indian Journal of Science and Technology, *Vol 9(37), DOI: 10.17485/ijst/2016/v9i37/98974, October 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Feature Selection for Automatic Categorization of Patent Documents

S. Don[1*] and Dugki Min[2]

[1]Department of Analytics, School of Computer Science and Engineering, VIT University, Vellore - 632014, Tamil Nadu, India; don.sasikumar@vit.ac.in
[2]Department of Computer Science and Engineering, Konkuk University, Seoul, South Korea; dkmin@konkuk.ac.kr

Abstract

Objective: With the rapid increase in the number of patent documents worldwide, demand for their automatic categorization has grown significantly. The automatic categorization of patent documents is the organization of such documents in digital form, thus replacing the manual time-consuming process. In this work, we proposed a system that can automatically categorize patent document by considering the structural information of the patents. **Methods:** We propose a three-stage mechanism for automatic categorization. In the first stage, we apply a pre-processing mechanism to reduce unwanted noise that can influence the categorization process. Such noise includes terms that have less structural meaning in the document. In the second stage, feature selection is conducted based on the term frequencies. Feature vectors are constructed from the structural information of the patent. In the third stage, classifications are conducted using a Random Forest (RF), Support Vector Machine (SVM), and Naïve Bayes (NB) classifier. **Findings:** It was found that the semantic structural information of a patent document is an important feature set in constructing the terms of a document for the categorization. The experimental results also show that feature reduction using Information Gain (IG) is beneficial for obtaining a higher accuracy rate in a reduced dimensional space. **Applications:** The results reveal the importance of the proposed method for automatic categorization of patent documents.

Keywords: Classification, Feature Selection, Patent categorization, Structural information

1. Introduction

With the exponential growth of information in the digital world and the increase in the filing of new patents, an efficient way for organizing such documents is becoming a necessity. Traditional approaches require human labour and a time-consuming process[1]. Patent categorization is the process of classifying patents into specific classes that help the patent examiner to evaluate new patents. The patent examiner evaluates a new patent by comparing it with the most similar patents from the database. Further refinement is required to reduce the number of patents for comparison based on their structural information. Thus, patent categorization is an important step in the processing life-cycle of a new patent registration. According to the World Intellectual Property Organization[2], there has been a consistent growth in the number of patent applications. The latest report shows a 5.1 % growth for the year 2013, as compared to the previous year. The United States and China had the highest number of patent filings in 2013. The United States Trademark Office has granted eight-million patents. From these figures, it is clear that there has been an exponential growth in the amount of patent information. Many different approaches for facilitating and streamlining the patent classification process have been suggested. Most of the text classification approaches are based on traditional machine learning techniques[3-5]. Data mining and natural language processing techniques play a pivotal role in this area. Many methods have been reported[6-9] and successfully implemented in different text categorization applications. However, the semantic structural information of each patent differs from traditional

*Author for correspondence

text processing applications. We used the word *"semantic"* in this paper to denote the patent document structure. Patents are documents structured based on their semantic structure such as <title>, <abstract>,<claim>, <description>,etc. Thus classifying patent from these structures into specific domain of interest is challenging. Traditional classification systems[2] are based on either International Patent Classification (IPC), or a keyword search, or the title search, or any such combinations. A comparison of such methods shows that having a human expert read the patents is the most effective way of analyzing the documents. However, this is a time-consuming and difficult process for larger documents. Because patents are structured documents, many questions may arise, such as what structural information that we need to consider (i.e., title, abstract, claim, or description), how efficient will the categorization system be when considering different structural information? To utilize the full potential of the structural information of a patent for categorization, we propose a three-stage representation model for classification purposes. In this process, the semantic structure of the patent document, such as the <title>, <abstract>, <claim>, and <description>, is used and represented as a term in the feature vector space. Each document has a frequency value for each term. A major problem in constructing the feature vector space is its high-dimensional vector space. Since each document contains unwanted noise information, it is necessary to remove such noises to reduce the computational complexity. This includes removing stop-words. In the case of patent data processing, we have created additional sets of words common to all patents that have less meaning in the feature vector construction. These words are also selected during the stop-word construction phase. In the feature selection process we construct the feature vectors based on term frequencies. Based on the reconstructed feature vectors, we designed a classifier based on Random Forest (RF)[10,11] Support Vector Machine (SVM)[12-16] and Naïve Bayes (NB)[16] for the classification and validation phases. With our proposed system, it is more applicable to handle large volumes of patents for categorization. We evaluated the performance of the proposed system using publically available patents. With the increase in new patent documents being filed globally, there has been a high demand in protecting patent information. Different techniques exist to manage such protection. The World Intellectual Property Organization defined the taxonomy for organizing patents[17,18]. The major patent filing organizations include the United States Patent and Trademark Office (USPTO)[i], the European Patent Office (EPO)[ii], the Japan Patent Office[iii], the Chinese Patent Office[iv], and the Korean Patent Office[v] (*[i]www.uspto.gov; [ii]www.epo.org; [iii]www.jpo.go.jp;[iv]www.sipo.gov.cn;[v]www.kipo.go.kr).*

These organizations semantically structure their patents in different ways. For example, Japanese patent documents[19] are structured based on the <bibliography>, <abstract>, <claims>, <description>, <explanation of drawings >and <drawings >. Each of these sections provides a detailed subjective description of the findings. In contrast, US patents are semantically structured based on the <abstract>, <background summary>, <detailed descriptions>, <claims>, etc. The purpose for using the text information inside these structures includes strategic planning, as well as technology and knowledge management. In[1], the author addresses how patent information can be beneficial to competitor monitoring, technology assessments, R&D portfolio management, etc. The contribution of this work has led to two main important assessments: (a) the use of patent information by senior management for decision-making purposes is an important area of technology management, and (b) external stakeholders of a firm have a growing interest in assessing the firms technological competence for future competitiveness. Traditional methods for classifying patents are based on IPC[18,20]. The United States Patent and Trademark Office classification scheme contains 400 classes and 135,000 subclasses. It is therefore a tedious job for the patent examiner to match a newly filed patent with an existing patent to determine the similarity between the two documents. Many papers have addressed this issue, and different methods have been proposed for organizing and finding the similarities among documents, and visualization tools for analyzing patent documents have been developed[21]. Natural Language Processing (NLP) has played a prominent role in the analysis and evaluation of patent documents. The authors of[22] presented the use of natural language processing techniques used in the European Patent Office. This method achieves a sufficient level of accuracy in classifying newly filed patents, thereby reducing the workload of the human examiner. In[23], a multi-classification method for the classification of large documents based on the winnow algorithm was proposed. The test corpora applied for the experiment were taken from the European Patent Office. Searching the text written inside the patent sections requires efficient search engine mechanisms. In[24], the author reviews the impact of

