Patent Mining to Predict Class using Decision Tree and Naïve Bayes Algorithm

Darshana A Naik, Brunda C J

Abstract*: The number of patents that are being filed across the world is increasing day by day. With the increase in patents being filed the process of segregating the patents based on their class becomes even more difficult. There is no prior work that has been done to increase the efficiency of this process, therefore patent mining is done. There are a set of features that are extracted from the dataset that is previously present. The features that are being extracted will vary for each document and based on the feature that is extracted the following steps are carried out. After the feature extraction is done there are two steps that need to be carried out, namely: Classification and prediction. For this purpose, decision tree algorithm is used which makes use of the most prominent feature and classification is done using those features. Therefore, for classification a hierarchical decision tree algorithm is used along with the probability of patent conversion. Based on the classification that is done a model will be created and whenever a new entity is brought it is compared with the model file that was created using the available datasets and is predicted as a particular class. Thus, both classification of existing dataset and the prediction for any new dataset based on previous inputs can be achieved thereby facilitating the patent mining process.

Keywords—Classification; Prediction; Decision Tree; Naïve Bayes; Patent Mining; Feature Exracction; Attributes.

I. INTRODUCTION

Patent documents are very important for research purpose. It is therefore required to maintain these documents for the betterment of the society. There are several communities that will get benefitted by these policies making. Therefore, it is required to monitor these documents for better maintenance of these. Duplication of patents or overlapping can be easily avoided. Segregation of the patent data can help in classifying them and bringing under the same banner. So that there is not much of an impact. It is always required that more manpower with expertise is required for segregating these patents. This is where the patent mining can be of greater use. In the long run the patents can be easily segregated and could lead to more clarity in the new documents that are required to come. Patent analysis and mapping go hand in hand, the mapping is required to do the patent analysis as there is no use in doing either one of these. The problem here is that it takes more time for the patent analysts to retrieve a certain amount of information from the paper in order to segregate it into a particular class.

It is also not possible for the experts to have a multi-disciplinary expertise. They need to have knowledge in all the field that they review in in order to segregate it into the required class. To Rad the documents and then to analyze it at the same time will take more time and thus is time consuming. In order to reduce the amount of time that is spent in this type of analysis it is better to have a generalized methodology that can easily segregate all the data and thereby produce the necessary result. Therefore, the automated technologies can be used to assist the experts that analyse the system in order to increase the efficiency of the process and make the segregation process easier.

Typically, patent documents are evaluated based on the following parameters:

1. Identification of the task.

2. Iterative searching and filtration.

3.Segmentation Cleaning and Normalization

4.Abstracting and summarizing the documents.

5.Clustering based on the domain.

6.Visualization of the document.

7.Interpretation and correlation with modern trends.

II. PROBLEM STATEMENT

Patent documents are of high value and importance when it comes to the research fields. It is thereby required to maintain it properly. In order to segregate these documents, it is required to go through these documents. He patents documents will be lengthy normally and on top of which it will also require an higher level if technical expertise in the particular field as well. If the person that is reviewing the document does not have a particular level of expertise in the field, then it will be difficult for him to segregate the document.

Revised Manuscript Received on July 22, 2019.

Sneha V.V.
PG Scholar, Dept. of ECE, MEA Engineering College, Perinthalmanna
Ismayil Siyad C.
Assistant Professor, Dept. of ECE, MEA Engineering College, Perinthalmanna, India
S.Tamilselvan
Associate Professor, Pondicherry Engineering College, Puducherry, India
Email: tamilselvan@pec.edu.in, 17pmce08@meaec.edu.in, ismayilsiyad@meaec.edu.in

Even if the expert has enough knowledge in the field it will take a lot more time for him to segregate this and the chance of there being any human errors are also more. Therefore, here is a need to segregate all the data through an automated process that can be relied upon so that not only does the time taken for segregation reduce but also there will be a significant extent to which the bias in the process of evaluation of the document will reduce.

It is therefore required to prepare an automated process using a machine learning process so that efficiency of the system will increase. There needs to be an evaluation pattern that needs to be devised so that the evaluation process takes place much faster and in a more efficient way than how it is done by the experts. Domain wise analysis must be done and the mapping of these needs to be done accordingly as well. This classification system needs to be more efficient than the already existing and available classification techniques.

