PATENT DATABASE: A METHODOLOGY OF INFORMATION RETRIEVAL FROM PDF

Pawan Sharma*, R.C.Tripathi**

*Research scholar, Indian institute of information technology, Allahabad India.
**Professor, Indian institute of information technology, Allahabad India.

ABSTRACT

Patent document holds wealth of information in itself. A brief detail of Indian patent application information is published as eighteen month publication by Indian patent Office, in electronic gazette weekly. To date, a proper database of Indian patents specifically for research determination has not been available, making it complicated for researcher to use this data for measuring any kind of research activities in terms of patents in India. To facilitate this, we constructed a comprehensive patent database which incorporates the information presented in the electronic gazette. This database includes information such as technology class, applicant, inventor, country of origin etc., of the patent submitted. We present the methodology for the creation of this database, its basic features along with its accuracy and reliability in this research paper. Patent based database has been developed and can be used for various innovation researches and activities.

KEYWORDS

Patent document, Indian patent, Electronic gazette.

1. INTRODUCTION

Despite the widely shared view that technological knowledge is one of the most important factors in economic development, many difficulties have been encountered in the measurement and counting of technological knowledge. The sources of data capable of serving as a basis for studying technological knowledge have been limited, stemming from the difficulties inherent in measuring knowledge. Patent data has been considered one of the few precious sources of standardized information for technological knowledge. However the lack of user friendly databases, coupled with the sheer volume of the data has kept researcher from exploiting this rich and valuable data mine. Since the national bureau of economic research of the United States made public its NBER patent citation data file[1], thanks to heroic efforts on the part of Bronwyn hall and her associates, research on innovation has made major strides. The rapid progress of computing has also aided this research. Many important papers have been written based on this valuable database, providing significant insight into the process of innovation.

Following this example, we developed a patent database based on patents applications filed with the Indian patent office. We call this database the PRC patent database. PRC stands for patent referral cell a research section based in IIIT-Allahabad that works in patent filing and related activities. Access to patent database is limited. There are hardly any databases for Indian patent

specifically apart from Ekaswa developed by Patent Facilitating Centre (PFC), Department of Science and Technology under Technology Information Forecasting and Assessment Council (TIFAC) in 1995. But the database is restriction in its use for research as it can be only used for single component keyword search. An instant substitute is to develop a database in a widely used package that can be updated and transmitted easily for research. Using this database the materialization of new areas of research and the level of activity in other important areas can eventually be traced over the last decade.

The rest of this paper is organized as follows. The next section 2 explains the basic features of this database, section 3 describe the field use as parameters of the database, section 4 explains the methodology underlying the construction of the database, and section 5 presents the database of granted patent database followed by section 6 & 7 which explains the graphical tool and future work with conclusions.

2. BASIC FEATURES OF DATABASE

The Indian patent office disseminates a patent electronic gazette weekly[2]. This gazette contains detail of patent application filed by the applicants in Indian patent office and published after 18 months of application filling. This electronic gazette is freely and easily available in form of Pdf year wise and also monthly and weekly on the official website of patent India. This pdf contains value of information for researcher. But since the gazette being in a pdf form with a varieties of difference in pdf it is almost impossible to be used by simple available tools which can convert this pdf into any other format. For using this data for any kind of research activity, it should be easily convert able to any other database programs like SQL, MYSQL etc. It is also desirable to have plotting and data analysis capabilities to improve individual sorting tasks. Microsoft excel word sheet, meets enough of these criteria and was identified as useful for this purpose. The conversancy and short learning curve connected with excel assist to compliment the preliminary necessities. The basic database formation of excel makes it a good choice for researcher who would like to incorporate this database with one of their own, since excel database is easily adaptable by other database programs.

3. FIELDS

The main goal in developing the PRC database was to store this information in flexible, useful form. The gazette is brief in its form and so only limited fields are available to be used. The final database contains eleven fields excluding claims and figures. A brief description of each field along with its categories is presented in table 1.

