Analysis on Data Mining Tools Used in Business Intelligence and Inventions

[1]Madeleine Wang Yue Dong
[1]School of Design, University of Washington, Seattle, WA.
[1]yuedongwang@hotmail.com

Article Info
Journal of Computing and Natural Science (http://anapub.co.ke/journals/jcns/jcns.html)
Doi: https://doi.org/10.53759/181X/JCNS202101010
Received 15 November 2020; Revised form 12 January 2021; Accepted 22 January 2021.
Available online 05 April 2021.

Abstract – This paper will evaluate data mining tools for competitive intelligence and technology. Data analyzers i.e. Thomson and OmniViz are the tools for completing diversified and sophisticated mathematical analyses of data. AnaVist and Aureka are considerable for modest visualization of statistics and itoplistsi is used for creating maps that are stylish. Novel features of OmniViz during the comparison of other tested tools are used for visualizing clustered data from difference viewpoints, which makes it possible to assess the attributes using patent map animation. The Thomson data analyzer provides effective tools that compare various subsets for data, such as the identification of unique attribute values. In citation assessments, Aureka is used as well as in illustrative patent maps. AnaVist is the best in retrieving basis statistics smoothly and quickly. The findings from four tools were similar, despite the fact that various databases for data retrieving were utilized. Superior investors and assignees list were the same, since they were an annual trend for geographical and technological business segments. Nonetheless, the conclusions from the findings were that business decisions are made using their tools to enhance competitive intelligence.

Keywords – Thomson, Omniviz, Anavist, Aureka, Business Intelligence, Patent Documents, Data Mining.

I. INTRODUCTION

About 75% of technical and scientific data can be location in patent documents alone as the research done by the Patent Office of Europe. Patents provide a unique source of data since they can be published, screened and collected based on globally acknowledged standards. Apart from being a major source of technological intelligence, these documents provide competitive intelligence to businesses by defining the strategies, strengths and weaknesses of businesses. Information attained from patents can also aid in identifying the partners for collaboration and licensing. Since the patent system has been created, approximately 55 million patent application can been documented. It can be challenging to identify and evaluate critical documents in a manual format [1]. The necessity to evaluate and analyze patent tools has been accepted by wide-range remedies from producers. Novel remedies are still being recommended into the market, and include the tools for assessing individual tools, reading patent data and evaluating a collection of patent documents. Remedies for evaluating a collection of patent documents can be classified into two: Tools meant to retrieve and prepare basic statistical data for patent documents; and the Tools meant to visualize and progressively evaluate patents. The former handles data in a structured manner, whereby the latter evaluates unstructured information.

This paper evaluate, four tools for evaluating patent documents have been analyzed and tested. These include AnaVist, OmniViz, Thomson and Eureka. The tools evaluate both unstructured and structured data. They also visualize the findings attained from all the clustered data fields from the patent document and project basis statistical graphs and incorporate the filters meant to perform particular remedies. These tools have been tested with respect to two fundamental cases, assessing their capacities to provide business technology and intelligence from the patent documents for businesses on a daily basis. The knowledge of enhanced techniques and forestall of products overlap the Research and Development (R&D) projects and hence prevents unessential investments. Of equally significance is the identification of essential actors that operate in the same technological field. Assessing and benchmarking the competitors in business is a market and R&D approach, which helps in the management of processes and identification of parties for cross licensing and collaboration.

This contribution focusses on the perspective of a patent expert with the competent understanding of patent information but no particular competency of data mining methods or tool testing. All these tools assessed and tested are critical for the tasks and efficient for the adoption of daily business works. All these tools have some weaknesses and strengths when compared to each other. In conclusion, it can be argued that Thomson and OmniViz tools for analyzing data are for diversified and sophisticated mathematical evaluation of data. On the other hand, AnaVist and Aureka are effective for modest visualization of statistical data and for creating patent maps that are stylish. Novel features of OmniViz, if distinguished with other tools, are possible in the visualization of clustered information from wide-range perspectives and possibilities to assess some attributes based on patent animated maps [2]. Thomson

provides effective tools for distinguishing various sets of data, such as identifying novel values of attributes. Aureka is a tool that permits citation evaluation and incorporate illustrative patent maps. AnaVist is considered the best in retrieving basis statistical data quickly and efficiently.

The results retrieved with all these tools are similar, despite the fact that various databases for data retrieval were utilized. Top investors and assignees lists were similar, since this was a yearly trend and both geographical and technological business segments reported the same results. Patent documents in reciprocal order were varied. Nonetheless, the conclusion gotten from the findings, and decision-making process by businesses would be the same irrespective of the tools utilized. This Section I has introduced the four tools: Thomson, Aureka, and AnaVist, OmniViz. The remaining section further evaluates the tools, patent documents, and their significance to business intelligence and technology. In that case, the remaining part of the paper is organized as follows: Section II provides a brief analysis of the research. Section III provides a study analysis. Section IV results for the research. Section V concludes the paper.

II. BRIEF OVERVIEW

Business Intelligence and Technology for the Patent Documents

The patent scheme is based on the protocol that to attain the aspect of monopoly, innovation need to be comprehensive with upmost accuracy for individuals that are skilled in a specific art to enhance implementation. Typically, patent application incorporate 18 months following the filing scheme and are presentable for anyone typically in an electronic manner. The significance of patent information as a technological source and business intelligence has been accepted for many decades now. Researchers have differentiated four major types of technical data carriers: processes/products, people, technical publications and patents. In the research, it is argued that, in the context of patent data, irrespective of comprehensive inadequacies, it can be considered as a novel source of technical data. It is gathered, analyzed and published based on standardized protocols. It progressively generates an evaluation of sophisticated measure of metrical technological transformation. It also allows transparent accumulation of competencies on an international scale.

Based on the comments evaluated above, patent data is fundamental for R&D and business intelligence, oven for industries without the projections to utilize patent for various tasks. Evaluations of patent documents show the current sophistication, including helping in identifying white spaces and technological fields that lack innovation. Because of the need to disclose an invention, the creation of patented methods and products can be evaluated in their entirety and this might be challenging. Annual trends in R&D may be identified by evaluating annual data in patent documents. Emphases on R&D also vary based on geographical regions that might be evaluated through patent information.

Annual trends in patenting represent the current hot segment, which are the areas with multiple inventions and technological segments that are declining. Evaluations for technological intelligence have to be completed by professionals in the technological segment to attain maximum profits and benefits. The patent documents also provide business intelligence meant to support the firm's business decisions [3]. The knowledge of other firm's patent portfolios provide critical data on the rivals and aids in the allocation of potential competitors for licensing technologies and collaborations, which includes the identification of novel entrants in the present market. Identifying another firm's or rival's patent activities identifies its strategies and business trends.