III. LITERATURE SURVEY

Marko Velic et al,(2013) had proposed a system in which the collaborative method can be used to analyse how the stock market is fluctuating. The implemtation of the collaborative filtering in th stock market shows that based on data analysis on the filtering process it is possible to increase the efficiency of the system. Several collaborative methods such as the wisdom of crowds can be used to initiat this process. There are several other methods such as web-based methods to evaluate such patterns. The mathematical model behind the solution is not exposed, but still can produce a certain amount of idea on the topic. The algorithmic model has been explaind in this paper.[1]‘

Gu Chengjian et al.(2009) have developed a patenting roadmap used for the analysis of e=several developments in technologies that is based on solid patent information. This paper also exposes some research done in the global level in the same field. Patent management map one of the widest used method has been adopted in this paper. Based on this paper it is also required to increase the efficiency of the system. It can also be used to increase the efficiency in the R and D sector of our nation. It also shows about a method of flat panel display that can be used in the future.[2]

Zhenbao Liu et al.(2018) There is a method that is used for reliability assessment and prediction in the field of product manufacturing that is used. Physics based mathematical modelling is used which can help in further development of several processes in the same field. This paper presents some key research based on the various other parameters that need to be handled and addressed. The number of patents that are used in several other parts are also noted and the necessary changes are done.[3]

Hongshu Chen et al.(2018) This paper explains about changing various literature contents of paper into technical insights. Although much effort is being put in technological trends a regular monitoring of these systems needs to be done and the processes need to be carried out. A new time series framework used in the paper can be used for the same processes and thereby will increase the efficiency of the system drastically. Technological intelligence will therefore enable in predicting the technological datapoints in a futuristic manner. This will overcome the problem in any futuristic predictions that is done.[4]

Hayley Beltz et al,(2017) The network of patents need to be used to increase the detection citations that are used in papers is done using this method. Whenever a patent cites another patent, it simply means that there is an overlap in the content of the patents. Thus, based on the aforementioned method this overlapping detail in the patents can be easily found out. A reinforcement learning algorithm is used to create a rank-based setup in order to increase the efficiency of the system. By combining regression with other several clustering methods it is evidently found out that there is an improvement in the detection. Some patterns were also found out and were exposed in this paper.[5]

Xiang Ji et al,(2017) In this paper a collaborative filtering recommendation approach is used in order to attain similarity between these documents that are used as patents. The problem of data sparsity has reduced to a greater extent by using the methodology carried out in this paper. Preliminary level analysis has showed that the efficiency has increase and also similarity index in between patents have been used in order to facilitate the process. By using these methods, the patent utilization can be increased considerable over a long period of time. [6]

Xin Jin et al.(2011) In businesses patents are of high value as they will provide the legal information to properties and therefore help in creating a lot of value for the system. Patents provide legal protection for any company that owns the patents. This will help in increasing the extent to which it is enforceable on the people. If the patent is not deemed to be valuable also it can be made use of in order to reduce the loss to the company, I terms of securing the capital spent on the maintenance of the patent. Therefore, benefitting the company to a greater extent. Thus, this paper explains the need for proper maintenance of patents in the line of business.[7]

IV. KNOWLEDGE DELIVERY PROCESS

The knowledge discovery has some process to lead the data. It is mandatory that these processes are carried out in order to get the required output. The several techniques involved are as follows: data selection, data transformation, data cleaning, data integration, data mining, pattern evaluation and knowledge representation.

A. Data cleaning

Any set of data will have certain amount of information that is not required for the training and classification processes. Therefore, there is a need to refine the data and only extract the required information. This process is facilitated by data cleaning and will enable in making the upcoming steps easier.

B. Data integration

The entire dataset is taken from several sources in order to get good input dataset. If there isn't substantial amount of data present the training process will become difficult. All the data taken from several sources has to be integrated together in order to get uniform information.

C. Data selection

Of all the data that is refined the required data needs to be selected and only this dataset is used for further processing. This actually is a second step of dataset refining.

D. Data transformation

The dataset will be present from several sources and will have various other features and attributes that are not required. Therefore, the data is transformed into the necessary format in order to make the further steps easier.

E. Data mining

Data mining is the step in which the data is segregated into the various classes in which it is required. The data mining will help in further prediction of the new data that are presented.