the historical development of online searching and document preparation from the resulting database. Part one discusses the challenges of using the content inside a patent for retrieval. The structure of the content categorization includes the <title>, <abstract>, and <claim>. Part two of the paper describes the role of implementing new methods for the document preparation and retrieval process that can improve the quality of a patent search. Identifying patents that are related to cross-disciplinary areas is difficult to define through the patent class. A keyword search is an appropriate mechanism for identifying such patents. Limited research has been conducted on selecting the appropriate keywords from the patent document, including the <title>, <abstract>, <claim>, and <descriptions>. In[25] built an advanced patent processing service called PAT Expert that can meet users on-demand requirements of patent processing services. PAT Expert introduced a content-representation scheme for patent documentation. Two types of techniques were introduced in this work. The first technique provides access to the content of a document, and the second type shows the content representation. The service was tested in two technological areas: optical recording devices and machine tools. Since patent information is important to modern businesses, choosing a specific tool for a patent analysis is the most important task. The authors of[26] presented a reference collection of patent documents for automated categorization by applying various machine learning algorithms. The results reveals that automation can help users those who are unfamiliar with IPC based classification. In[27] has used three different methods for CLEF-IP based patent classification by combining semantics and statistics driven techniques. The evaluation was carried out on both English document and German documents. The need for automatic patent classification is increasing due to the growing number and diversity of inventions. In[28] dwelled on the needs of automatic classification of patents, its issues, state-of-the art technologies and evaluation methods.In[19], considered the structural information of Japanese patents for the categorization process. The system processed the patents in three phases: indexing phase, retrieval phase and categorization phase. In addition, knowledge on which part of the patent section should be considered for the classification process is lacking. An overview of the patent information and innovative solutions in the area of patent informatics was provided in[17]. This study starts by identifying the actual requirements of different users of patent information and the management tasks they require. Innovation covers all important layers, from the database to the algorithm and online services. The study concentrates on intelligent and semantic solutions proposed in recent years. Whereas most of the works mentioned thus have focused on a text-based analysis of a patent document. The authors in[29] studied the potential benefits, requirements, and challenges involved in patent image retrieval. They proposed a framework with the potential capability of an advanced image analysis, and indexing techniques to address the need for content-based patent image search and retrieval. To evaluate their framework, a search engine called PatMedia was developed. The results generated from PatMedia have been encouraging, and a comparison was made with an existing system called PATSEEK. The results were evaluated in terms of the precision and recall measurements. In[30] provided the importance of information retrieval in the area of patent classification. Document categorization is considered an active area of research in machine-learning communities[31]. Most of the existing research has focused on text categorization from news feeds. Such studies have mostly selected benchmark datasets, including Reuters 21578[32,33], 20Newsgroups[vi] , and Classic3 [vii] ([vi]*http : //qwone.com/ jason/20Newsgroups/*[vii] *ftp : //ftp.cs.cornell.edu/pub/smart*). A two-stage feature selection method for text categorization was conducted in[6]. Here, documents are ranked depending on their importance to the classification. This is achieved using Information Gain (IG) methods. A Genetic Algorithm (GA) and a Principle Component Analysis (PCA) were considered for feature selection and feature extraction. The classification was conducted using the k-nearest neighbor (KNN) and C4.5 decision tree algorithm. The results reveal that this approach is able to achieve high precision, recall, and f-measure scores. The authors in[3] proposed the use of inductive learning to categorize documents into predefined categories. Here, a Bayesian classifier and a decision tree learning algorithm show a reasonable level of performance. Text categorization refers to the labeling of an unclassified document using a classifier that has some labeled documents as a training set. The authors in[4] proposed a two-level representation model (2RM) to represent text data. This includes both syntactic and semantic information. Two classifiers are considered for syntactic and semantic information. The resultant outputs from these two classifiers are given to a third classifier as input. Experiments conducted on publically available datasets show that the proposed method improves the classifica-

tion rate as compared to other text representation models. Feature selection is an important step of text categorization. The main purpose of feature selection is to determine which candidate features are the most relevant attributes for classification purposes. An accurate selection of these features plays a pivotal role in the accuracy of text classification. An empirical study on selecting different feature selection methods is described in[34]. In this study, twelve sets of features are used as candidates, and are evaluated on a benchmark of 229 different text classification problems. The results are analyzed in terms of accuracy, precision, recall, and F-measure. The authors of[31] presented a comparative analysis of five different feature selection methods for text categorization. These features include information gain, X2 statistics, document frequency, term strength, and mutual information. This paper concludes that information gain is the most appropriate candidate for the purpose of text categorization. The authors in[35] provide a detailed overview of several popular feature selection methods considered for text classification. Feature selection from imbalanced data for categorization is more difficult than from balanced data. In[36], a feature selection framework for text categorization is suggested. Three cases were used for the scenario. The first one considers positive features only using a one-sided matrix. The second implicitly combines positive and negative features, and the third combines two kinds of features explicitly, and chooses the size ratio empirically. One of the main conclusions is that the feature selection can significantly improve the performance. The authors in[37] proposed feature selection method for text classification via global information gain method. In[38] a novel filter based probabilistic feature selection method namely distinguishing feature selector(DFS) for text classification was proposed. And the result indicate that the performance of the DFS is competitive compared to other traditional chi square, information gain and deviation from Poisson distribution. A two stage term reduction based on information gain and geometric particle swarm optimization(GPSO) is described in[39]. In[40] a novel projected prototype based classifier was proposed for text classification.

As the above survey indicates, choosing an appropriate feature selection and classification algorithm for categorizing patent documents is not an easy task. The authors are interested in filling this gap using different parts of a patent document with different clusters of keywords. Keyword clusters are taken by measuring the recall and precision of each keyword from different parts of the document. The results show that the most efficient method for identifying patents in a specific domain through a keyword search is to select the text information from the patent sections, such as the <title>, <abstract>, and <claims>. We applied this procedure for our design model. Motivated by the above mentioned literatures, we propose a method for automatic categorization of patent documents. We evaluated the proposed method using various feature sets, and measured the performance based on the precision, recall, and f-measure.

3. Proposed Method

In patent document categorization, an input consists of a collection of documents that are split into a training set and classification set. Each document is represented in a vector space model, sometimes referred to as a bag of words. The problem addressed in this paper is to evaluate the effect of the different section of patent for the classification. We aim to study the classification accuracy in two levels. In the first level, we consider the features without reduction and in the next level by reducing the number of features using a standard set of evaluation technique. We consider the second level as an iterative process such that the performance of the classification rate variation depends on the size of the feature set. The basic preliminary of our proposed method is introduced in this section, an overview of which is shown in Figure 1. Figure 1(a) shows the representation of patent documents in the form of term document matrix and its categories. Figure 1(b) is the feature matrix with its class labels for the purpose of classification and dimensionality reduction. Figure 1(c) represents the information gain value for reduction stage. Figure 1(d) describes different cases of dimensionality reduction and finally Figure 1(e) is the classifier used for the purpose of classification without feature reduction and with feature reduction.

3.1 Document Categorization

3.1.1 Case Selection and Dataset Acquisition

We define the patent document categorization process as follows. Given a set of patent documents d with terms t in vector space t, we assume that there exists a class label that assigns each document into one of the c classes. Then, d can be represented as $\psi(d) = (w_{(t1,d)}, w_{(t2,d)}, ...w_{(tn,d)})^T \in R^D$,

where $w_{(tn,d)}$ is the weight of term t_n in *d*, *T* represents the transpose operator, and ψ(d) denotes the term weight of *d* in dimensional vector space *n*. To represent the whole corpus of *n* documents, the matrix TD_m of the number of terms versus the number of documents is defined using an M *N matrix. By transposing ψ(d) into $[\psi(d)]^T$, the rows represent the terms, and the columns represent the documents. Thus, each document will have term t_n and class label *c* for the purpose of categorization.

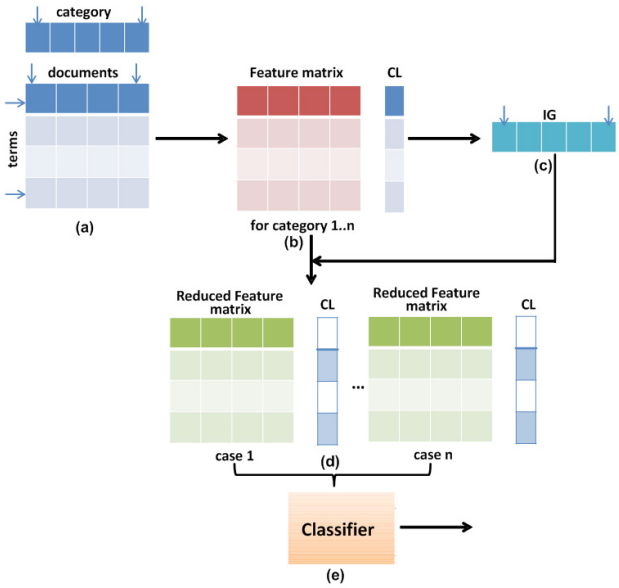

Figure 1. Classification process of the proposed methodology.

In our experiment, we applied Google Patents as a source for extracting the patent documents. Google Patents collects patent documents from USPTO, EPO and WIPO, and is freely available to users. For the current study, we developed our own standalone module that can automatically download a patent based on the users given the topic of interest. Rather than using Google-provided search facilities that can first visualize a patent and download it manually based on the users choice, in our module, the program downloads patents on a given keyword list for each user topic from the Google patent (<*https: // www.google.com/?tbm=pts&gws rd=ssl#newwindow=1& q=keyword&tbm=pts&tbs=ptst:u*>). By downloading the searching results from this URL, we are able to obtain the patent documents that contain the keywords and store them in HTML format. After collecting all of the patents, we extract the title, abstraction, claims, and descriptions by selecting data from the patent-title, patent-abstract section, patent-claims section, and patent-description sections. We then remove the stop-words from the extracted patent content based on Onixs stop-word list(*www.lextek.com/manuals/onix/stopwords1.html*). We then select verbs, nouns, adjectives, and adverbs from the remaining contents. Finally, we apply a lemmatisation algorithm that attempts to find the lemma of the words based on the vocabulary, along with a morphological analysis method provided by the Stanford Core NLP toolkit (*http : //nlp. stanford.edu/software/corenlp.shtml*) for normalizing all selected words. Figure 2 shows the proposed data acquisition system. We have been selected ten different topics of interest for the categorization.