Annual tends indicate critical technological segments that have been abandoned by the firm and the segment it is focusing on at the present. Application activities for various patent firms, in every country's patent office, reveal a firm's geographical business approach. It is however considered that some data is always confidential to firms. To receive patent, innovation should be novel and essential requirements have to be considered as well. This shows that patent application should be the first place when inventions are disclosed. This provides a significant possibility of identifying novel techniques and product before they are seen in the market. The evaluation of patent investors and assignees identifies collaboration between different actors and provide essential data that facilitates recruiting.

Patent Data

Data from patent documents includes both unstructured and structure formats. In most patent documents, the first page normally includes bibliographic information that is structured and the abstract and title that are unstructured. The document also includes a definition of claims, which is an illustration of legal protection for patents, inventions and essential drawings. Publication and numbering practices vary significantly between various patent firms [4]. The major attribute utilize for the evaluation; nonetheless, identified in all patent documents, are standardized in some form by various database distributors.

The various fields in patent documents are illustrated using global numbering system. The implication of these fields utilize in the analysis have been evaluated in upcoming sections of this paper. Patent documents integrate the identification of numbers. Priority information integrates the priority numbers that have been assigned for the first application applicable for patents, which also integrates the corresponding dates. The number of publication is provided to a document during the period of publication i.e. eighteen months following its filing. The date of publication denotes to the corresponding dates. The date of issue represents the dates granted to the patent and this is typically three to five years after filing.

However, this depends on the patent departmental office. The patent applicants or assignees are the individuals and organizations that hold the rights for application and invention for patents. Inventions structured in collaborations are allocated to the various parties involved. The investors include the research developing the inventions. Inventions have been categorized based on technologies attached to them. One of the most commonly

utilized is the International Patent Classification, 51 (IPC International Patent Classification) – 51) [5]. Most of the patent firms have formed custom categorization systems such as the USA patent classification. A number of database producers have also formed individual categorization system to create analyses and searches of essential document easily.

In this research, two of these categorizations have been utilized and provided by Thomson; these are Derwentis manual code and Derwent categorization. The abstract (57) and title (54) signify the illustrations of inventions of the natural languages. These represent the unstructured data and its informative format of it varies significantly based on the author.

III. STUDY ANALYSIS

The general number of patent documents that have been published is 60,000,000 and this incorporates the applications that never produced patent. It might be challenging to identify and evaluate all the essential patents without automatic application [6]. The necessity for evaluation and analysis of the patent tools has been accepted by most of remedy providers. Novel remedies are being considered into the market; the tools used for reading and assessing each patent e.g. PatBase, STN-Viewer, ScioSphere, and the tools for evaluating multiple patents. Remedies of the former type can be grouped into two segments: tools meant to retrieve and create basic statistics for patents e.g. Patbase, Qpat and Lexis Nexis, and the tools meant to visualize and progressively evaluate patents. The latter type handles unstructured data type while the former also evaluates unstructured data [7]. The former one has been focused in this contribution.

This research was done based on the perspective of patent analysts. The analysts, as focused in this paper, provide business data and technology for businesses and R&D departments. This research focused on clarifying the type of assistance the established remedies provide for normal works of analysts. Four remedies significant utilized by patent evaluations have been tested. These tools included AnaVist. Thomson, Aureka and OmniViz. This paper presents an observation made during the process of testing and presents the findings of two testing instances. In this section, the study framework and set of data utilized have been defined.

Data Mining

The application of terminologies in literature relating to the evaluation and processing of patent-based information can be confusing. Data mining, visualization, text mining, and patent mining all relate and are the terms used when explain documents processing. In this section, we will focus on providing the definition of terms and provide an explanation as to why data mining was selected as the theme topic of this research.

J. Bacardit and X. Llorà in [8] have defined the terminology data mining as the evaluation of massive observable data meant to identify unspeculated connections and make a summary of information in a comprehensive and fundamental manner to the analysts. The researchers have defined text mining as the skill-intensive procedure for user interaction with documents and the collection process based on the application of analysis tools. In a way analogous to the process of data mining, text mining focusses on the extraction of critical data from the various sources of information through the exploration and identification of the best patterns. For text mining, the sources of data are information collection, where patterns are identified in the standardized databases but in an unstructured form.

R. Pérez-Castillo, D. Caivano and M. Piattini in [9] defined the clustering process that categorizes different objects into different classes. This is executed through the categorization of different objects. The variation between categorization and clustering is that in categorization, issues are presented based on the collection of pre-categorized training samples and system tasks is obliged to master the definition of different categories for it to be capable of classifying novel objects that have not been labelled. As for clustering, the issue is to classify particular unlabeled collections into comprehensive clusters without any data submitted in prior. Any individual labels that are linked to particular objects are retrieved from the information.

Researchers in [10] defined the terminology patent data as the processing of information to process patent documents. The ideology is to visualize information and how it is explored. Visualization is based on the presentation of information is a visual manner, which allows individuals to get significant insight into information, create conclusions, and interact with information. The enhancement of visualization methods over the past few decades has been fulfilled to enhance visualization of data that is of low dimension, and this incorporates making histograms a significant attribute of data and to formulate novel visualization of high-degree textual data. The patent documents include both unstructured and structured data, which are semi-structured. Bibliographic data of patents is structured and is based on a significantly strict format. For instance, it incorporate the assignees of the patents and the names of the inventors, various identifiers e.g. number of publication and priority, categorization and years.

H. XIE and X. CHEN in [11] focused on unstructured data as text illustrating the inventions and the dimension of protecting patents such as claims, abstracts and titles. According to the researchers, result visualization of patent evaluation is identified as patent graphing whenever it has been structured from structured information, and patent map for unstructured data.

Nonetheless, H. Han and C. Zhang in [12] admit that, the terminology map has been utilized in typically all the cases. The four tools evaluated and tested analyzed both the unstructured and structured data, and the results of visualization of all the classifications. Visualization is possible for both the high-degree and low-degree data e.g. the pie and bar graphing for patent landscape and structured sets of data i.e. mappings for the unstructured sets of data. An evaluation of textual information is done based on the first categorization and it is based on frequent texts.

In this paper, the terminology data mining is utilize as an upper level terminology for all the activities that handle massive amounts of data. Text mining on the other hand denotes to the evaluation of unstructured data.

Testing Case Analysis

In this paper, tools have been tested based on two critical cases, assessing their capacities to provide business intelligence and technology from different patent documents for firms on a daily basis. Being versed with the latest technological segments is fundamental for firms and this is a critical process of innovation. The skillset of developed technologies and product forestall that overlap R&D efforts prevents unessential investments. Similarly, other essential actors operate in the same field. Evaluating and benchmarking competitors is a market and R&D approach that helps in the management of processes and identifying potential parties that are used in cross-licensing and collaboration.