F. Pattern evaluation

There is prone to be a pattern that is observed in the classification of data. In this process the pattern is identified and the required classification can be done. This will help in comparing with the pre-existing dataset.

G. Knowledge representation

This is the final step in which the knowledge is discovered from the selected data for the user with visualization so that the process of getting the required data does not become a tedious process.

V. CLASSIFICATION USING DECISION TREE ALGORITHM

Decision tree algorithm comes under the banner of supervised learning algorithms. The decision tree algorithm is easier to comprehend in comparison with other learning algorithms. The main motive of a decision tree is to create a set of learning rules that are created from the previous datasets using which it can predict the target class into which the the data that is presented will fall. Decision tree algorithm solves using a tree like representation where each node represents an attribute. These attributes can be taken from the dataset.

In this case we make use of decision tree algorithm to classify into the respective classes based on the set of attributes present in the dataset. Each internal node will be represented by an attribute that is present in the dataset. Using this the decision tree algorithm will classify and produce the class into which that particular patent will fall. The patents are classified based on the learning rules that are generated by the algorithm which can also be used for further prediction of data.

VI. PREDICTION USING NAÏVE BAYES ALGORITHM

This algorithm is supervised learning method, it's a collection of various probabilistic classifiers that uses Bayes theorem presuming that the features are independent of each other. It as packages called sklearn which contain methods for Naïve Bayes. The dataset is divided into two clusters-training data and testing data. Training data serves as basis for creating training model and this provide knowledge for naïve Bayes algorithm. The patent records are tagged with respective category and help the training model to form a rule by analyzing this association. The testing dataset is used for training purpose.

In this case the naïve Bayes algorithm would be applied on the training dataset to develop a training model and tested against the testing dataset. Predicted class_code is a parameter that is called as predicted category code of the given patent. Numerical format is used for this purpose.

VII. ASSOCIATION RULE MINING WITH CENTRALITY

The traditional clustering methods rely only on similarity (or distance) measures, which can disregard important characteristics of the database. The use of complex networks to find groups can include important features on the clustering process, such as the elements position and the data density.

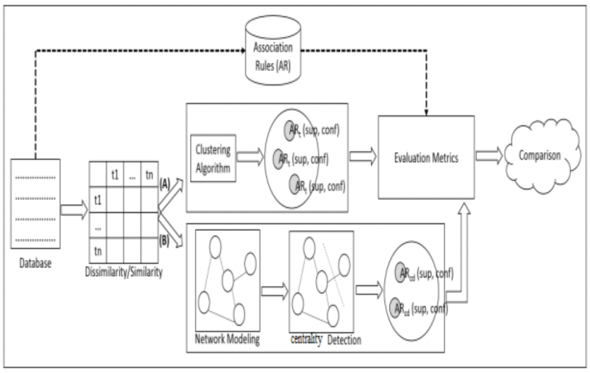


Figure 1: Assessment Methodology

Based on this, this paper proposes an analysis on centrality detection algorithms, which uses the network structure to find groups, to aid the association rule mining process. The proposed analysis is illustrated in Figure 1. The analysis consists of comparing the results obtained by the traditional clustering methods with the results obtained by the centrality detection algorithms. Therefore, some metrics proposed in [6] are used (described below). To compute the metrics, each process is executed: the traditional (A) and the process that uses centrality detection (B). First, the original association rules set (AR) is mined from the databases. This set is used to compute the evaluation metrics (see below). Then, in all the databases, the similarity among all the transactions is calculated. After that, in (A), a traditional clustering algorithm is applied. Inside each obtained group, an AR extraction algorithm is executed, generating groups of rules. All of these rules form the clustered AR set (ARcl).

In (B), based on the computed similarities, the data is modeled through a network – vertices are transactions and edges are the computed similarity among them. Then, a centrality detection algorithm is applied. Inside each obtained group, an AR extraction algorithm is executed, generating groups of rules. All of these rules form the centrality detection AR set (ARcd). Considering the obtained AR sets, the evaluation metrics are computed and the comparison done. Some of the proposed metrics evaluate the benefits obtained from the data grouping, which means that some measures needs to consider the original association rule set (AR) and an association rule set generated after the grouping (ARcl or ARcd). Besides, the authors use the concept of h-top best rules, that consists of selecting the h% best rules, based on an objective measure, from the AR set and the h% best rules from the grouped set that will be analyzed (ARcl or ARcd).

