International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-2 Issue-4, April 2013

# Classification of Patents by Using the Text Mining Approach Based On PCA and Logistics

**Manmeet Kaur, Richa Sapra**

***Abstract-Analysis of patent data is important tool for industrial research. Patent analysis has been used in many research fields and applied for rich topics in technology management. Patents are often used as the source of inspiration for new ideas. Patents contain detailed technical information about technical problem and the preferred technical solution. This information can be used for example to assess the state of the art or as a basis to identify possible gaps in a technology field. But often it is a very time consuming process to analyze the information provided by patents, because huge amounts of patents have to be considered. This paper proposes an intelligent system for classification based on Principle component analysis (PCA) and logistics. The intelligent system is designed to extract the features from the patents database and classify them according to the predefined categories as software, biological, business and chemical. Three different stages are designed to classify the content of patents such as (1) text pre-processing (2) PCA based features extraction and (3) classification using logistics. The main advantage of this approach is that the user need not to read whole patent documents but able to retrieve the relevant parts of the text in short time for further analysis process.***

***Keywords - Classification, Data mining, Logistics, PCA, Text mining.***

## I. INTRODUCTION

Patent information has an increasing value for companies during the whole technology management process starting from technology creation up to technology usage [10]. Thus for a variety of strategic business decision patent data analyses can be necessary [1] [2]. In today's competitive environment technology has become the most important weapon of the enterprises. Acquiring competitive advantages can only be succeeded through management of innovation and technology. The rapid changes in the technology have also transformed the structure of competition in business world. With the change in technology, more opportunities are created to invest.

A deeper understanding of technological change has been an essential need to avoid unnecessary investment and beyond to find promising investments. Thus, understanding technology and its evolution over time, forecasting and tracking technology has become extremely important for managing the technology. In this regard, different source of data and their processed form are employed to manage these important problems.

**Manuscript published on 30 April 2013.**
***** Correspondence Author (s)
**Manmeet Kaur,** Computer Science, Lovely Professional University, India
**Richa Sapra,** Computer Science, Lovely Professional University, India.



Therefore, intelligent selection of the information source along with valid framework has been an essential task to reduce the failure risk of wrong technology selection and investment. Using a valid framework has also been a very crucial part of technology selection management.

In case of right framework and data resource selection, the gathered and processed data through a framework can be used to formulate a technology vision and strategy.

Patents have been one of those important and robust data resources where they are the documents which protect an inventor's invention by a particular given monopoly, so that others can't duplicate and commercialize it. Patents documents enclose an archive with millions of papers. These papers witness the progress of technologies through the history. Therefore, patents documents are one of the most valuable and rich technology information resources.

It is well known that the patents are used to protect the innovative ideas during research and development. And it is well known that patents are not only used by companies for reasons of protection, they are also very powerful tool to generate new ideas and solve technical problems. So the analysis of patents information is also very important for various patent analysis reasons [11] [12] [13]. However in order to develop such a text-classification system, many researchers have devoted their work for automating the text classification task. A patent categorization system is developed where a rule base is generated by human expertise. Building such a system takes more than couple of months because of the enormous number of rules.

On the other hand, a statistical approach based on keywords extraction from training texts is popular method of generated a knowledge base [14]. This has the advantage that the knowledge base can be quickly generated without much cost. However, we need guidelines on how to gather a large quantity of training texts. As for automated text categorization, our results are interesting. It shows that relatively knowledge-poor machine learning algorithm outperforms human beings in a text classification task. This suggests that automated text categorization are reaching a level of performance in which they can compete with humans not only in terms of cost-effectiveness and speed, but also in terms of accuracy of classification

This rest of the paper is organized as follows section 2 describe related works, section 3 provides the proposed methodology, section 4 displays the experiment work and conclusion and future work.