This analysis based on the perspective of patent analysts with the better comprehension of patent information but do not have particular skills on methods of data mining or testing of tools. Moreover, to assess the findings on the study and the value for business intelligence and technology, their application rate was compared based on particular questions:

- Are the tools easy to use?
- Do the tools require the users to read manuals before focusing on the evaluation?
- What forms of information formats are supported by the tools?
- Are there any reasons for manipulating information and can they be evaluated by the tools?
- Can research be enhanced using the tools; focusing on the findings by eliminating the texts, which have been included due to extensive search profiles?

1st Case Analysis: Technological-centered information: The first case analysis assessed the tools with the technology-centered information. The patent texts are the most accurate and comprehensive source of technological-based data. In reference to D. Gagliardi in [13], approximately 75% of technological data can be identified from these patent texts. in the advent of temporary monopolies of inventions, investors are required to disclose this in enough and satisfactory details. The 1st case evaluated the patent texts that handle technological application for evaluating friction between road surfaces and motors. Any potential restrictions by the time of the patent tests were not incorporated. The profile of search was structured based on the restricted searches of the essential texts into a single IPC category i.e. to measure and test. Therefore, the patent texts includes works such as road, vehicle, friction and conditions, including the various synonyms related to them.

In the study, particular answers to the following queries were provided:

- What are the patent trends in the technological areas at the present?
- How potential is it to evaluate particular groups of technology?
- How can invention be done with a particular firm where X is related to the Y firm's invention?
- Can it be easier to locate particular documents from multiple patent texts?

2nd Case Analysis: Firm-Based Information: The 2nd case assessed the portfolio of the patent of a particular firm. Identifying another competitor or firm patenting an activity can reveal its business strategies and business strengths. Annual trends indicate the technological segments the firm has eliminated and identifies the segments being focused on at the present. Applying the various activities to various patent firms, ie, various nations and offices reveals a firm's geographic organizational strategies. Patent evaluation provide a significant possibility that reveal novel techniques and products before their arrival in the market. The evaluation of patent inventors and assignees reveal the available collaborations between different actors and provide data for head-hunting. In the 2nd case analysis, Fraunhofer Gesselschaft was selected. The firm is based in Europe as the largest company applicable for this research. The company filed its applications and patents from 1995.

This 2nd case focusses on analyzing how various queries can be answered using different tools. These queries include:

- What are the specific technologies a firm can focus on?
- How has its level of concentration transformed the company for the last decade?
- What is the regional segment that the company operates in?
- What forms of cooperation has been established by the company?

Datasets and Databases in Testing

The different sets of data were retrieved from four various databases commercially used: PCT Full, USPat Full, DWPI, and Micro Patent, and three various sets of data were created. The dataset content is formed with the same profile. An integration of PCT Full, USPat Full and DWPI information was applied in the tests of AnaVist. The DWPI information was utilized to test OmniViz and Aureka. The application of these tools in actual life is limited based on the application of various datasets and databases. Aureka utilizes Micro Patent and AnaVist makes use of four various databases. Thomson evaluates information from various sources; however, it is based on particular filters meant to handle DWPI information.

OmniViz makes utse of information in a different format and provides wizards that import information in Excel format in Microsoft, which explains why Micro Patent information was utilized to assess it. The evaluation was done out based on the application of bibliographic, abstract and title data of the texts. Micro Patent represent the Patent Web, which is an internet repository that is created by Thomson. It incorporates more than 50,000,000 texts

and records of the front page. The Patent Web incorporate the documents retrieved from 6 different patent firms and PCT: complete documents from patent and trademarks offices in the U.S.; and the patent office in Europe.

DWPI represents the databases created by Thomson. It is known for its value addition that implies to the patent texts that wrote abstracts and documents in English for the purposes of being standardized and informative. The database has patent texts from more than 38 authorities globally. Information from different members of the patents has been integrated into a single document. These documents are integrated in the DWPI categorization. The codes and manuals are meant to enhance the search for the documents. US Patfull represents the databases from the patent and trademark offices in United States. It incorporates the complete documents from the offices from 1971. US Patfull integrates the complete tests of the patent cooperation treaty, which publishes applications the globes intellectual properties since the late 1970s.

Due to the fact that information incorporates two various cases and three different databases, this amounted to a total of 6 sets of data. All these sets incorporate patent applications and patent tests. Micro Patent information was saved on a excel file in Microsoft.

Searches have been restricted to the texts categorized into IPC category: 'to measure and test'. Texts had to incorporate a single word: Car or vehicle, and automotive, include the following combinations:

- Friction or skid (5N)
- Highways or roads
- Roadways or freeways (condition)
- Roads, freeways or highways (condition)

These searches created more than 649 tests from the Micro Patent and 1080 from the DWPI databases. In STN, tests from DWPI were combined with the texts that have been taken from US Patfull and PCT full. These texts were evaluated through the limitations of fields for claims, abstracts and titles to create a more accurate search. This amounted to 1340 tests. For the firm-based data, the searches have been restricted to the applications and patent documents filed after the early 1990s. These searches produced more than 4500 texts from the DWPI databases and 4625 texts from Micro Patent. The variation in figures is because of the few days that elapsed.

Analysis with Tools

The wide-range tools and remedies for evaluating patents and the novel ones are gradually being presented into the market. These are tools for evaluating and reading individual tools and patents for evaluating different patent texts. Remedies can be grouped into two segments: tools for preparing and retrieving statistical data for patent texts; and the tools used to visualize and evaluate patents. The former category handles information in an unstructured manner, whereby the latter evaluates both the unstructured data and other sets of data.

The tools used in the analysis were tested and presented by past researchers. In this research, tools have been tested with two actual-life cases to study how they effectively fit to their actual cases available in a firm in its normal business operations. The main purpose of this research was to assess the manner in which tools are to utilize and how applicable and informative these analyses are. The research considered the two instances, one with technological-based information while the other is a firm's portfolio for patents.

Section IV below focusses on the results and remarks of the evaluation done by the technological-based data, while Section V focuses on the firm-based data. The section will present data on the tools and the relative observations concerning their applicability. The tools included the processes meant for clustering unstructured data, in this instance, abstracts and titles for patent tests and to visualize different clusters. They also created tools meant to retrieve basis data on various attributes identified in different sets of data. Tools that have been test might be classified into two classes. The first class incorporates Anavist and Aureka that are efficient and easier to utilize and provide basic analyses with minimal efforts and evaluations requires by an individual.

The available data is limited to minimal patent databases: Aureka utilizes data that has been retrieved from the Micro Patent databases, while AnaVist from the four various databases for patents. Alongside the highly applicable interfaces, the efficiency of application is based on the restrictions of individuals from affecting the strategies and algorithms utilize in dealing with data. The tools in the other class, TDA-Vantage-Point and OmniViz represent those tools that empower statistical evaluations done on different forms of data. The application of the tools necessitates some degree of study.