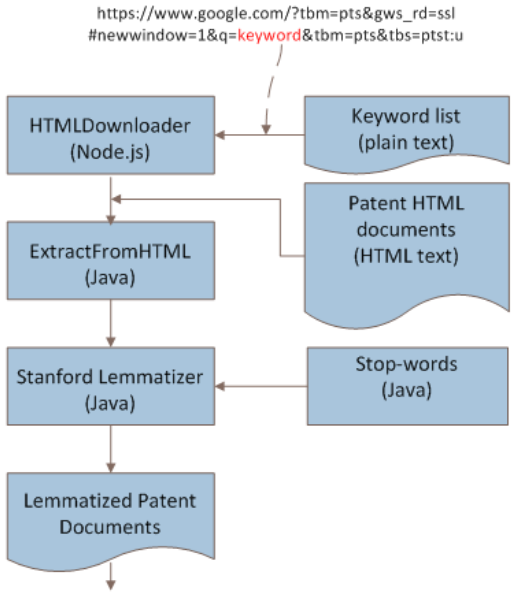

Figure 2. Control flow of the data acquisition system.

3.2 Pre-processing

For patent categorization, each patent document is converted into a set of terms called feature vectors. The general approach to representing these terms is in the form of bag of-words. In this method, the terms present in a document are represented by a Bernoulli or multinomial distribution. In the construction phase of the feature vector, many additional processes need to be conducted. The first problem that arises, however, is that the document may contain noisy information that includes the words and symbols used for the sentence construction. To remove these words and symbols during the preprocessing stage, techniques such as stop-word removal are performed[39]. The second problem is that the original

feature vector generates a high-dimensional vector space. This can occasionally lead to a decrease in classification performance. To improve the performance of the classifier, we need to reduce the size of the feature vector in a lower dimension. For our study, we used information gain to reduce the size of the feature vector and evaluate the performance of the classifier. A more detailed explanation is provided in the experiments section. The following sub-section provides detailed pre-processing steps that are taken into account for the construction of the dataset for the purpose of classification

3.2.1 Removing Stop-Words and Special Symbols

During the patent categorization phase, removing stop-words means removing common words such as a, an, and, the, and for, which frequently appear in a sentence. Removing these stop-words do not affect the categorization process. It also helps to reduce the dimension of the feature set. For this study, we used the stop-words list from Onix, which consists of 429 words. Apart from this, we created an additional set of patent stop-words manually that are used in the general construction of a patent document structure (comprise, invention, etc.). Another important step during the pre-processing phase is the removal of whitespace delimiters. This includes spaces, tabs, and special symbols. We also removed uninformative tokens such as (|∧, |0 − 9|, |i ± |i°|i| ± |.|, |[|/|; |(|, |)| − "|0|?|&| = | + | ∗ |%|@|) from the document.

3.2.2 Lemmatisation

The term lemmatisation means conducting the process properly through the use of a vocabulary and morphological analysis of the words, normally aiming at removing only inflectional endings and returning the base or dictionary form of a word, which is known as a lemma (*http : //nlp.stanford.edu/IR − book/*) . The lemmatisation process transforms words into their basic forms using a set of rules and a dictionary. We used Stanford parser to perform lemmatisation.

3.2.3 Term-Frequency Weighting

During the pre-processing stage of the patent categorization, once the terms are extracted after lemmatisation, the next step is to perform term weighting. Each document can be represented in a vector form depending on the number of terms it contains. In a binary vector representation, the presence or absence of terms in a document can be represented as a 0 or 1. Thus, the document is represented by d= {1,0,1,1,0}. In term-frequency weighting, the document is represented as

$$d = \{W_{t1}, W_{t2}, ...W_{ti}, ...W_{tn}\} \quad (1)$$

where W_{ti} is the term weight with index *i* in document d. To obtain the term-frequency weighting, term frequency (tf) is required, where *tf* is the number of terms in a document[39,6]. In this study, we used tf as the feature vector model for the categorization process.

3.2.4 Global Term Selection

The global term selection is a method for reducing the dimensions of the feature vectors applied during the document processing. This process removes the terms that are less important to the patent categorization. In our proposed method, we prune terms that appear fewer than three times in a document, which helps reduce the dimensions of the feature vector.

3.3 Local Term Selection

One of the main challenges in text categorization is the high dimensionality of its feature space. Most of the terms in the feature space are irrelevant to the categorization process. To preserve the performance of the classifier, it is important to construct the feature vector with a reduced space without sacrificing the classification Accuracy(Figure1(d)). The feature vectors obtained after pre-processing is sufficient for performing the classification properly; however, the computational cost for executing this process is high. Many techniques are available to reduce this cost by considering the size of the feature space[31]. In this work, we consider IG based methods to reduce the desired amount of terms from the feature vector, and thus evaluate the classification accuracy in a reduced feature vector space during our class-specific categorization process(Figure 1(c)).

3.3.1 Information Gain (IG)

Information gain is the most popular method used for the informativeness criterion of terms in the field of machine learning[19]. IG measures the number of bits of information obtained for a category prediction by knowing the presence or absence of a term in a document. Let $\{c_i\}_{i=1}^{m}$ denote the set of categories in the target space. The information gain of term t is defined by first providing four dependency tuples:

(i) (t, c_i), the presence of *t* with membership in c_i

(ii) $(t, \bar{c_i})$, the presence of *t* with non-membership in c_i

(iii) $(\bar{t}, c_i)$, the absence of *t* with membership in c_i

(iv) $(\bar{t}, \bar{c_i})$, the absence of *t* with non-membership in c_i

In these definitions, *t* and c_i represent the term and category, respectively. The first and third tuples represent the positive dependency between *t* and c_i, whereas the second and fourth tuples represent the negative dependency. Thus, the mutual information of term t and category c_i is defined in eq. (2). *P* is the probability of a term that is present or absent in a specific class considering the occurrence of the term within the document that belongs to that class[39].

$$IG(t,c_i) = \sum_{c\in\{c_i,\bar{c_i}\}} \sum_{t'\in\{t,\bar{t}\}} P(t',c).\log\frac{P(t',c)}{P(t').P(c)} \quad (2)$$

3.4 Patent Categorisation Algorithm

In this study, we applied three different classifiers for the purpose of patent categorisation. RF, SVM, and NB classifier methods are used owing to their accuracy and efficiency in document categorization (Figure 1(e)). These algorithms are applied separately to the dataset during the training and testing phases. The performances of these algorithms were evaluated based on different evaluation criteria. In the following sections, we provide brief descriptions of these three algorithms.

3.3.2 Random Forest

Breiman[10,11] first suggested the RF classification algorithm. RF is an ensemble learning algorithm that has received wide popularity in the machine learning community because it can handle high-dimensional classification, and the results are more accurate and robust to noise. RF obtains a class vote from each tree and then classifies the documents using the majority vote. These trees are typically grown using the CART methodology[41]. A good split is required that pushes the input data from a parent tree node to the child node[11]. Thus, based on a given collection of document features, a decision tree will be grown. RF considers a random subset of document features in the division of each node within the tree. To grow trees from different training documents, RF uses a method called bagging[42]. For classification, RF combines individual decision trees into large ensembles, where each tree contributes with a single vote for the assignment of the most frequent class to the input dataset[34].

3.3.3 Support Vector Machine

An SVM is a supervised classification algorithm that has proven to be an efficient learning algorithm for document categorization. It has an excellent performance for large datasets. The SVM method is defined over a vector space where the classifier is used to find the decision surface that separates the data into two classes[43,44]. The essential point of an SVM classifier is the idea of margin maximisation[13]. In the case of linear separable data, the SVM computes a hyperplane that maximises the margin between two classes, whereas in the case of separable nonlinear data, the SVM computes a soft maximum margin that separates the hyperplane. Thus, given direction w of the hyperplane and d the position in space, the linear SVM is then defined through eq. (3).

$$f(x) = w^T x + d \quad (3)$$

Then the region between the hyperplane $w^T x + d = 1$ and $w^T x + d = -1$ that separate two classes called the margin. The width of the margin is equal to $2/\sqrt{w^T w}$. Maximization of the margin can be solved by equ(4)

$$\min\left\{\frac{1}{2} w^T w + k\sum_{i=1}^{p} \varepsilon_i^2\right\} \quad (4)$$

Which then subject to $y_i(w^T x_i + d) \geq 1 - \varepsilon_i$ and $y_i(w^T x_i + d) \geq 1 + \varepsilon_i$, where i=1 and $\varepsilon \geq 0$. Thus given the training data set$\{x_i, y_j\}_{, i=1,2,---p,}$ where $x_i \in R^n$ are the training document values and y_i are the class labels, p is the number of samples and 'n' is the number of features in each samples.