## II. RELATED WORKS

### *2.1 Automated text categorization*

In the recent years a variety of software solutions have been developed for different applications, ranging from special patent search tools to complex patent analysis tools providing for example with TRIZ [4], the theory of problem solving, the inventor is able to systematically analyze his problem and find out appropriate solutions build upon the experience of former successful inventors [3] [6]. Another example of how patents can be used for idea generation processes is the White Spot Analysis of Fraunhofer IAO [5]. As the patent databases world –wide grow continuously, there is a growing need for software solution assisting the user is a handle the patent analyses, because the analysis of hundreds of patents is very complex and time consuming. For methods like TRIZ special software tools have been already been developed. Also for bibliographical or citation analyses various tools have been developed. But there are still tools missing for adapted or new methods of patent analysis like the White Spot Analysis of Fraunhofer IAO. Also for the analysis of patents Luxid is used, especially for the generation of a patent map is presented in the recent research, as it gives satisfactory results but the accuracy level is not much.

In comparison to TRIZ the white spot analysis fraunhofer IAO is not only based on solving technical problems by analyzing them and assigning them to a small set of basic inventive principles. The general objective of white spot analysis is to discover gaps for usually incremental technology development is patent map that is built of specific problems and solutions described in patents. Thus the core element of the White Spot Analysis is the analysis of a patent map that is built of problem and solution described in patents.

### *2.2 Manual text classification*

Much of the useful information is in the form of texts. This ranges from patents, emails, web pages, newspapers articles, market research reports, and customer complaint letters. In earlier days the classification and indexes is done manually. Classifying and indexing patents by hand were found to be an expensive, slow and labour -intensive activity. Also consistent accuracy was difficult to obtain with human indexers, and the work to cause high staff turnover. With these issues we have generated, an automated patents categorization system based on a fuzzy rule-based text.

## III. PROPOSED METHODOLOGY

In the proposed paper we develop patent classified system. The basic idea behind the proposed system is that firstly the input data is tokenized that is the patent data. In Tokenize step we use input String tokenize, specify an input string that contains all the delimiters. Delimiters are the characters that separate tokens. The words obtained from tokenize form the basis for the feature space of the training data. The main aim of the pre-processing is that is to make the data set clean. The various functions of pre-processing are- data cleaning, relevant analysis, data transformation and data reduction in order to increase accuracy, scalability and reliability of the classification algorithms. After the pre-processing techniques then feature extraction technique is applied that is PCA. After the extraction of features training set is obtained. Then classification technique is applied that is Logistics which classify the patents.

### *3.1 Feature extraction based on PCA*

Basically, PCA is a method that reduces data dimensionality performing a covariance analysis between factors. The original data will be turned into a new coordinate system based on the variance in the data. PCA applies a mathematical procedure for transforming a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

PCA is useful when there is data on a large number of variables, and (possibly) there is some redundancy in those variables. In this case, redundancy means that some of the variables are correlated with one another. And because of this redundancy, PCA can be used to reduce the observed variables into a smaller number of principal components that will account for most of the variance in the observed variables.PCA is recommended as an exploratory tool to uncover unknown trends in the data. The technique has found application in fields such as face recognition and image compression, and is a common technique for finding patterns in data of high dimension.[15]

### *3.2 Classification based on Logistics*

Logistic regression can be binomial or multinomial. Binomial or binary logistic regression refers to the instance in which the observed outcome can have only two possible types (for example, "dead" vs. "alive"). Multinomial logistic regression refers to cases where the outcome can have three or more possible types (e.g., "better" vs. "no change" vs. "worse"). In binary logistic regression, the outcome is usually coded as "0" and "1", as this leads to the most straightforward interpretation. The target group (referred to as a "case") is usually coded as "1" and the reference group (referred to as a "non case") as "0". The binomial distribution has a mean equal to np where n is sample size and p is the proportion of cases, and has a variance equal to the product of sample size cases and non cases, np(1-p).Thus, the standard deviation is the square root of np(1-p). Logistic regression is used to predict the odds of being a case based on the predictor(s). The odds are defined as the probability of a case divided by the probability of a non case. The odds ratio is the primary measure of effect size in logistic regression and is computed to compare the odds that membership in one group will lead to a case outcome with the odds that membership in some other group will lead to a case outcome. The odds ratio (OR) is defined as the odds of being a case for one group divided by the odds of being a case for another group. An odds ratio of 1 indicates that the odds of a case outcome are equally likely for both groups under comparison. The further the odds deviate from one, the stronger the relationship. The odds ratio has a floor of 0 but no ceiling (upper limit) – theoretically, the odds ratio can increase infinitely [16]