Table1. Database fields defined in PRC database main table

Field title	Description and Categories
Application number	Number granted during filling patent application.
Date of filing application	Date on which application is filed.
Publication date	Date on which first information after 18 months is published.

Title of invention	Title provided by a patent applicant.
International classification	International patent classification.
Priority Document No	Number granted by a country where the application is first filed.
Priority Date	Date on which the patent application was first filed in a country.
Name of priority country	Country name where the patent application is first filed.
Name of Applicant	Name of applicant or person who will be assignee of the patent in future.
Name of Inventor	Name of person involved in invention. This could be a single name or many depending on the number of people involved in an invention.
Abstract	A brief description of the invention.

4. METHODOLOGY

Many portable document format (pdf) converters are available on internet, which simply provide a format conversion, with the absence of structural information in the final format, except at the very low level. Also these available converters have little or no understanding of the document's structure with probabilities for errors and omissions. Information contained in the original Pdf file is in a way simply translated in another format. OCR oriented software's provides a functionality to convert pdf through scanning and OCR. But the drawback is that, information present in the pdf file is clear in its nature and the image scrutiny step can introduce noise for example unrecognized text as text zone; or unrecognized images. Also, this strategy is useful only when the pdf file only contains images from scanning.

4.1 PDF Extraction

Data from Pdf to excel cannot be directly converted. In order to convert data from PDF to Excel, the PDF must have visible tables i.e., visible table borders. This is because the conversion process takes the contents in the tables of the PDF and places them into corresponding cells of the Excel spreadsheet. If the PDF data is not in a tabular form it is pretty hard to convert it to Excel sheet. It is always possible to convert the PDF to a DOC file regardless of tables but not to an excel sheet. Excel spreadsheets could hold only 65,000 rows. Also in cases of very large PDF documents that require conversion from PDF to Excel, it is possible that the number of rows in Excel could exceed. This may cause Excel to crash or freeze. In such case, it is required that the Excel conversion take place in two or three parts. This will ensure that the 65,000 row limit is not exceeded and a new file for every year will be formed. The methodology we adopted involves conversion of pdf to an intermediate XML [4]file following the basic technological principles of information extraction, and then extracts information from the intermediate file and store it into a Microsoft excel sheet.

Also we have developed, a database for granted patent based on the url retrieval technique which is discussed in the later part of methodology section. The system consisted of three major components: the basic input file, the rule method and the extraction system. The input of the basic module is the pdf file. The intermediate output was the xml file[6], stored in buffer describing the content and structure of the pdf file. The rule module analyzed and processed the pdf file, based on the rules set to extract the information from pdf files and extraction module stores it into xml files. Every page of the e gazette, contains a patent description header which starts with the heading "(12) Patent Application Publication". We use this valuable information to identify the pages of our purpose from the e gazette. This heading has been used as header identification for using the information enclosed in that page. A weekly electronic e gazette published by patent office contains 300-400 pages and sometimes extended to 1200 pages, of which few pages in the beginning and at the end are redundant as they contain generic information in every e gazette pdf which is not required for our database. Setting a header rule, will filter out such pages. An automated system is prone to result in errors if applied, without setting header rules. With this condition applied, only pages containing patent application information are retrieved. Each pdf is analyzed, according to the rule module and information is extracted. This rule module is based on the string matching characteristics. Each field in the pdf has identified string and is in a fixed format. These strings are matched, on the basis of string matching algorithm and the data is stored in xml files.

4.2 String matching

The purpose of using string searching is to find the location of a specific text pattern within a larger body of text. The main consideration for selecting a string matching algorithm for search is speed and efficiency. There are a numerous strings searching algorithms in existence today, but the one we used for our database construction is Rabin-Karp[7]. The Rabin-Karp string searching algorithm calculates a hash value for the pattern, and for each M-character subsequence of text to be compared. If the hash values are unequal, the algorithm will calculate the hash value for next M-character sequence. If the hash values are equal, the algorithm will do a Brute Force comparison between the pattern and the M-character sequence. In this way, there is only one comparison per text subsequence, and Brute Force is only needed when hash values match.