3.3.4 Naïve Bayes

Is a probabilistic classifier that is most commonly used in text classification. The algorithm assumes a particular generative model for classifying a text. It considers the conditional probability of the document terms *t* and their categories c_i to calculate the probabilities of the terms that

belongs to a document d_j by considering the Bayes rule. Thus, the probability of test patent document d_j being for category c_i is given in eq. (5). A detailed explanation can be found in[45,39].

$$P(d_j \mid c_i) = \prod_{k=1}^{N} \frac{P(w_{jk} \mid c_i)}{P(c_i)} \quad (5)$$

$P(d_j \mid c_i)$ is the probability of patent document d_j belonging to category c_i. $P(w_{jk} \mid c_i)$ corresponds to the probability of term t_k of patent document d_j belonging to category c_i ,and n represents the number of terms that belongs to document d_j and the category c_i.

3.5 Performance Evaluation

In this paper, we used the precision (P), recall (R), and F-measure (Fm) to evaluate the performance of the patent categorization classifier. P is the ratio of the number of correct categorization of the patent document to the total no of predictions. R is the ratio of correct classification of patent document into categories to the total number of labeled data in the test set[39]. Fm indicates the harmonic mean of P and R. Thus, the equations for P, R, and Fm are as follows:

$$P = \frac{TP}{TP + FP} \quad (6)$$

$$R = \frac{TP}{TP + FN} \quad (7)$$

$$F = \frac{2..P..R}{P + R} \quad (8)$$

4. Experiments

In this section, we evaluate the performance of the proposed method empirically. We first explain the dataset considered in our experiments. Next, we illustrate the classification performance of the RF, SVM, and NB algorithms under different parameter settings. During this process, we studied how different-sized feature vectors affect the accuracy of the classification. We also considered a validation method for calculating the accuracy of the patent categorization in a reduced feature vector space. Finally, we analysed the time complexity of the algorithms to perform the classification on the constructed data.

4.1 Datasets

The proposed patent document categorization system considers Google patents(http : //www.google.com/ patents) originating from the United States Patent and Trademark Office, the European Patent Office, and the World Intellectual Property Organisation[46,47]. The system collects patent documents from the year 1780 for US patents, and for EPO and WIPO, it collects documents from 1978. The Google Patent search engine operates continuously in retrieving both newly filed and published patents. Over eight-million documents are available, and the search engine is continuously growing with files from patent organisations such as USPTO, EUO Patent, and WIPO. In our current study, we consider only patents that have been published and submitted from the USPTO and EUO patents, and exclude patents filed from other countries. Ten different topics were used for the categorization process. These include topics from computer science and health-related areas. The topics taken from computer science files include MapReduce, cloud computing, semantics, context awareness, databases, virtualisation, clustering, and data mining, whereas in the case of the healthcare domain, ECG and mammogram related patents are selected. A total of 1,040 patents have been selected for the experiment. These documents are the resultant of the search query. From each document, we extracted the text from the semantic structural information automatically. This structure includes the <title>, <abstract>, and <claim>, and the first 200 words collected from the <description>after performing stop-words removal and lemmatization. We used only 200 words from the <description>due to the fact that patent examiners rarely focus much attention on the <description>part . The feature attributes and characteristics of the ten different topics are listed in Table.1 and Table.2. Table.2 shows the datasets, the number of documents presented in each topic, number of words and the number of unique words that are present in each topic. These sets are constructed according to the feature sets shown in Table.1.

To evaluate the performance of the proposed system, we constructed a series of evaluations under different parameter settings. For this purpose, we used ten different terms from the constructed dataset for the categorization. This includes database, ECG, context, clustering, mammogram, mining, semantic, cloud, virtualisation, and Mapreduce. The difference between topics and terms in our approach is as follows: topics are the general keywords that are used to retrieve the patent documents. A topic can be a word or a phrase. However, for the classification, we applied only single terms rather than topics for greater effectiveness in constructing the training and classification dataset. In addition, the lemmatisation process carried

out is based on morphological operators and a dictionary based approach that lemmatises in a single word-by-word manner rather than by multiple words or sentences. During this process, the words in the documents will start the lemmatisation process, and certain words such as clustering, mining, and virtualisation will be processed as is rather than in the root form during lemmatisation. The experimental evaluation consists of four different types of feature sets. In Table.1, the first set consists of features constructed from the <title>and <abstract>, the second set consists of the <title>, <abstract>, and <claim>. The third set contain features constructed from the <claim>, <abstract>, and <description>, and the final set contains features taken from the <title>, <abstract>, <claim>, and <description>from the collected patents. We evaluated the text categorization performance against three widely used classifiers: RF, SVM, and NB. We performed precision, recall and f-measure of proposed method for two different settings. In the first evaluation, we applied all features of four different sets, and performed the classification individually without a dimensionality reduction in the feature vector space. For the second case, we used IG-based feature reduction method and perform the classification. In both cases we evaluated the precision, recall and f-measure. To evaluate the success ration of the classifier in the reduced dimension, we ran the experiment with different sized feature vectors in the training dataset with an increased population of (20%,30%,40% and 50%) from the ranked index. All experiments were performed on a machine with a 3.40 GHz Intel Core i7-4770 CPU with 8 GB of RAM, a 2 TB HDD, and the 64-bit version of the Windows 7 operating system.

Table 1. Feature Attributes

Feature Sets	
F1	Title + Abstract
F2	Title + Abstract + Claim
F3	Claim + Abstract + Description
F4	Title + Abstract + Clam + Description

4.2 Overall Accuracy Evaluation

To measure the overall accuracy of the classifier RF, SVM, and NB were applied to four different data sets. Each of these sets contains term frequency of 1040 patent documents. The experiment using the RF-based classifier was conducted by setting the parameter for growing 10, 20, 30 and 40 sized trees. And presented the performance result for tree with minimum size 10 and maximum size 40 thus by excluding 20 and 30.The multiplicative factor was set using the values {0.5, 1, 2}. Table 3. shows the corresponding performance of the classifier in terms of precision, recall, and f-measure, whose average values were obtained by conducting the experiment on different tree sizes with different multiplicative factors. The results obtained after taking the average values of the multiplicative factor with respect to the number of trees are provided in Table 3. The experimental results from using the SVM and NB classifiers for the same dataset are also shown. As shown in Table 3., the f-measure for the SVM is considerable low compared to that for the NB. In addition, the f-measure value for the mammogram, mining, cloud, and virtualization categories is slightly better than the f-measure

Table 2. Dataset descriptions

Date Set	No.of docs	No. of words				No. of unique words			
		F1	F2	F3	F4	F1	F2	F3	F4
Database	105	6813	69836	90204	90836	827	1359	2686	2682
ECG	137	9322	70419	96830	97620	976	1545	2795	2791
Context-aware	105	6668	65838	85915	86554	592	938	1949	1952
Clustering	103	6283	62771	82834	83371	755	1149	2453	2458
Mammogram	117	8210	86573	109289	109973	713	1276	2276	2270
Data-mining	102	7446	74136	93890	94536	832	1336	2386	2387
Semantics	107	6801	72505	92837	93506	744	1198	2437	2441
Cloud-computing	98	6904	66726	85589	86127	693	1135	2016	2016
Virtualization	103	6737	74005	93748	94315	646	1045	1978	1977
Mapreduce	63	3765	37936	50033	50336	479	795	1504	1506

value obtained using the RF classifier. The f-measure values of the SVM and NB for the <title>, <abstract>and <claim>shown in Table.4 have varying measures compared to the <title>and <abstract>datasets. In this case, the f-measure value for the category mining, cloud and mapreduce related categories obtained an f-measure value of 1 for the SVM, whereas for all test data, the mapreduce category obtained an f-measure value of 1 for the SVM. The f-measure values for the dataset constructed from the patent sections such as <claim>, <abstract>, and <description>with the RF classifier are also provided in Table 5. The overall f-measure values for all ten different categories with a tree size of 40 are slightly better; in addition, the minimum f-measure value is 97, and the max f-measure value received for the mammogram category is 1. The best result for all cases is obtained by the SVM classifier, which is also shown in same table for a comparison with the NB and RF.