Because of the possibilities of assessing data from unrestricted sources, a number of preparations before focusing on analyses is required. Tools generate wizards and filters to aid in the importation of information. Various analyses provided by the tools are efficient to render through the application of more default values. Due to the influence of the users, there are various potentials for selections e.g. clustered algorithms. A preparation of enough search profiles, mostly in creating technological-based patent texts, is normally challenging. Searchers has to identify the potential balance between eliminating some essential tests and incorporating essential ones. The applicability of the evaluation, nonetheless, is based on the data validity. All these tools tested allow the redefinition of information by creating subsets of them and handling they as new analyses. Data visualization is made easier through the location of documents, which were no related to subjects and eliminate them.

Aureka

Aureka is a Thomson tools used to evaluate and visualize patent information and sharing it within the firm. The most effective element of Micro Patent includes the representation visualization, interface and basic statistics. The tool is incorporated for the Micro Patent database. The tool is more flexible and permits modification and broadening of sets of data even in the midst of evaluations, irrespective of the fact that it limits some evaluation to

just the complete patent document from some global patent offices. Operating with the tool require minimal preparations of information of tool learning. The tool focusses on the sharing of data within the firm and signification of findings of analyses. Statistics might be visualized with custom reports or data transfer to the excel file. The tool do not provide more interactivity at that level as other tools have been tested; however, patent texts may be identified from visualizations and comprehensive application.

OmniViz

This tool is Bio-wisdom and utilized as a significantly powerful tools for data mining. It has been creased significantly for evaluating biological information and data. However, it is created for handling patent texts from other fields of technology also. The most effective element of the tool is its efficiency, interactivity, flexibility and supply of various visualization methods. The tool is a significant for users familiar with the algorithms and methods of data mining. Whereas preparing information, users can select, for instance, the algorithms and stop texts utilized for clustering e.g. Hierarchical and K-means. The tools has been created to be easy to utilize using default figures and has been suited for creating typical statistics and visualization for patent texts.

OmniViz focusses on visual representation and evaluations of patent information. There are eight various forms of visualization for handling various data perspectives, searching for clusters created from them, terms associated, and the connections between special visualization and numerical elements for network connection in biological texts. Different data formats can be formed using the tool. It provides the filters for data importation and for has no limits for capable individuals. Information utilized in evaluation might be interlinked with various sets of data, such as publications and patents. Due to the data format being non-predetermined, it require a number of prior preparations before the process of clustering.

Cleansing of information may be created before the process of clustering through the edit of auxiliary file or utilizing the tool. The tool concentrates on the visual evaluation and more complex for creating basis statistics. The tool provides high-dimensionality of interactivity in the wide-range tools that are utilized in data analysis. The process of selection in a single tool significantly influences different aspects in the workplaces. It is potential to visualize animations of categorical and numerical information such as annual patent filling and any related records in a visualization.

AnaVist

AnaVist is a tool created by the Chemical Society of America to enhance and facilitate flexible analyses during the process of retrieving data from the STN information bank. The most effective element of the tool is its representative visualization, seamless connection within various analyses, preparation of different statistics effectively and an interface that is friendly to users. The tool evaluates information that has been accessed from 4 STN databases. Two of these databases are complete data patent databases i.e. PCT Full and US PatFull, and two are based on value addition databases i.e. CA Plus and DWPI that integrate the scientific and patent references for biochemistry and chemistry. In 2008, the amount of databases were boosted by one, during which the EP Full database incorporating the complete data and patent texts in Europe were filed in the late 1970s.

Data can be imported with ease from STN and transferred to enhance advance processing. On the contrary, it is restricted to the patents in various fields compared to biochemistry and chemistry. Formulating visualization of data using the tool is prompt that the preparation of statistics from various perspectives. AnaVist has an incredible dimension of interactivity between various forms of analyses.

Thomson

The tool is created by the Thomson Reuters that utilizes search technologies of the VantagePoint software for data mining and analysis. The most effective element of the tool is its efficiency and flexibility, including its macros for formulating various tools and reports that compare various segments created by data. Thomson evaluates information in almost every format. It further generates filters that aid in the importation process, and critical tools utilized in processing and handling information from DWPI. The tool focusses on data processing and integrates the most basic forms of visualization despite including sophisticated calculations for backup. It provides various forms of tools for data analysis, matrices, and lists for modest statistics and the maps meant to assess the connection between clustered data and terminologies.

The Thomson analyzer tool also permits individuals to effectively reveal different data records, which are the same or the assignees with similar texts i.e. patent application that have been documented based on collaborations. These approaches are significant for the identification of statistical divergences from information i.e. identifying the technological segment, which are mostly patented, critical segments and firms, which are active in most technological areas. The tool provides three various forms of reports that are predefined. These reports incorporate modest and processed data concerning the subject matter. They are filed in an Excel file format and incorporate graphs and tables. Reports might not be updated or upgraded. The three forms of reports include firm reports, for accessing data on a specific firm, Firm comparison reports used to compare 2 to 5 firms and Technological reports that include different metrics concerning the areas of technology.

IV. RESULTS

The tools of analysis were evaluated and tested based two critical cases. This section focusses on the findings of the evaluations with technological-based information. Some of the data have been eliminated due to their similarity in findings to the ones in other Sections. Various specialized inventions patented and individual frequencies in information may be assessed through the cluster of terms that appear in the patent texts. In this paper, abstracts and titles have been utilized for the process of cluster. Tools have been used to visualize the findings by mapping clusters and documents in equal proportion to each other, i.e. formulating maps for patents. The texts with the same subjects were in the same place in those maps. As a result, it is easier to identify the most sophisticate segment of invention. It also indicates the data outliers, texts, which do not relate that much with the subjects.

The tools evaluated and tested allow closer analysis of custom clusters and documents. Some modest statistical evaluations were done on information to gain effective comprehension. Answers to the queries: When did? What did? and Who did? provide an overall idea about the subject. These queries were responded to by investigating the patent inventors and assignees, priority year and the various categorization provided for the document.

Aureka representation was identified to be the most illustrative and clearest. The document frequencies have colors and lines of contours. Each document is signified by dots. Three of the most essential terms in the cluster differentiate the documents from other tests. The top most frequency of the text was of the methods that relate to the motor breaks, which has been illustrated using words systems, braking force, and brake in the process of visualization. Fig 1 represents the statistics from Aureka data. Aureka provides two fundamental tools used in the retrieval of data; predefined filters and reports for exportation of data to excel. The reports that are predefined are provided for all the wide-range elements in data. Reports are difficult to edit and this is normally require for attracting data into displayable manner.