*Published By:*
*Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)*

## IV.PRACTICAL EXPERIMENT

### *4.1 EXPERIMENT ENVIRONMENT*

The main software used in our experiment is weka (Waikato Environment for Knowledge Analysis). The Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality. The original non-Java version of Weka was a TCL/TK front-end to (mostly third-party) modelling algorithms implemented in other programming languages, plus data pre-processing utilities in C, and a Makefile-based system for running machine learning experiments. This original version was primarily designed as a tool for analyzing data from agricultural domains, but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research. Advantages of Weka include:

- free availability under the GNU General Public License
- portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform
- a comprehensive collection of data pre-processing and modelling techniques
- ease of use due to its graphical user interfaces

Weka supports several standard data mining tasks, more specifically, data pre-processing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka. Another important area that is currently not covered by the algorithms included in the Weka distribution is sequence modelling.

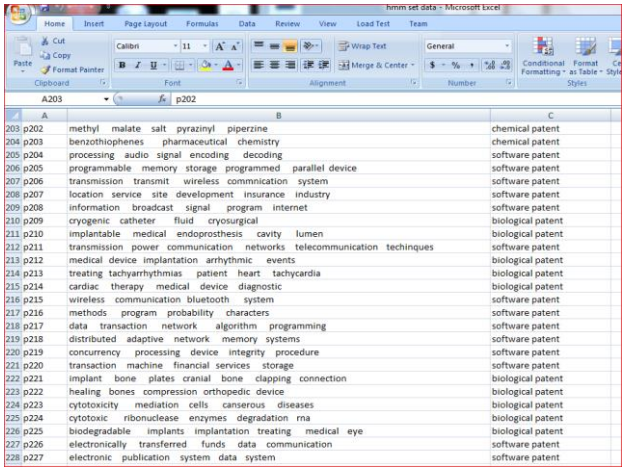
Weka's main user interface is the Explorer, but essentially the same functionality can be accessed through the component-based Knowledge Flow interface and from the command line. There is also the Experimenter, which allows the systematic comparison of the predictive performance of Weka's machine learning algorithms on a collection of datasets.

The Explorer interface features several panels providing access to the main components of the workbench:

- The Pre-process panel has facilities for importing data from a database, a CSV file, etc., and for pre-processing this data using a so-called filtering algorithm. These filters can be used to transform the data (e.g., turning numeric attributes into discrete ones) and make it possible to delete instances and attributes according to specific criteria.
- The Classify panel enables the user to apply classification and regression algorithms (indiscriminately called classifiers in Weka) to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, ROC curves, etc., or the model itself (if the model is amenable to visualization like, e.g., a decision tree).
- The Associate panel provides access to association rule learners that attempt to identify all important interrelationships between attributes in the data.
- The Cluster panel gives access to the clustering techniques in Weka, e.g., the simple k-means algorithm. There is also an implementation of the expectation maximization algorithm for learning a mixture of normal distributions.
- The Select attributes panel provides algorithms for identifying the most predictive attributes in a dataset.
- The Visualize panel shows a scatter plot matrix, where individual scatter plots can be selected and enlarged, and analyzed further using various selection operators.[17]