Consider an M-character sequence as an M-digit number in base B, where B is the number of letters in the alphabet. The text subsequence t[i .. i+M-1] is mapped to the number

$$X(i)=T[i].BM-1+T[i+1].BM-2+.....+t[i+M-1]$$

Furthermore, given X (I) we can compute X (I+1) for the next subsequence T [I+1.. I+M] in constant time, as follows:

$$X (I + 1) + T [I+1] . BM-1 + T [I+2] . BM-2 + ……+ T [I+M]$$

$$X [I+1] = X (I) . B \text{ one digit shifted to left}$$

$$-T [I] . BM \text{ subtract left most digit}$$

$$+T [I+M] \text{ add new right most digit}$$

In this way, we never openly calculate a new value. We merely regulate the existing value as we move over one character. To elaborate more let's say that our alphabet holds 10 characters. Our alphabet = a, b, c, d, e, f, g, h, i, j. Let's say that "a" corresponds to 1, "b" corresponds to 2 and so on. The hash value for string "cah" would be ...3*100 + 1*10 + 8*1 = 318. If M is large, then the resulting value (~bM) will be enormous. For this reason, we hash the value by taking it mod a prime number q. The mod function (% in Java) is particularly useful in this case due to several of its inherent properties:

$$(x \bmod q) + (y \bmod q)] \bmod q = (x+y) \bmod q$$

$$(x \bmod q) \bmod q = x \bmod q$$

For these reasons:

$$h(i)=((t[i]\cdot bM\text{-}1 \bmod q) +(t[i+1]\cdot bM\text{-}2 \bmod q) + ...$$

$$+(t[i+M\text{-}1] \bmod q))\bmod q$$

$$h\ (i+1) =(\ h(i) \cdot b\ \bmod q$$

Shift left one digit

$$-t[i] \cdot b\ M \bmod q$$

Subtract leftmost digit

$$+t\ [i+M] \bmod q\ )$$

Add new rightmost digit

$$\bmod q$$

If a adequately large prime number is used for the hash function, the hashed values of two dissimilar patterns will generally be different. In such case, searching takes O (A) time, where A is the number of characters in the larger body of text. If the prime number used for hashing is small a complexity of O (MA) case can take place but this is likely to happen if the prime number used for hashing is small.

Applying this string matching algorithm, we match strings of each required field which we have described earlier in the table. Words located after the strings are tagged with XML and stored in the buffer memory. In the next step of our execution, we retrieved these tags and stored it in the specific cells of excel file. A new excel sheet is automatically utilized once the limit of 65,000 is exhausted and data are stored in the new excel file. A flow chart of the overall process is depicted in figure 1. The overall research process consists of several steps. In the first phase, patent data is collected from the pdf. This data is analyzed with string matching algorithm and stored in xml format. Finally the tagged xml data is stored in excel sheet for final output.