These are the rules that are considered as the most interesting to the user according to the objective measure that was chosen. In this work, h was set to 1% of the total of the rules and the objective measure used was Lift.

VIII. PROPOSED SYSTEM

Several analyses were done on patent mining with the techniques of data mining. The proposed system makes use of a filtration algorithm for classification and decision tree algorithm is used for prediction. It is required to create a dashboard for fetching data and preprocessing. The preprocessing module includes the cleaning, feature extraction, selection and finally training set.

In the proposed system there is a need do classification and prediction of patent mining. Before proceeding to classification and prediction, we make use of priority attributes in order to check the association frequency and also the check patent is granted or not.

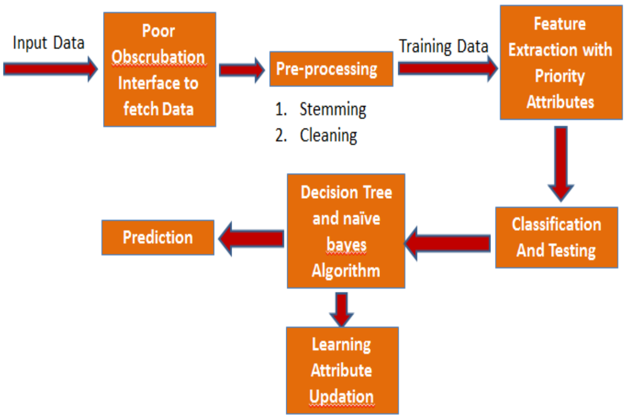

Input Data
Poor Obscrubation Interface to fetch Data
Pre-processing
1. Stemming
2. Cleaning
Training Data
Feature Extraction with Priority Attributes
Classification And Testing
Decision Tree and naïve bayes Algorithm
Prediction
Learning Attribute Updation

Figure 2: Classification done using Neural Filtration Algorithm

In experimental analysis, graphical user interface needs to be created for loading data, preprocessing, classification and prediction of patent mining. Firstly, it is required to fetch the data and preprocess the data and then check the association and centrality with respect to country. In classification it is required to check whether the country is associated with patent filed or not, the filtration algorithm is used to find best patent in top countries among all countries. In prediction a new tuple is entered which is decisively compared with target tuple used for predicting into which group it will fall into.

A. Fetching Data

In order to extract the required information for the input dataset a poor obstruction interface is used. Using this only the required information will be updated extracted from the existing dataset thereby making the dataset refining process much easier.

B. Pre-processing

Pre-processing is a set of processes that need to be carried out only once in a system. Once the pre-processing step is done it is not required to carry out this step repeatedly. The pre-processing steps that are carried out are stemming and cleaning respectively. This will help in facilitating in making the extraction process easier.

C. Feature Extraction

For the process of feature extraction, a certain set of attributes can be used and these attributes are set as priority attributes. Feature extraction is done accordingly keeping in mind the several other priority attributes.

D. Classification and Testing

After the required features are extracted the classification of these features need to be done. The classification of respective classes using the attributes is done using Neural Filtration Algorithm.

E. Prediction using Decision Tree Algorithm

A decision tree algorithm is used in facilitating the prediction process of the respective algorithms. Based on the classification of the previously available dataset, whenever there is a new tuple that is presented it can be predicted and slotted into the same class.

F. Learning Attribute Updating

Whenever a new dataset is brought into the picture that will change the existing dataset and the learning attribute will also change. As the learning attribute changes there is a need to update the learning attributes. Thus, the learning attribute updating needs to be done after each cycle.

IX. RESULT AND CONCLUSION

A user interface has been created for user interaction. It results in fetching data, preprocessing, classification and prediction. The data is collected and pre-processed with response to the requirements of the project. Then check for the centrality, which uses the network structure to find groups, to aid the association rule mining process.

The below shown figure 3 shows the home page of the user interface. In this all the necessary processes are represented in the same screen. Figure 4 shows the classification process that is done after patent mining and collaboration. In this figure there is a bar graph representation of the classification that is done. The final result will be displayed in a window along with the domain it will belong to. A sample of this predicted result is shown in figure 5.