Table 6 shows the results for the dataset constructed from the patent sections such as the <title>, <abstract>, <claim>, and <description>. From Table 6, the f-measure values obtained for all categories after applying the RF algorithm have no change compared to the previous dataset, except for a slight improvement for the ECG category. The cases for the SVM and NB classifier are also shown in the same table. The results of each experiment show that, depending on the type of patent sections we are considering, the f-measurement value also varies for certain iterations. Furthermore, by considering all features for the classification, the f-measure value is almost similar when categorizing a patent using the three features, i.e.,<claim>,<abstract>, and <description>. From Table 3-6, it is clearly observable that a single term based categorization with patent information such as <title>and <abstract>perform better in precision(P) than recall(R) in RF and SVM. Where as in NB the recall have significant improvement than precision. This is because both precision and recall vary inversely. And this is similar to cases such as <title>, <abstract>, <claim>or<claim>,<abstract>, and <description>or <title>, <abstract>, <claim>,or<description>. The best scoring number for f-measure is shown in bold. The overall performances listed from Tables 3-6 verify the success rate of our proposed system achieved by creating a dataset with different semantic structural information.

4.3 Dimensionality Reduction Evaluation

Dimensionality reduction is an important aspect of feature selection. In our approach, we applied the most popular and powerful IG-based feature selection method to reduce the high dimensionality of the feature vector. To evaluate the performance of the classifier, we applied two step procedures. First we applied four different feature sets with an increased population in their feature vectors of size 20,30,40, and 50% repetitively from the IG based ranked index. Consecutively, in the second step we cross validated the classification accuracy with K-Fold method by choosing the value of K=5. Table 7 shows the f-measure value achieved from a dataset created from patent section such as <title>and <abstract>. And for each feature size we reduced the training samples to a fraction of n. We also observe how changing the value of mean and SD for the different features sets. These values show the average mean and SD for a specified dataset that contains

Table 3. The performance (averages over classes of P, R, Fm) of RF, SVM, NB classifier with <title>and <abstract>

	RF						SVM			NB		
	Tree=10,*			Tree=40,*								
Category	P	R	Fm	P	R	Fm	P	R	Fm	P	R	Fm
database	95.2	86.2	90.5	99.2	94.9	97.1	95.9	81.6	88.2	95.0	100	**97.4**
ecg	93.3	88.4	90.8	100	95.8	**97.8**	100	88.3	93.8	100	87.9	93.6
context	98.8	87.9	93.0	100	100	**100**	95.9	87.7	91.6	98.8	100	99.4
clustering	91.8	89.3	90.5	100	100	**100**	100	75.4	86.0	97.0	100	98.5
mammogram	94.3	93.0	93.6	100	95.8	97.8	93.2	87.3	90.2	96.9	100	**98.4**
mining	93.1	90.5	91.8	98.6	97.3	98.0	100	76.6	86.7	98.5	100	**99.2**
semantic	98.6	88.5	93.2	100	97.4	98.7	95.2	86.8	90.8	93.2	100	**96.5**
cloud	91.7	88.0	89.8	98.6	97.3	98.0	100	86.2	92.6	100	100	**100**
virtualization	92.4	84.7	88.4	97.3	98.6	97.9	100	86.6	92.8	98.5	100	**99.3**

mapreduce	93.3	82.4	87.5	100	100	**100**	100	76.9	87.0	100	100	**100**

*mtr [0.5,1,2]

Table 4. The performance (averages over classes of P, R, Fm) of RF, SVM, NB classifier with <title>, <abstract>and <claim>

	RF						**SVM**			**NB**		
	Tree=10,*			**Tree=40,***								
Category	P	R	Fm	P	R	Fm	P	R	Fm	P	R	Fm
database	93.1	92.5	92.9	99.0	98.0	**98.5**	86.8	75.4	80.7	80.5	100	89.2
ecg	91.6	90.6	91.1	96.9	97.9	**97.4**	100	86.4	92.7	96.6	94.9	95.7
context	93.5	91.7	92.6	100	97.2	98.6	79.5	76.1	77.8	98.0	100	**99.0**
clustering	93.5	90.5	92.0	98.9	97.9	98.4	90.2	76.7	82.9	97.6	100	**98.8**
mammogram	90.7	90.7	90.7	97.4	100	98.7	98.0	98.0	98.0	98.5	100	**99.3**
mining	93.2	86.1	89.5	100	98.7	**99.4**	100	100	**100**	95.7	100	97.8
semantic	94.9	89.3	92.0	100	97.6	**98.8**	82.2	86.0	84.1	90.4	100	94.9
cloud	95.9	89.9	92.0	100	100	**100**	100	100	**100**	93.8	100	96.8
virtualization	91.7	89.2	90.4	100	95.9	**97.9**	87.5	87.5	87.5	78.6	82.5	80.5
mapreduce	100	87.5	93.3	100	100	**100**	100	100	**100**	100	100	**100**

*mtr[0.5,1,2]

Table 5. The performance (averages over classes of P, R, Fm) of RF, SVM, NB classifier with <claim>, <abstract>and <description>

	RF						SVM			NB		
	Tree=10,*			Tree=40,*								
Category	P	R	Fm	P	R	Fm	P	R	Fm	P	R	Fm
database	95.5	96.4	95.9	99.5	99.8	**99.7**	99.8	99.8	99.8	71.5	100	83.4
ecg	95.0	94.1	94.5	100	98.0	98.6	100	98.4	**99.2**	100	90.0	94.7
context	94.6	97.0	95.8	99.7	99.7	99.7	100	100	**100**	60.9	100	75.7
clustering	95.8	89.6	92.6	99.3	98.7	99.0	100	100	**100**	60.4	93.3	73.4
mammogram	94.5	90.8	92.6	100	100	**100**	100	98.1	99.0	100	98.5	99.3
mining	93.9	87.0	90.3	100	96.7	98.3	100	99.4	**99.7**	39.0	100	56.1
semantic	92.0	86.7	89.3	98.3	95.8	97.0	100	100	**100**	23.8	100	38.5
cloud	92.9	91.2	92.0	100	99.1	99.6	100	100	**100**	73.9	97.7	84.2
virtualization	88.3	89.2	88.8	100	96.1	98.0	100	100	**100**	95.9	100	97.9
mapreduce	92.1	83.3	87.5	100	97.6	98.8	100	100	**100**	94.3	100	97.1

*mtr[0.5,1,2]

Table 6. The performance (averages over classes of P, R, Fm) of RF, SVM, NB classifier with <title>, <abstract>, <claim>, and <description>

	RF						SVM			NB		
	Tree=10,*			Tree=40,*								
Category	P	R	Fm	P	R	Fm	P	R	Fm	P	R	Fm
database	95.5	96.4	95.9	99.5	99.8	99.7	99.6	100	**99.8**	71.5	100	83.4
ecg	95.0	94.1	94.5	100	98.0	99.0	100	98.4	99.2	100	100	**100**

context	94.6	97.0	95.8	99.7	99.7	99.7	100	100	**100**	60.9	100	75.7
clustering	95.8	89.6	92.6	99.3	98.7	99.0	100	100	**100**	65.5	100	76.1
mammogram	94.5	90.8	92.6	100	100	**100**	100	98.1	99.0	100	98.5	99.3
mining	93.9	87.0	90.3	100	96.7	98.3	100	99.4	**99.7**	39.0	100	56.1
semantic	92.0	86.7	89.3	98.3	95.8	97.0	100	100	**100**	23.8	100	38.5
cloud	92.9	91.2	92.0	100	99.1	99.6	100	100	**100**	73.9	97.7	84.2
virtualization	88.3	89.2	88.8	100	96.1	98.0	100	100	**100**	95.9	100	97.8
mapreduce	92.1	83.3	87.5	100	97.6	98.8	100	100	**100**	94.3	100	97.1

*mtr[0.5,1,2]