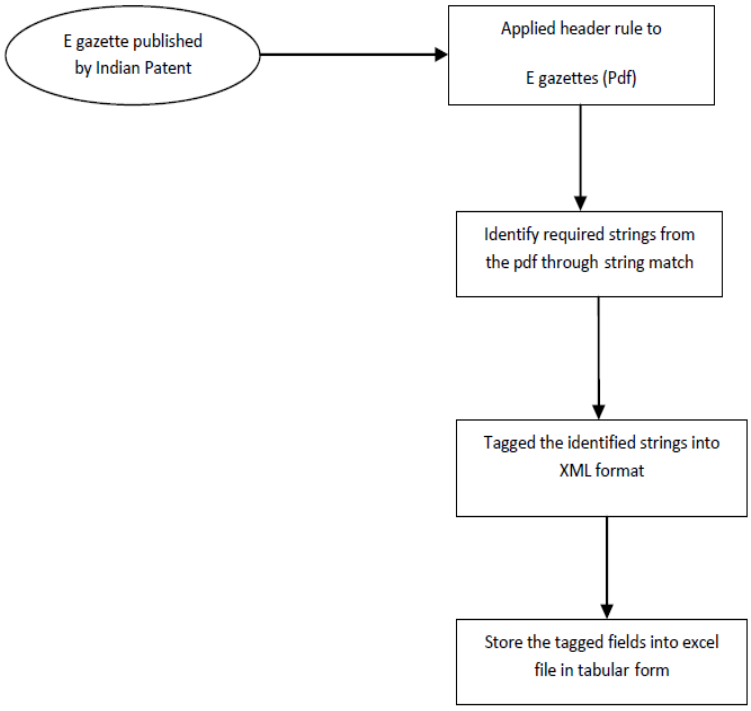

Figure 1: Work flow model of the process

5. DATABASE OF GRANTED PATENTS

The database which we developed is formed from the patent application published by the patent office India weekly. These patent applications do not result into a patent. As most of them are left unprocessed, some which proceed with FER results into case of similarity or infringement issues. On an average, of the total application only 40-45% applications are granted as a patent. For example in year 2008-2009 total numbers of applications for patents filed was 36,812 where as the number of Patent granted were 16,061. In such a situation it is important to analyze the patents which are unprocessed or which have been left abandoned, since they are merely applications and cannot be termed as a granted patent.

To filter such application from our main database we used the URL get method. Data was sent as a get method and was matched with the URL of granted patent. This URL is available on the website of Indian patent once we search for any specific patent in form of HTML page request and is in unencrypted format. Input data is the application number. For example the URL for html pages was in the following format http://ipindiaservices.gov.in /patentsearch/GrantedSearch/ completeSpecification.aspx? ID=491/MUMNP/2008[3]. In this URL we see that the last digits are application number which has been assigned an electronic id. We sent the input data ranging from 001/MUM/2008 till the last digit which was stored in our excel sheet in application number field. The program filtered out all the granted and not granted patents. For granted patents, the result was obtained in form of HTML format, complete specification along with details which we stored in buffer memory. Thereafter with the help of brute force string matching algorithm[8], we matched strings like application number, applicant name, assignee etc. Once these strings were

found the data was retrieved till it finds a new string which was set as next input. These retrieved strings were, stored in an intermediary XML file[4], which tagged every field. In the next step of process these tags were stored in various fields of excel sheet. The accuracy percentage is 97% since the html pages are in format of running text and the accuracy rate cannot be expected to be 100% compared to pdf files which are published by patent office. The benefit of applying a string matching and retrieving the data from html pages is that, we got addition information about patents, the claim part which is not made available in the published PDF. Also using a brute force algorithm for string matching was that it is more flexible in error and exception handling.

6. THE GRAPHICAL USER INTERFACE

We developed a tool to allow the operator to set up the whole conversion chain for a given gazette pdf file published weekly and validate and/or correct the processing output. The conversion takes approximately 2-3 hours for conversion of one pdf with 350 pages approximately. A snap shot image of tool is depicted below as figure 2.