Fig 1: Aureka - Evaluation of patents that relate to the evaluation of road surfaces and friction

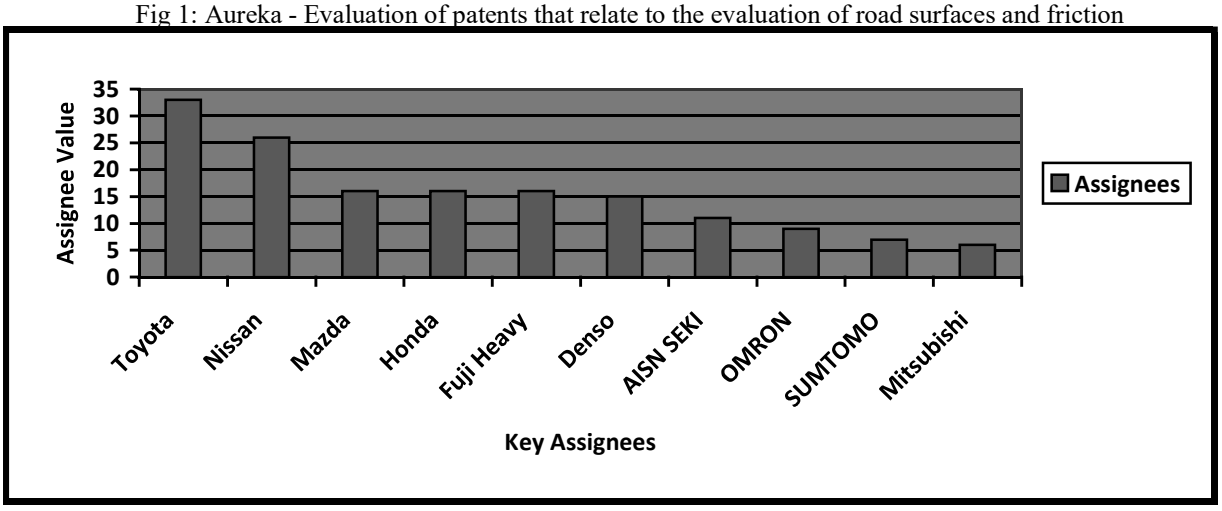

Fig 2: Number of years against Annual trends in filling up to 2006

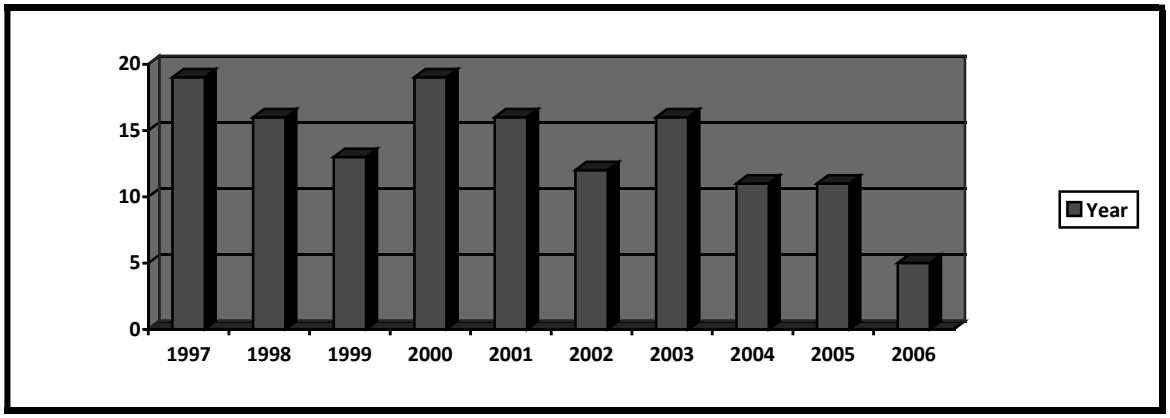

Table 1: IPC for the patents

IP CATEGORY	Text Count	%
B60-T000-8172	16	7.2
B60-R000-1602	12	6.4
G01-N00-1902	11	4.9
B60-T000-858	11	4.5

G01-B00-1130	8	4.5
B60-G000-170165	7	3.4
B60-G000-2300	5	3.0
B62-D000-0600	5	2.3
G01-M00-1700	5	2.2
B62-D000-0601	18	2.2
Occurrence of the 10 best patent categories	98	41
The remaining IPC	155	60
Overall primary IPC	265	
Application No.	189	
Patent No.	75	

Because of this, the tool provides tools used to import information to excel and the macros used to creating effective statistical presentation. The top most assignees for the patents and the document amounts filed relate to the technological application in question and this is represented in the Fig 1 above. Mazda, Nissan and Toyota are three of the most active firms that patent in this field and handle with the invention that relate to roads. Annual trends in filing are indicated in Fig 2. Patent activities varied significantly until 2006, and this seemed to have diminished significantly. Table 1 indicate the global patent categorization provided by patent firms. The data indicates a division of various technology for various documents compared to what clustering does. IPC in the reports that have been predefined are non-informative since they are based on the statistics by excel and Aureka's macros.

Fig 3 represents the visualization of datasets from various points-of-view created by OmniViz. The same way, in the texts visualized from the graphics on the left side, galaxy visualizations have been engaged. The tool incorporates the documents in a cluster form that are close to one another. Every document incorporates the squares and dots that represent cluster centroid. All these clusters represent three frequent terminologies; however, they have been centered on a few images that are clear. Graphics included in the right side have been created using ThemeMap visualizations. This indicate similar patent landscapes; however, they are visualized from a different angle. It is thus more effective to evaluate the document amounts in every cluster.

Graphics can also be zoomed or rotated in OmniViz to visualize them from the various directions. The clusters with the terminologies friction, surface and wheel has high frequencies close to one another at the top most segment of the representation. In this case, three clusters are included at the bottom, are separated, and have minimal texts. There are possibility that they do not relate to the subject. A close evaluation of the clusters was done. This evaluation revealed that they assess the methods of navigation, not assessing the relative conditions of road surfaces, but outliers. The evaluation was done using data by eliminating them out to retrieve the most accurate set of data for the purpose of the assessment.

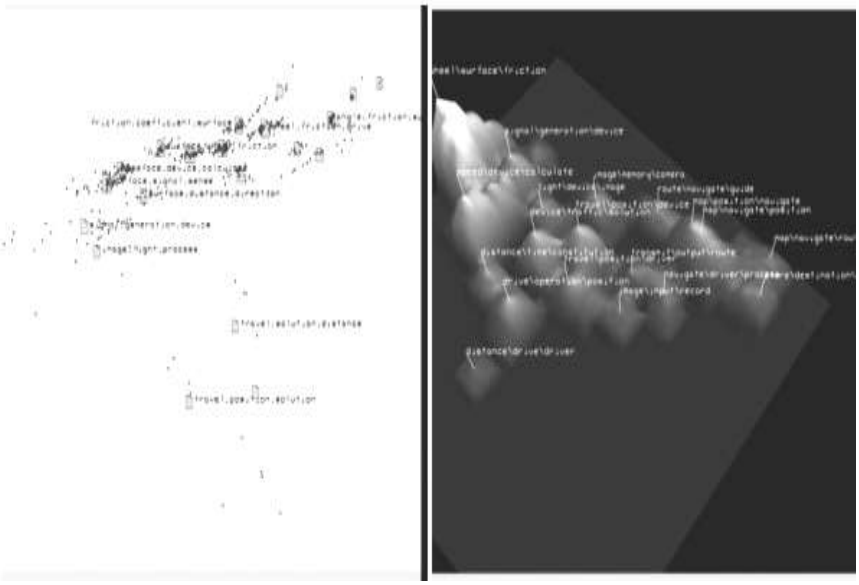

Fig 3: OmniViz visualizations of patents that related to evaluation of road surface friction by dual visualizations (ThemeMap – right and Galaxy – right).