Table 7. Comparison of F-Measures (Mean ± SD) for <title>and <abstract>

Category	Algorithm	No.of Features			
		20%	30%	40%	50%
database	RF	0.9542 ± 0.0032	0.9506 ± 0.010	**0.9578 ± 0.0022**	**0.9738 ± 0.0041**
	SVM	0.7734 ± 0.0379	0.5798 ± 0.0418	0.4504 ± 0.0458	0.3176 ± 0.0382
	NB	**0.9821 ± 0.0130**	**0.9752 ± 0.0071**	0.7481 ± 0.1297	0.5706 ± 0.0063
ecg	RF	0.9132 ± 0.0158	**0.9662 ± 0.0050**	**0.9844 ± 0.0015**	**0.9784 ± 0.0005**
	SVM	0.7854 ± 0.0281	0.5804 ± 0.0347	0.4028 ± 0.0837	0.3062 ± 0.0621
	NB	**0.9206 ± 0.0067**	0.8534 ± 0.0043	0.8494 ± 0.0058	0.4768 ± 0.0022
context	RF	0.9302 ± 0.0011	**0.9758 ± 0.0032**	**0.9824 ± 0.0009**	**1.0000 ± 0.0000**
	SVM	0.7526 ± 0.0337	0.4968 ± 0.0841	0.3678 ± 0.0582	0.2480 ± 0.0651
	NB	**0.9932 ± 0.0004**	0.9627 ± 0.0012	0.7350 ± 0.0071	0.5936 ± 0.0019
clustering	RF	**0.9052 ± 0.0011**	**0.9572 ± 0.0013**	**0.9878 ± 0.0016**	**1.0000 ± 0.0000**
	SVM	0.7156 ± 0.0291	0.5461 ± 0.0346	0.3786 ± 0.0640	0.3072 ± 0.0435
	NB	0.8498 ± 0.0118	0.5246 ± 0.0071	0.5234 ± 0.0052	0.5234 ± 0.0052
mammogram	RF	0.9358 ± 0.0011	0.9642 ± 0.0018	**0.9648 ± 0.0011**	**0.9774 ± 0.0019**
	SVM	0.7281 ± 0.0374	0.5822 ± 0.0617	0.3922 ± 0.0577	0.2556 ± 0.0218
	NB	**0.9924 ± 0.0019**	**0.9826 ± 0.0036**	0.9516 ± 0.0013	0.8622 ± 0.0039
mining	RF	**0.9182 ± 0.0011**	0.9648 ± 0.0004	**0.9778 ± 0.0022**	**0.9812 ± 0.0016**
	SVM	0.7618 ± 0.0268	0.5461 ± 0.0391	0.4674 ± 0.0371	0.2958 ± 0.0623
	NB	0.8732 ± 0.2544	**0.9742 ± 0.0058**	0.8722 ± 0.0044	0.4166 ± 0.0032
semantic	RF	**0.9356 ± 0.0033**	**0.9662 ± 0.0013**	**0.9674 ± 0.0005**	**0.9869 ± 0.0004**
	SVM	0.7494 ± 0.0291	0.5478 ± 0.0512	0.3332 ± 0.0359	0.2738 ± 0.0459
	NB	0.8242 ± 0.0024	0.7448 ± 0.0035	0.6772 ± 0.0104	0.4542 ± 0.0041
cloud	RF	0.8972 ± 0.0013	**0.9534 ± 0.0022**	**0.9588 ± 0.0011**	**0.9822 ± 0.0029**
	SVM	0.8098 ± 0.0410	0.5634 ± 0.0632	0.4663 ± 0.0260	0.2936 ± 0.0594
	NB	**0.9234 ± 0.0005**	0.8052 ± 0.0044	0.6510 ± 0.0022	0.5121 ± 0.0277
virtualization	RF	0.8844 ± 0.0009	0.9532 ± 0.0044	**0.9738 ± 0.0031**	**0.9778 ± 0.0018**
	SVM	0.7778 ± 0.0768	0.5512 ± 0.0210	0.3632 ± 0.0358	0.1938 ± 0.0472
	NB	**0.9746 ± 0.0014**	**0.9541 ± 0.0031**	0.8844 ± 0.0044	0.7546 ± 0.0149
mapreduce	RF	0.8742 ± 0.0013	0.9369 ± 0.0026	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**
	SVM	0.8592 ± 0.0459	0.6596 ± 0.0463	0.3861 ± 0.0934	0.2962 ± 0.0843
	NB	**0.9946 ± 0.0074**	**0.9814 ± 0.0255**	0.8248 ± 0.0091	0.6726 ± 0.0041

Table 8. Comparison of F-Measures (Mean ± SD) for <title>, <abstract>and <claim>

Category	Algorithm	No.of Features			
database		20%	30%	40%	50%
	RF	0.9254 ± 0.0005	0.9672 ± 0.0016	**0.9834 ± 0.0047**	**0.9854 ± 0.0009**
	SVM	**1.0000 ± 0.0000**	**0.9984 ± 0.0052**	0.9812 ± 0.0061	0.9462 ± 0.0132
	NB	0.9356 ± 0.0008	0.9376 ± 0.0425	0.8912 ± 0.0250	0.9136 ± 0.0307
ecg	RF	0.9232 ± 0.0061	**0.9742 ± 0.0030**	0.9738 ± 0.0020	0.9856 ± 0.0017
	SVM	0.9778 ± 0.0496	0.9600 ± 0.0894	**0.9666 ± 0.0179**	0.9674 ± 0.0177
	NB	**1.0000 ± 0.0000**	0.9728 ± 0.0333	0.9592 ± 0.0333	**1.0000 ± 0.0000**
context	RF	0.9250 ± 0.0173	0.9606 ± 0.0031	0.9676 ± 0.0013	0.9846 ± 0.0013
	SVM	**1.0000 ± 0.0000**	**0.9956 ± 0.0032**	**0.9824 ± 0.0140**	**0.9974 ± 0.0058**
	NB	0.9876 ± 0.0008	0.9792 ± 0.0049	0.9858 ± 0.0049	0.9726 ± 0.0098
clustering	RF	0.9112 ± 0.0011	0.9416 ± 0.0013	0.9626 ± 0.0005	0.9742 ± 0.0004
	SVM	**1.0000 ± 0.0000**	**0.9988 ± 0.0027**	**0.9820 ± 0.0205**	0.9694 ± 0.0191
	NB	0.9924 ± 0.0008	0.9868 ± 0.0036	0.9852 ± 0.0052	**0.9854 ± 0.0013**
mammogram	RF	0.9214 ± 0.0026	0.9678 ± 0.0018	0.9824 ± 0.0005	**0.9884 ± 0.0009**
	SVM	0.9582 ± 0.0540	0.9148 ± 0.1196	0.9666 ± 0.0171	0.9734 ± 0.0185
	NB	**1.0000 ± 0.0000**	**0.9832 ± 0.0094**	**0.9730 ± 0.0092**	0.9788 ± 0.0004
mining	RF	0.9076 ± 0.0019	0.9832 ± 0.0044	0.9866 ± 0.0009	**0.9896 ± 0.0053**
	SVM	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**	**0.9920 ± 0.0076**	0.9740 ± 0.0067
	NB	0.9768 ± 0.0043	0.9672 ± 0.0085	0.9778 ± 0.0033	0.9574 ± 0.0023
semantic	RF	0.9114 ± 0.0389	0.9628 ± 0.0029	0.9738 ± 0.0025	0.9864 ± 0.0176
	SVM	0.9422 ± 0.0235	0.9326 ± 0.0701	**0.9856 ± 0.0123**	**0.9890 ± 0.0111**
	NB	**0.9830 ± 0.0014**	**0.9788 ± 0.0091**	0.9784 ± 0.0077	0.9720 ± 0.0067
cloud	RF	0.9276 ± 0.0005	0.9836 ± 0.0036	**0.9846 ± 0.0049**	**1.0000 ± 0.0000**
	SVM	**0.9934 ± 0.0026**	**0.9944 ± 0.0040**	0.9760 ± 0.0166	0.9826 ± 0.0122
	NB	0.9806 ± 0.0026	0.9944 ± 0.0040	0.9760 ± 0.0166	0.9826 ± 0.0122
virtualization	RF	0.9038 ± 0.0011	0.9584 ± 0.0005	0.9792 ± 0.0004	0.9802 ± 0.0016
	SVM	0.9600 ± 0.0894	0.9989 ± 0.0031	**0.9796 ± 0.0118**	0.9642 ± 0.0171
	NB	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**	0.9666 ± 0.0747	**1.0000 ± 0.0000**
mapreduce	RF	0.9336 ± 0.0026	0.9778 ± 0.0018	0.9802 ± 0.0016	**1.0000 ± 0.0000**
	SVM	0.9500 ± 0.1118	**1.0000 ± 0.0000**	0.9742 ± 0.0233	0.9846 ± 0.0213
	NB	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**

Table 9. Comparison of F-Measures (Mean ± SD) for <claim>, <abstract>and <description>