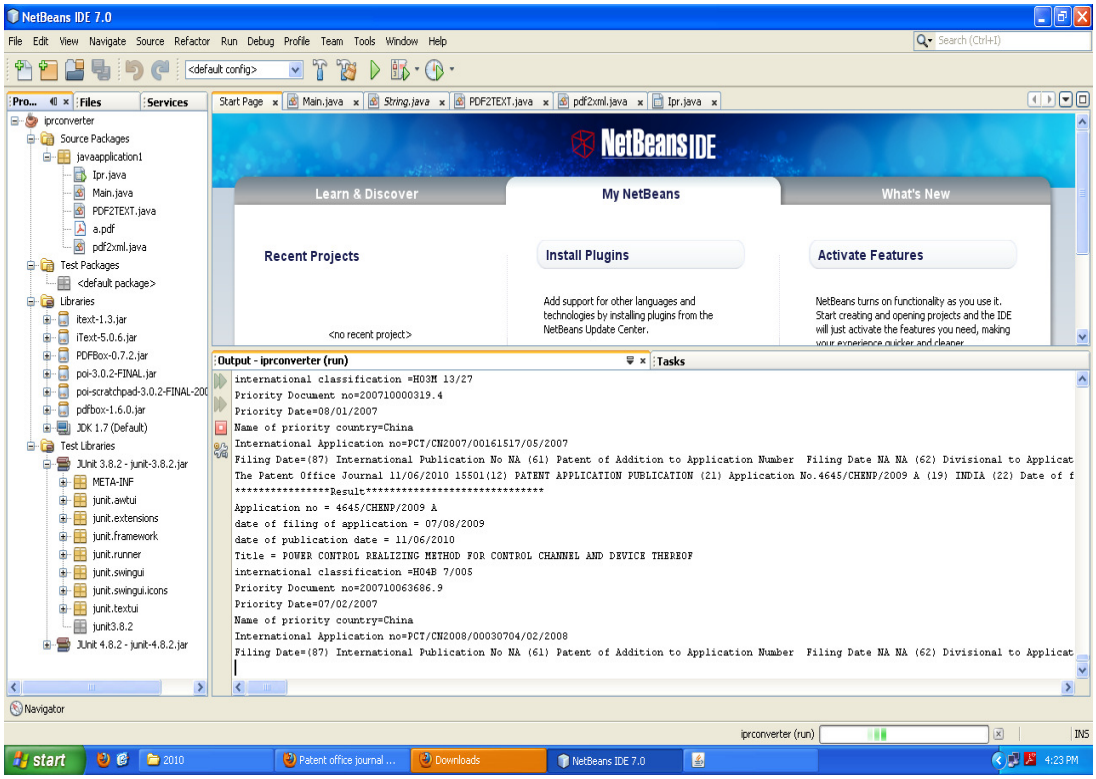

Figure 2: Snap Shot image of tool

7. CONCLUSION AND FUTURE WORK

We have presented a system to convert patent pdf into logically structured database in excel sheet. The specificity of our approach relies in the exploitation of the native internal pdf objects rather than using image based techniques on a pdf document. This database helps to solve the problem which a researcher faces in his research regarding the study of Indian patents since there are no data bases readily available which can provide such solution. Manual entry of such data is

time consuming and prone to errors. Commercial data bases like Delphion, Ekaswa etc., provides solutions for patent search rather than for research work. In future we wish to form a citation network between patents based on semantic information retrieval which will help a patent examinee to find citation for filed patents in India and will reduce time in clearing the reports. The database formed is solely used for research purpose and has no commercial applicability.

REFERENCES

[1] Hall, B., jaffe, A. and M. Trajtenberg (2000), Market Value and Patent Citations: A First Look, NBER WP 7441

[2] http://ipindia.nic.in/ipr/patent/patents.htm managed by Indian patent office, India.

[3] http://ipindiaservices.gov.in/patentsearch/search/index.aspx managed by Indian patent office, India.

[4] Herve Dejean, Jean-Luc Meunier, " A system for converting PDF documents into structured XML format document analysis system VII Pages pp 129-140 Book Subtitle 7th International Workshop, DAS 2006, Nelson, New Zealand, February 13-15, 2006. Proceedings Publisher Springer Berlin Heidelberg Print ISBN 978-3-540-32140-8.

[5] L.A. Ochoa-Franco, C.T.Haas, C.M.Dailey, A.E.Traver, Automation and Robotics in Construction Xi Proceedings of the 11th International Symposium on Automation and Robotics in Construction Copyright © 1994 Elsevier B.V. All rights reserved. *Edited by: D.A. Chamberlain* ISBN: 978-0-444-82044-0

[6] WENDE Zhang, (2008),"converting PDF files to XML files", The Electronic Library, Vol. 26 iss:1 pp. 68-74.

[7] Karp, Richard M.; Rabin, Michael O. (March 1987). *Efficient randomized pattern-matching algorithms* 31 (2). Retrieved 2008-10-14

[8] Christof Paar, Jan Pelzl, Bart Preneel (2010). *Understanding Cryptography: A Textbook for Students and Practitioners* Springer p. 7 ISBN 3-642-04100-0