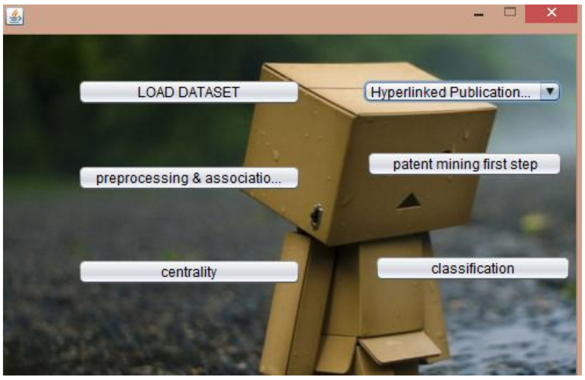


Figure 3: User Interface

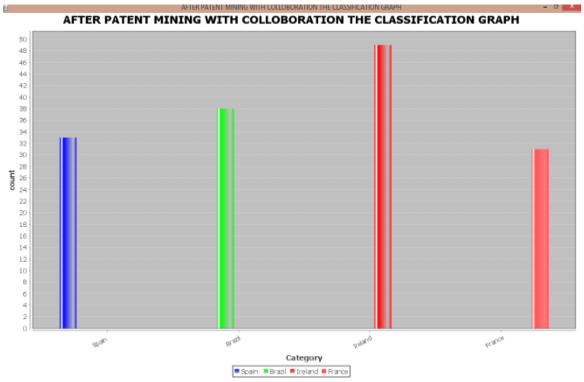
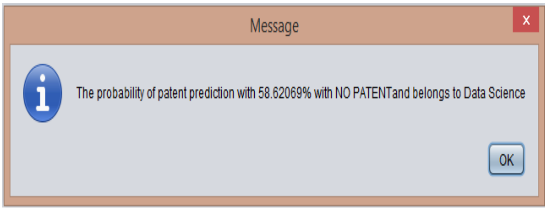


Figure 4: Sample's Classification Process



The probability of patent prediction with 58.62069% with NO PATENTand belongs to Data Science

Figure 5: Sample Predicted Output

Patent classification along with prediction is an important process as it not only helps in management but also in further maintenance of the existing patent documents. Every year it is obvious that the number of patent documents keeps increasing at a very rapid rate which also demands for a more sophisticated system that would help in automatically categorizing the patents.

In this proposed system, a decision tree algorithm is exposed and is implemented for the patent classification along with prediction. Naïve Bayes algorithm is used for predicting the data in case of any new instance. Here we can say that each tuple present in the dataset is a patent, and will also determine if this patent can also be extended to other existing ones.

(1) REFERENCES

[1] Marko Velic ; Toni Grzinic ; Ivan Padavic ," Wisdom of Crowds Algorithm for Stock Market Predictions" Information Technology Interfaces, 27 June 2013.

[2] Gu Chengjian ; Huang Lucheng," The study on CNT-FED for emerging technology forecasting by using patent Management Map", International Symposium on Information Engineering and Electronic Commerce,2009.

[3] Zhenbao Liu ; Zhen Jia ; Chi-Man Vong ; Junwei Han ; Chenggang Yan ; Michael Pecht," A Patent Analysis of Prognostics and Health Management (PHM) Innovations for Electrical Systems",IEEE,2018.

[4] Hongshu Chen ; Guangquan Zhang ; Jie Lu," A Time-series-based Technology Intelligence Framework by Trend Prediction Functionality",IEEE International Conference on Systems, Man, and Cybernetics,2013.

[5] Hayley Beltz ; Raoul R. Wadhwa ; Peter Erdi," From ranking and clustering of evolving networks to patent citation analysis",IEEE,2017.

[6] Xiang Ji ; Xinjian Gu ; Feng Dai ; Jixi Chen ; Chengyi Le," Patent Collaborative Filtering Recommendation Approach Based on Patent Similarity",Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD),2011.

[7] Xin Jin ; Scott Spangler ; Ying Chen ; Keke Cai ; Rui Ma ; Li Zhang ; Xian Wu ; Jiawei Han," Patent Maintenance Recommendation with Patent Information Network Model", IEEE International Conference on Data Minin