Category	Algorithm	No.of Features			
database		20%	30%	40%	50%
	RF	0.9644 ± 0.0050	**0.9872 ± 0.0023**	**0.9926 ± 0.0030**	**0.9960 ± 0.0020**
	SVM	**0.9924 ± 0.0032**	0.9850 ± 0.0033	0.9822 ± 0.0048	0.9728 ± 0.0045
	NB	0.9046 ± 0.0871	0.8418 ± 0.0017	0.8180 ± 0.0514	0.8358 ± 0.0116
ecg	RF	0.9502 ± 0.0100	0.9776 ± 0.0045	0.9868 ± 0.0030	**0.9950 ± 0.0019**
	SVM	**0.9796 ± 0.0080**	0.9736 ± 0.0050	0.9600 ± 0.0126	0.9272 ± 0.0164
	NB	0.9592 ± 0.0372	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**	0.9592 ± 0.0372

context	RF	0.9236 ± 0.0166	0.9614 ± 0.0111	**0.9868 ± 0.0046**	**0.9928 ± 0.0062**
	SVM	**0.9896 ± 0.0036**	**0.9866 ± 0.0044**	0.9786 ± 0.0031	0.9664 ± 0.0040
	NB	0.8002 ± 0.0394	0.7574 ± 0.0008	0.8056 ± 0.1087	0.7464 ± 0.0237
clustering	RF	0.9106 ± 0.0186	0.9512 ± 0.0033	**0.9886 ± 0.0100**	**0.9952 ± 0.0052**
	SVM	**0.9884 ± 0.0011**	**0.9804 ± 0.0069**	0.9686 ± 0.0095	0.9480 ± 0.0094
	NB	0.8536 ± 0.0033	0.8554 ± 0.0008	0.8344 ± 0.0483	0.8474 ± 0.0192
mammogram	RF	0.9142 ± 0.0284	0.9568 ± 0.0071	**0.9856 ± 0.0107**	**0.9858 ± 0.0065**
	SVM	0.9746 ± 0.0148	**0.9700 ± 0.0076**	0.9662 ± 0.0092	0.9514 ± 0.0135
	NB	**0.9608 ± 0.0277**	0.9608 ± 0.0004	0.9936 ± 0.0143	0.9670 ± 0.0017
mining	RF	0.9236 ± 0.0078	0.9698 ± 0.0131	0.9912 ± 0.0104	**0.9938 ± 0.0041**
	SVM	0.9896 ± 0.0050	**0.9786 ± 0.0071**	0.9708 ± 0.0037	0.9528 ± 0.0071
	NB	**1.0000 ± 0.0000**	0.9586 ± 0.0008	**1.0000 ± 0.0000**	0.9918 ± 0.0183
semantic	RF	0.9196 ± 0.0179	0.9582 ± 0.0193	0.9890 ± 0.0028	**0.9930 ± 0.0057**
	SVM	**0.9884 ± 0.0029**	0.9814 ± 0.0047	0.9598 ± 0.0139	0.9486 ± 0.0163
	NB	0.9600 ± 0.0894	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**	0.9418 ± 0.1301
cloud	RF	0.9222 ± 0.0265	0.9694 ± 0.0152	**0.9768 ± 0.0111**	**0.9914 ± 0.0074**
	SVM	**0.9904 ± 0.0067**	**0.9866 ± 0.0097**	0.9730 ± 0.0100	0.9502 ± 0.0078
	NB	0.8314 ± 0.2715	0.9868 ± 0.0004	0.9568 ± 0.0675	0.9568 ± 0.0675
virtualization	RF	0.9306 ± 0.0124	0.9638 ± 0.0129	**0.9892 ± 0.0058**	**0.9908 ± 0.0075**
	SVM	**0.9870 ± 0.0035**	**0.9804 ± 0.0047**	0.9688 ± 0.0122	0.9488 ± 0.0217
	NB	0.9714 ± 0.0103	0.9758 ± 0.0004	0.9714 ± 0.0103	0.9576 ± 0.0103
mapreduce	RF	0.9198 ± 0.0482	0.9496 ± 0.0156	**0.9832 ± 0.0066**	**0.9976 ± 0.0054**
	SVM	**0.9958 ± 0.0058**	**0.9956 ± 0.0060**	0.9824 ± 0.0100	0.9608 ± 0.0169
	NB	0.9636 ± 0.0033	0.9636 ± 0.0033	0.9390 ± 0.0151	0.9594 ± 0.0148

Table 10. Comparison of F-Measures (Mean ± SD) for<title>, <abstract>, <claim>and <description>

Category	Algorithm	No.of Features			
		20%	30%	40%	50%
database	RF	0.9586 ± 0.0027	0.9854 ± 0.0031	**0.9934 ± 0.0030**	**0.9972 ± 0.00314**
	SVM	**1.0000 ± 0.0000**	**0.9920 ± 0.0043**	0.9460 ± 0.0341	0.8584 ± 0.0770
	NB	0.9348 ± 0.0027	0.9358 ± 0.0011	0.9248 ± 0.0425	0.9356 ± 0.0009
ecg	RF	0.9474 ± 0.0056	0.9760 ± 0.0066	**0.9896 ± 0.0042**	0.9974 ± 0.0005
	SVM	**1.0000 ± 0.0000**	0.9904 ± 0.0215	0.9660 ± 0.0179	0.8472 ± 0.0334
	NB	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**	0.9728 ± 0.0372	**1.0000 ± 0.0000**
context	RF	0.9220 ± 0.0069	0.9694 ± 0.0126	**0.9874 ± 0.0029**	**0.9914 ± 0.0037**
	SVM	**1.0000 ± 0.0000**	**0.9948 ± 0.0036**	0.9678 ± 0.0233	0.8662 ± 0.0687
	NB	0.9862 ± 0.0027	0.9876 ± 0.0009	0.9858 ± 0.0049	0.9876 ± 0.0005

clustering	RF	0.9260 ± 0.0157	0.9674 ± 0.0130	**0.9844 ± 0.0133**	0.9838 ± 0.0011
	SVM	**1.0000 ± 0.0000**	**0.9968 ± 0.0046**	0.9484 ± 0.0323	0.8172 ± 0.0321
	NB	0.9926 ± 0.0005	0.9942 ± 0.0018	0.9888 ± 0.0052	**0.9938 ± 0.0011**
mammogram	RF	0.9100 ± 0.0139	0.9702 ± 0.0090	0.9828 ± 0.0053	0.9760± 0.0208
	SVM	0.9860 ± 0.0313	0.9924 ± 0.0170	0.9666 ± 0.0171	0.8742 ± 0.0091
	NB	**0.9916 ± 0.0188**	**1.0000 ± 0.0000**	**0.9958 ± 0.0094**	**0.9916 ± 0.0115**
mining	RF	0.9298 ± 0.0220	0.9762 ± 0.0071	**0.9868 ± 0.0131**	**0.9960 ± 0.0007**
	SVM	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**	0.9500 ± 0.0504	0.8444 ± 0.0837
	NB	0.9770 ± 0.0027	0.9806 ± 0.0042	0.9736 ± 0.0087	0.9736 ± 0.0087
semantic	RF	0.9180 ± 0.0226	0.9738 ± 0.0111	0.9770 ± 0.0057	**0.9844 ± 0.0041**
	SVM	0.9638 ± 0.0395	0.9578 ± 0.0424	0.9640 ± 0.0413	0.8204 ± 0.0309
	NB	**0.9844 ± 0.0005**	**0.9832 ± 0.0011**	**0.9810 ± 0.0067**	0.9810 ± 0.0067
cloud	RF	0.9254 ± 0.0281	0.9658 ± 0.0189	**0.9790 ± 0.0042**	**0.9928 ± 0.0041**
	SVM	**0.9854 ± 0.0202**	**0.9950 ± 0.0050**	0.9542 ± 0.0273	0.8540 ± 0.0675
	NB	0.9812 ± 0.0004	0.9822 ± 0.0018	0.9772 ± 0.0085	0.9772 ± 0.0085
virtualization	RF	0.9386 ± 0.0206	0.9706 ± 0.0142	0.9814 ± 0.0031	**1.0000 ± 0.0000**
	SVM	**1.0000 ± 0.0000**	0.9986 ± 0.0031	0.9716 ± 0.0175	0.8440 ± 0.0154
	NB	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**
mapreduce	RF	0.9212 ± 0.0480	0.9606 ± 0.0299	0.9828 ± 0.0191	0.9868 ± 0.0027
	SVM	**1.0000 ± 0.0000**	0.9960 ± 0.0089	0.9628 ± 0.0232	0.9390 ± 0.0413
	NB	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**

information taken from a patent structure. The best scoring number of features for a specific category is shown in bold. From this evaluation, one can easily observe the importance of IG in feature selection for the purpose of classification with a reduced dimensional dataset. Table 8 shows the f-measure for a dataset created from patent sections such as the<title>, <abstract>,and <claim>. It can be seen that as the structure of the information increases, the accuracy of scoring the f-measure value of different classification algorithm also improves slightly. This slight improvement is very important in patent categorization

Table 9 illustrates the f-measure value for a dataset created by considering the<claim>,<abstract>, and <description>. Table 10 shows the f-measure value for a dataset that considers the semantic structural information of a patent, such as the <title>,<abstract>, <claim>,or<description>. In summary the performance of categorization based on single term is applicable for patent categorization. This is due to the specific term selection and these terms correctly classify the patent document. The results are presented based on their average mean and SD after the cross validation. Overall, each of these evaluations demonstrates that IG is an important candidate for selecting a feature set for patent categorization.