Fig 4 indicates the modest data from OmniViz. The tool provides filters meant to export application for more evaluation. The OmniViz represent the group tools that show the document count of active assignee and the Galaxy visualization that indicate the assignees being coloured with the colors that correspond.

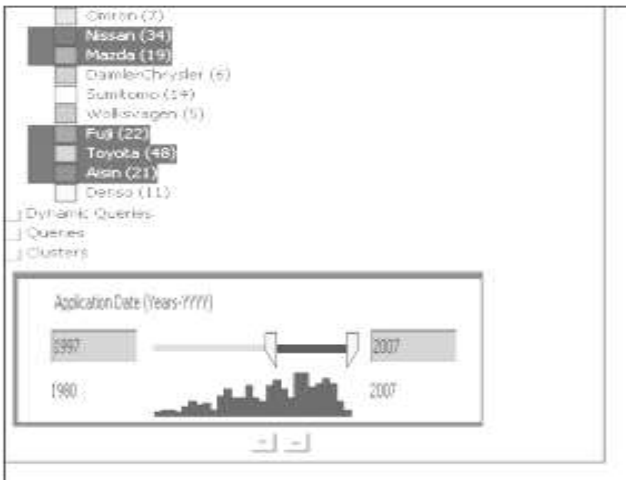

Fig 4: OmniViz data based on application date

Fig 5 represents the top assignees against various firms. The Galaxy landscape with patent texts of active assignees have been colored.

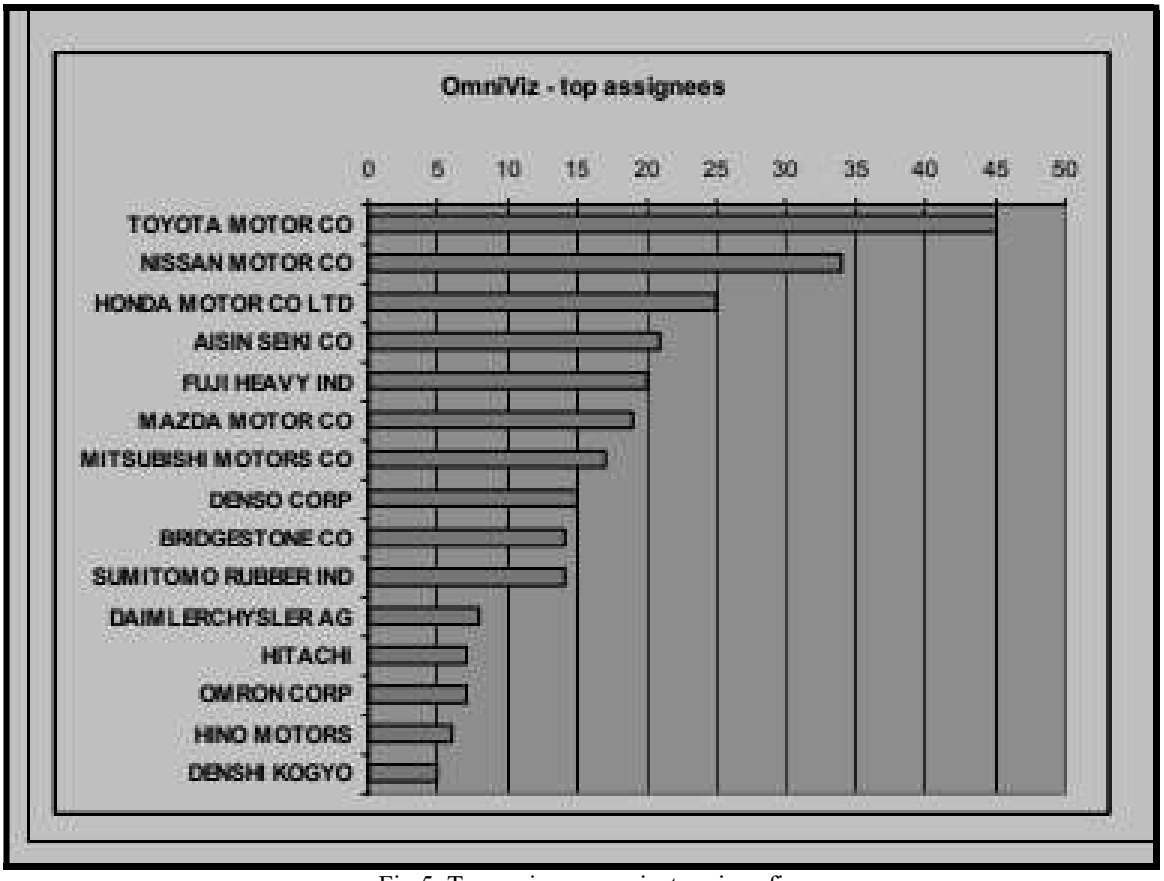

Fig 5: Top assignees against various firms

In Fig 4, the middle part of OmniViz represents the query tool that indicates an annual trend of participation from patents.

Fig 6 represents the evaluation created from AnaVist. The patent landscape visualizaiton can be visualized on the left side of Fig 6. The dots in the presented landscape signify the individual patent texts and documents. The

areas with dark color represent the clusters and the document frequency. Two words included besides the clusters show two most used words in the clusters. 10 most used terminologies have been visualzied by a shifting cursor on the landcape.

Fig 6 can also be rotaed to enhance visualization. AnaVist allows more effective formulation of basis statistical data. An annual trend of patents can be visualized from the graphs 7, 8 and 9 and are represented based on their years of priority. Fig 7 shows the mostly assigneed, with Honda, Nissan and Toyota on top. Fig 8 represents the geographical segment used for protection. Majority of the patents are filed in Japan and the US. Fig 9 shows the priority application years for patent documents.