### *4.2 EXPERIMENT AND RESULT*

**Table 1: Dataset for the model**



| | A | B | C |
|---|---|---|---|
| 203 | p202 | methyl malate salt pyrazinyl piperzine | chemical patent |
| 204 | p203 | benzothiophenes pharmaceutical chemistry | chemical patent |
| 205 | p204 | processing audio signal encoding decoding | software patent |
| 206 | p205 | programmable memory storage programmed parallel device | software patent |
| 207 | p206 | transmission transmit wireless commnication system | software patent |
| 208 | p207 | location service site development insurance industry | software patent |
| 209 | p208 | information broadcast signal program internet | software patent |
| 210 | p209 | cryogenic catheter fluid cryosurgical | biological patent |
| 211 | p210 | implantable medical endoprosthesis cavity lumen | biological patent |
| 212 | p211 | transmission power communication networks telecommunication techinques | software patent |
| 213 | p212 | medical device implantation arrhythmic events | biological patent |
| 214 | p213 | treating tachyarrhythmias patient heart tachycardia | biological patent |
| 215 | p214 | cardiac therapy medical device diagnostic | biological patent |
| 216 | p215 | wireless communication bluetooth system | software patent |
| 217 | p216 | methods program probability characters | software patent |
| 218 | p217 | data transaction network algorithm programming | software patent |
| 219 | p218 | distributed adaptive network memory systems | software patent |
| 220 | p219 | concurrency processing device integrity procedure | software patent |
| 221 | p220 | transaction machine financial services storage | software patent |
| 222 | p221 | implant bone plates cranial bone clapping connection | biological patent |
| 223 | p222 | healing bones compression orthopedic device | biological patent |
| 224 | p223 | cytotoxicity mediation cells cancerous diseases | biological patent |
| 225 | p224 | cytotoxic ribonuclease enzymes degradation rna | biological patent |
| 226 | p225 | biodegradable implants implantation treating medical eye | biological patent |
| 227 | p226 | electronically transferred funds data communication | software patent |
| 228 | p227 | electronic publication system data system | software patent |

After uploading the dataset then the pre-processing is done on the data set. After the pre-processing feature extraction algorithm PCA is applied, then the output of the PCA classification algorithm logistics is applied.

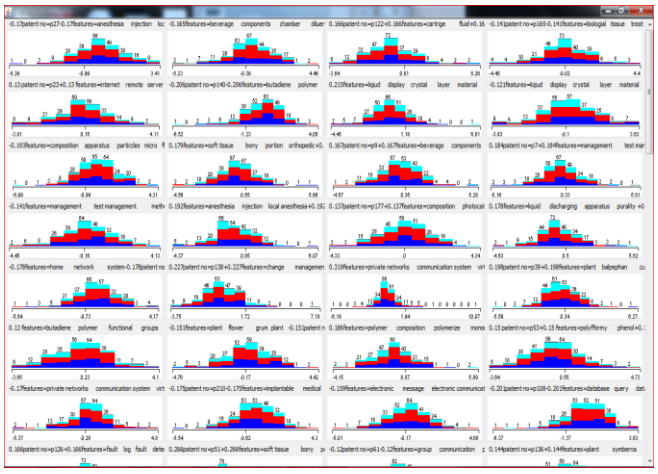


**Table 2: Visualization of features extracted by the PCA**


*Published By:*
*Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)*

Now the complete data is classified by the help of the logistics algorithm and patents are classified as software, biological, business and chemical patents

**Table 3: output of classification (a)**

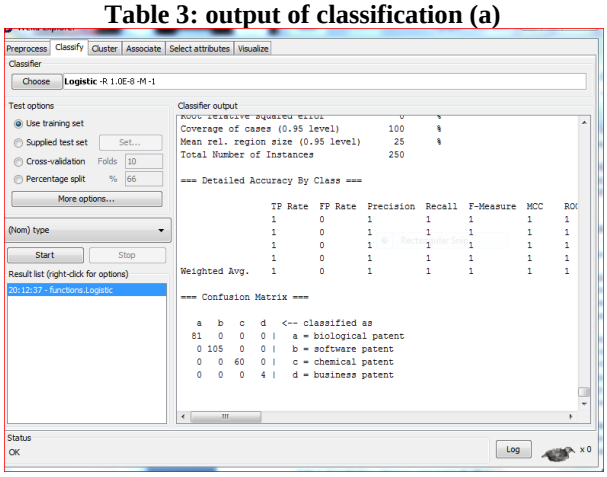
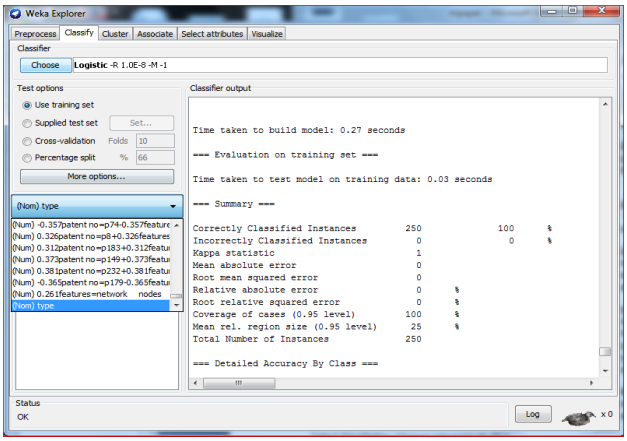