5. Conclusion

In this paper, we proposed a patent document categorization system on a patent dataset provided by Google Patent. The evaluation is carried out in three stages. In the first stage, terms are extracted from the patent documents, then pre-processing stages are performed and finally applied to the classifier for the purpose of classification. The classification is performed on both the original feature sets and the dimensionality reduced feature sets. The efficiency of the terms in both methods is tested using three different classifiers RF, SVM, and NB. The experimental results and their accuracy are evaluated in terms of precision, recall, and f-measure. It was found that the semantic structural information of a patent document is an important feature set in constructing the terms of a document for the categorization process. The experimental results demonstrate that the classification results vary depending on the patent sections selected for the classification, namely (F1)<title>and <abstract>, (F2) <title>,

<abstract>and <claim>, (F3) <claim>,<abstract>, and <description>and (F4) <title>,<abstract>,<claim>and <description>.The results reveals that the structural information of patent based categorization is an efficient method for analyzing the patents. The experimental results also show that feature reduction using IG is beneficial for obtaining a higher accuracy rate in a reduced dimensional space. This paper mainly focuses on categorizing the patents of ten different topics that are related to computer science and health related areas. As a future work, we would like to extend this work with increased topic and compare it with other source of methods for selecting the patents and propose different approach to improve the patent categorization.

6. References

1. Ernst H. Patent information for strategic technology management. World Patent Information. 2003; 25(3):233–42.
2. Christopher M. The world intellectual property organization. New Political Economy. 2006; 11(3):435–45.
3. David D, Ringuette LM. A Comparison of Two Learning Algorithms for Text Categorization. 3rd Annual Symposium on DAIR. 1994; 81–93.
4. Yun J, Jing L, Yu J, Huang H. A multi-layer text classification framework based on two-level representation model. Expert Systems with Applications. 2012; 39(2):2035–46.
5. Gomez JC, Moens MF. A Survey of Automated Hierarchical Classification of Patents. Professional Search in the Modern World. 2014; 8830:215–49.
6. Uguz H. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowledge-based Systems. 2011; 24(7):1024–32.
7. Verberne S, Dhondt E. Patent classification experiments with the linguistic classification system LCS in CLEF-IP In Proceeding of: CLEF labs and workshop, notebook papers. 2011; 19–22.
8. Hotho A, Nrnberger A, Paa G. LDV Forum – GLDV. Journal for Computational Linguistics and Language Technology. 2005; 29(1):19–62.
9. Sebastiani F. Machine learning in automated text categorization. ACM Comput Surv. 2002; 34(1):1–47.
10. Breiman B, Leo L. Bagging Predictors. 1996; 24(2):123–40.
11. Breiman L, Friedman J, Olshen R, Stone C. Classification and regression Trees, Monteret CA, Wadsworth and Brooks, 1984.
12. Joachims T. Text categorization with support vector machines. Technical report, LS, University of Dortmund, 1997.
13. Vapnik V. The nature of statistical learning theory, Springer, 1999.
14. Leopold E, Kindermann J. Text Categorization with Support Vector Machines. How to Represent Texts in Input Space. 2002; 46(1-3):423–44.
15. Ghate VN, Dudul SV. SVM Based Fault Classification of Three Phase Induction Motor. Indian Journal of Science and Technology. 2009 Apr; 2(4):1–4.
16. David D, Lewis L, Naive N. at Forty: The Independence Assumption in Information Retrieval. ECML 'Proceedings of the 10th European Conference on Machine Learning. 1998; 98:4–15.
17. Bonino D, Ciaramella A, Fulvio Corno F. Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics. World Patent Information. 2010; 32(1):30–8.
18. Leah SL. A Patent Search and Classification System. Proceedings of the Fourth ACM Conference on Digital Libraries. 1999. p. 179–87.
19. Kim JH, Choi KS. Patent document categorization based on semantic structural information. Information Processing and Management. 2007; 43(5):1200–15.
20. Lupu M, Hanbury A. Patent Retrieval. Foundations and Trends in Information Retrieval. 2013; 7(1):1–97.
21. Yun Y, Yang Y, Akers L, Thomas Klose T, Barcelon C, Yang Y. Text mining and visualization tools Impressions of emerging capabilities. World Patent Information. 2008; 30(4):280–93.
22. Krier M, Zacc F. Automatic categorization applications at the European patent office. World Patent Information. 2002; 24(3):187–96.
23. Koster C, Seutter M, Beney J. Multi-classification of Patent Applications with Winnow. Proceedings PSI (Springer LNCS. 2003; 2890:545–54.
24. Adams S. The text, the full text and nothing but the text: Part 2 The main specification, searching challenges and survey of availability. World Patent Information. 2010; 32(2):120–8.
25. Wanner L, Baeza-Yates R, Codina SJ, Diallo B, EnricEscorsa E, Giereth M, Yiannis Kompatsiaris Y, Papadopoulos S, Emanuele Pianta Piella G, Ingo I, Puhlmann P, Rao G, Rotard M, PiaSchoester P, Serafini L, Vasiliki Zervaki V. Towards content-oriented patent document processing. World Patent Information. 2008; 30(1):21–33.
26. Fall CJ, Trcsvri A, Benzineb, Karetka G. Automated categorization in the international patent classification. SIGIR Forum. 2013; 37(1):10–25.
27. Lewis DD. Text categorization Test Collection, Distribution. 1997.
28. Benzineb K, Guyot J. Automated Patent Classification, in Current Challenges in Patent Information Retrieval. 2011; 239–61.
29. Vrochidis S, Papadopoulos S, Moumtzidou A, Panagiotis Sidiropoulos P, Emanuelle Pianta E, Ioannis Kompatsiaris I.

Towards content-based patent image retrieval: A framework perspective. World Patent Information. 2010; 32(2):94–106.
30. Piroi F, Lupu M, Hanbury A, Zenz V. Clef-ip Retrieval in the intellectual property domain. In CLEF (Notebook Papers/Labs/Workshop), Austria. 2011; 1–16.
31. Yiming Y, Pedersen Jan O. A Comparative Study on Feature Selection in Text Categorization, ICML '97, 412–20.
32. Apte C, Damerau F, Weiss S. Towards language independent automated learning of text categorization models. In proceedings of the 17th Annual ACM/SIGIR Conference, USA. 1994. p. 23–30.
33. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. 2011; 173(3):693–4.
34. Dasgupta A, Drineas P, Harb B, Josifovski V, Mahoney MW. Feature Selection Methods for Text Classification. Proc 13th Int', l Conf Knowledge Discovery and Data Mining (KDD ',07), USA. 2007. p. 230–9.
35. Zheng Z. Feature selection for text categorization on imbalanced data. ACM SIGKDD Explorations Newsletter. 2004; 6(1):80–9.
36. Shang S, Changxing C, Min L, Shengzhong F, Jiang J, Qingshan Q, Fan F, Jianping J. Feature Selection via Maximizing Global Information Gain for Text Classification. Knowledge-based Systems. 2013; 54:298–309.
37. Uysal AK, Gunal S. A novel probabilistic feature selection method for text classification. Knowledge-based Systems. 2012; 36:226–35.
38. Karabulut M. Fuzzy unordered rule induction algorithm in text categorization on top of geometric particle swarm optimization term selection. Knowledge-based Systems. 2013; 54:288–97.
39. Zhang J, Chen L, Guo G. Projected-prototype based classifier for text categorization. 2013; 49:179–89.
40. Breiman L, Friedman JH, Olshen RA. Charles J Stone, Classification and Regression Trees, CRC Press, New York, 1999.
41. Breiman L. Bagging Predictors, 1996; 24(2):123–40.
42. Joachims T. A statistical learning model of text classification for support vector machine. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, publishing new Orleans, Louisiana, US. 2001. p. 128–36.
43. 44. Shang C, Li , Feng S, Jiang Q, Fan J. Feature selection via maximizing global information gain for text classification. 2013; 54:298–309.
44. Nigam N, Kamal K, Callum M, Kachites A, Sebastian T, Mitchell M, Tom T. Text Classification from Labeled and Unlabeled Documents using EM. 2000; 39(2-3):103–34.
45. WIPO, Strasbourg agreement concerning the international patent classification, Legislative text WOO26EN. Available from: http://www.wipo.int/treaties/en/classification/strasbourg/, Date accessed:28/09/1979.
46. Xie Z, Miyazaki K. Evaluating the effectiveness of keyword search strategy for patent identification. World Patent Information. 2013; 35(1):20–30.