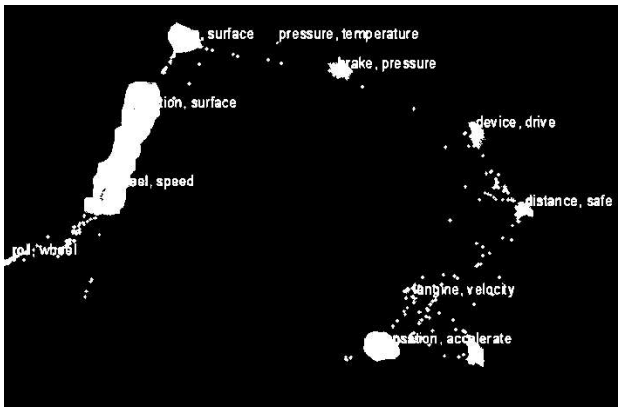

Fig 6: Patent landscape visualizaiton by AnaVist

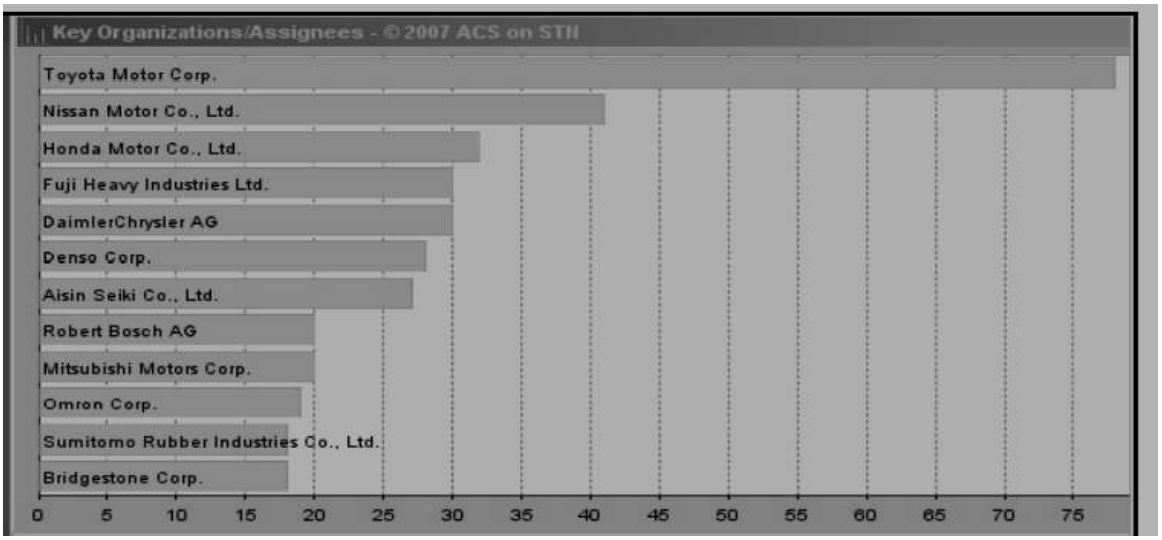

Fig 7: Key organizational assignees

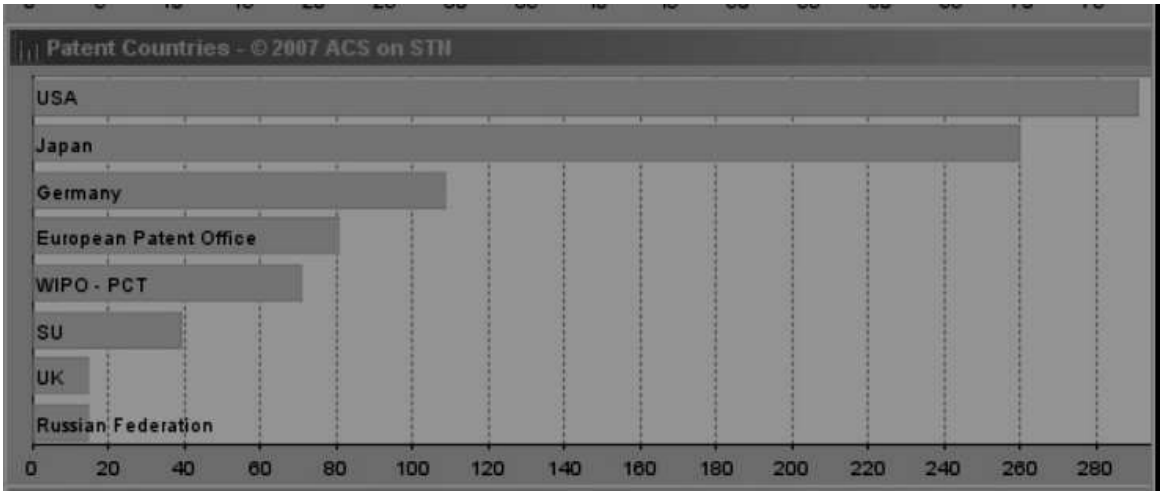

Fig 8: The geographical segment used for protection

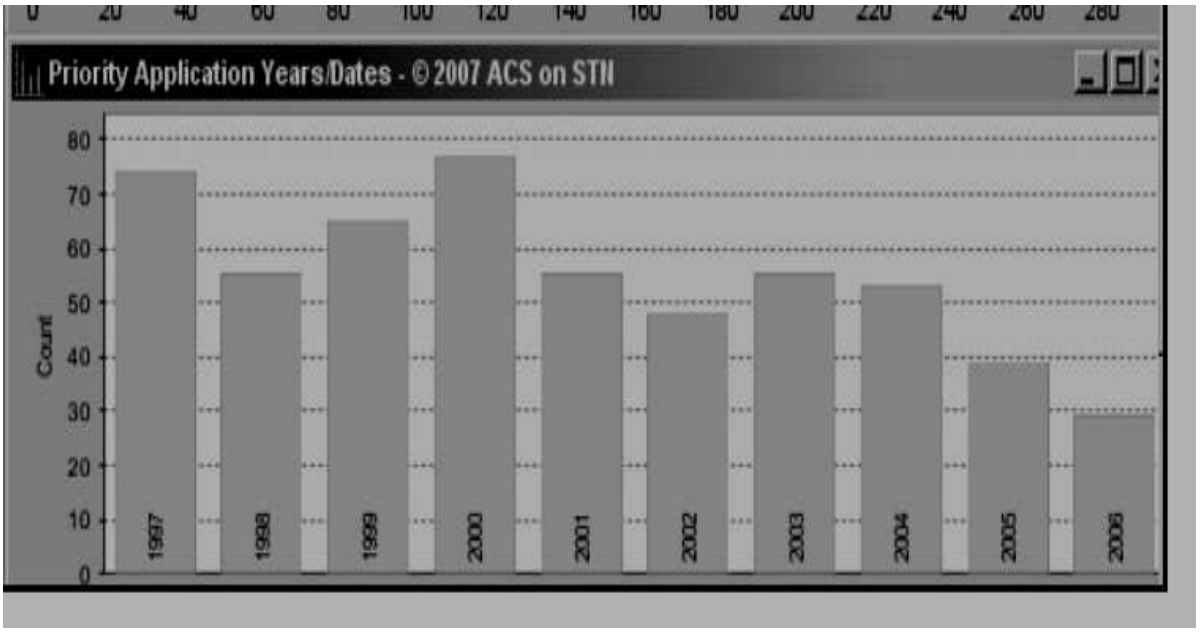

Fig 9: Priority application years

Fig 6 represents the AnaVist visualizations of patents that relate to the measurement of road surface friction. The applications of patent have been categorized to illustrate the technological concentration. IPC represents the global categorization given by authorities of patents. A number of databases also maintain individual categorization to aid in the process of discovering essential patent texts. The most utilized category includes the technologies that relate to the measurement of friction and braking control. Derwent manual codes indicate significant comprehensive inventions, in the lead and these are inventions that relate to the measurement of friction on road surfaces.