**Table 4: classifier output (b)**




## V. CONCLUSION AND FUTURE SCOPE

In this paper we have reviewed how to apply text mining techniques like PCA and logistics. PCA and logistics are very good techniques but if they used in combination then they will provide very good results because of characteristics of robustness, self organizing adaptive, parallel processing, distributed storage and high degree of fault tolerance. The result of the classifier is 100%. Noisy data is handled by pre-processing techniques. In this paper we have classifying the patents through text- mining techniques (PCA and logistics). The classification of patents is a very time consuming and tedious job. A text mining approach based on PCA and Logistics has been developed for the classification of patents based on quantitative and qualitative parameters.

The patents have been classified as software, business, chemical and biological. This research may extend by the applying the clustering algorithm and patents can be assigned to the domain experts to take decision according to their needs.

## REFERENCES

[1] Breitzman, A. F and Mogee, M.E. "The many applications of patent analysis," journals of information science, vol. 28 (3),pp. 187-205 2002.

[2] Mittelstadt, A, strategisches IP-Management- mehralsnurpatente: GestigesEigentumschutzen und alsWettbewerbsvorsprungnutzen. Wiebaden: gabler, 2009.

[3] Alshuller, G and L .Shulyak, 40 principles:TRIZ keys to technical innovation. Worcester, MA: Technical innovation centre, Inc, 1998-2002.

[4] Altshuller, G. and H. Altov, and suddenly the inventor appeared:TRIZ, the theory of inventive problem solving Worcester,MA: Technical innovation centre, Inc, 1994-2004.

[5] Siwczyk, Y. and J. Warschat, "White Spot Analysis: the potential ofpatent information for research and development," in: PICMET '11;Proceedings; Technology Management in the Energy-Smart World,eds.: D. F. Kocaoglu, T. R. Anderson, July 31 - August 4, 2011, pp.2097-2105.

[6] Mann, D.; "An introduction to TRIZ: The Theory of Inventive Problem Solving," Creativity and Innovation Management - TRIZ, vol. 10, pp. 123-125 2001.

[7] McCallum, A., Nigam, K.: A Comparison of Event Modelsfor Naive Bayes Text Classification. Sahami, M. (Ed.), Proc.of AAAI Workshop on Learning for Text Categorization (1998), Madison, WI, 41–48.

[8] Burges, C. J. C.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery Vol. 2 No. 2 (1998) 121–167.

[9] Joachims, T.: Advances in Kernel Methods-Support Vector Learning. chapter Making Large-Scale SVM Learning Practical MIT-Press (1999).

[18] Lin, S-H., Shih C-S., Chen, M. C., Ho, J-M.: Extracting

[10] Ernst, H:"Patent information for strategic technology management," World patent information, vol. 25, pp.233-242.2003.

[11] Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. European Conference on Machine Learning (ECML) (1998) Text Categorization with Class-Based and Corpus-Based Keyword Selection.

[12] Ozgur, L., Gungor, T., Gurgen, F.: Adaptive Anti-SpamFiltering for Agglutinative Languages. A Special Case for Turkish, Pattern Reconion Letters, 25 no.16 (2004) 1819– 1831.

[13] Yang, Y., Liu, X.: A Re-examination of Text Categorization Methods. In Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval, Berkeley, US (1996).

[14] D. Tao, X. Tang, X.li, and X. Wu. Asymmetric bagging and Random Sub spacing For Support Vector Machines-based Relevance feedback in image retrieval, IEEE Transactions on pattern analysis and machine intelligence, 2006