It can be summed up that the results obtained in this Section made by the various tools were the same. They provided also similar top assignees to patents. Their order however, varied. Nissan and Toyota are considered as the major actors in the sector. The deviation was that Thomson analyzer missed out in the analyses. The annual trend graphs indicated an enhanced patent since the early 1990s and 2000s and minor transformation to date.

V. CONCLUSION

In this paper, the data analyzer tools Thomson, OmniViz, AnaVist and Aureka have been applied and seen to be applicable for tasks in the business world. All these tools had their weaknesses and strengths compared on one another. AnaVist and Aureka were seen to be easier to utilize and their evaluation can be considered in preliminary data analysis activities. Aureka's strength is seen in its visual representation of findings and clear interface for users. The including of citation evaluation provide an added advantage to the evaluation. AnaVist is also credited for its easier user interface. The most efficient AnaVist feature include the potentials to visualize results of a particular analysis with ease from various viewpoints at a higher dimension and time of tool interactivity.

Despite the fact that OmniViz, Thomson and VintagePoint require significant time of preparation and mastering before an evaluation can be executed, they normally tend to compensate for all the significant effort for providing

more sophisticated evaluation and highly dimensional power of making decisions for the users. They are considered an effective tool for power users. However, they are useful and efficient for modest evaluations. The strength of the tools include the possibilities to utilize information from multiple sources and based on multiple formats and they also provide filters to help in the process of importing data and enhancing the process of data collaboration in various formats.

VantagePoint has a significantly special tool that is skilled in dealing with value-added patent information from DWPI that is a significantly unique database for patent data. The OmniViz data analyzer tool has focused on enhancing various remedies for data visualization. The high-dimensionality of interaction provided, such as, animation of annual trends is the best feature considered in this case. Both of these tools have various positive evaluating tools that have not been introduced in this research e.g. Thomson is considered as an auto correction matrices and map used to assess any items that are related within a specific field while the OmniViz represents the analysis of biological and numerical data. Limited presence of data sources includes AnaVist and Aureka weaknesses. Nonetheless, Aureka has limited this by permitting the data evaluated using Thomson and transferred for the purposes of visualization. Resultantly, this boosts the value of Aureka, even though the application of Micro Patent data might minimize the overall scope of evaluation.

AnaVist provides 5 different databases for evaluating and formalizing two of them, which are value-added databases i.e. DWPI and Chemical Abstract Services. The Chemical abstract plus service incorporates the bibliographic information from scientific and patent publications in the field of biochemistry. The preparation require to provide chemical and chemistry engineering services is also included in this case. The major weaknesses of OmniViz include the modest statistical data and the requirement to comprehend the state-of-the-art statistical techniques, such as, correlational matrices. The Thomson tool for data analysis had weaknesses such as visualization stiffness. The major statistics are more effective to access; however, the formation of presentations was challenging; for instance, record number restrictions. The model visualization formats minimized its clarity. In general all the tools reported their possibilities to create segments for closer assessment of information. In addition, they presented tools used to label cluster identifies and visualization attributes.

References

[1]. S. Ljaskovska, "DATA PROCESSING OF TECHNOLOGICAL PROCESSES IN MECHANICAL ENGINEERING", Scientific bulletin of the Tavria Agrotechnological State University, vol. 8, no. 2, 2018. Available: 10.31388/2220-8674-2018-2-43.

[2]. C. Liu and J. Yang, "Decoding Patent Information Using Patent Maps", Data Science Journal, vol. 7, pp. 14-22, 2008. Available: 10.2481/dsj.7.14.

[3]. S. Mishra, "Determining patent filing targets based on patent cost retrieval from Patent Examination Data System", World Patent Information, vol. 65, p. 102024, 2021. Available: 10.1016/j.wpi.2021.102024.

[4]. P. Pollick, "Processing of patent bibliographic data at chemical abstracts service", World Patent Information, vol. 3, no. 3, pp. 128-131, 1981. Available: 10.1016/0172-2190(81)90147-2.

[5]. M. Herz, "On-line data bases for chemical patent searches", World Patent Information, vol. 2, no. 3, pp. 119-124, 1980. Available: 10.1016/0172-2190(80)90055-1.

[6]. J. McDowall, "Prioritizing patent sequence search results using annotation-rich data", World Patent Information, vol. 33, no. 3, pp. 235-239, 2011. Available: 10.1016/j.wpi.2011.04.011.

[7]. M. Karvonen and K. Klemola, "Identifying bioethanol technology generations from the patent data", World Patent Information, vol. 57, pp. 25-34, 2019. Available: 10.1016/j.wpi.2019.03.004.

[8]. J. Bacardit and X. Llorà, "Large-scale data mining using genetics-based machine learning", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 3, no. 1, pp. 37-61, 2013. Available: 10.1002/widm.1078.

[9]. R. Pérez-Castillo, D. Caivano and M. Piattini, "Ontology-based similarity applied to business process clustering", Journal of Software: Evolution and Process, vol. 26, no. 12, pp. 1128-1149, 2014. Available: 10.1002/smr.1652.

[10]. "Methods and algorithm for analysis of patent statistics data series", World Patent Information, vol. 10, no. 4, p. 266, 1988. Available: 10.1016/0172-2190(88)90287-6.

[11]. H. XIE and X. CHEN, "Cloud storage-oriented unstructured data storage", Journal of Computer Applications, vol. 32, no. 6, pp. 1924-1928, 2013. Available: 10.3724/sp.j.1087.2012.01924.

[12]. H. Han and C. Zhang, "Color Map and Polynomial Coefficient Map Mapping", Journal of Software, vol. 5, no. 10, 2010. Available: 10.4304/jsw.5.10.1068-1076.

[13]. D. Gagliardi, "Material data matter — Standard data format for engineering materials", Technological Forecasting and Social Change, vol. 101, pp. 357-365, 2015. Available: 10.1016/j.techfore.2015.09.015.