International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 12, Number 9 (2019), pp. 1549-1555

The Design and Implementation of a Big Data Technology based Patent Analysis System

Junghoon Shin[1], and Joongjin Kook[2]*

[1]Ph.D., Technical Director, Wert Intelligence, Seoul, Korea.
[2]Assistnat Professor, School of Information Security Engineering, Sangmyung University, Korea.

**Corresponding Author*
ORCIDs: [1](0000-0002-3395-7365), [2](0000-0002-0033-388X)

Abstract

Today, the interest in intellectual property rights is growing, and the importance of patents, which is one of the intellectual property rights, is emphasized. In the case of a patent, it contains a lot of technical information like a paper, and it is possible to obtain useful information such as prior research and technology trends through patent analysis. However, patents have a large amount of data and it is difficult to collect patent data for the entire world that is continuously generated. In addition, there is a problem that patent data is difficult to process and analyze because the format of patent data provided for each country is different, and semi-structured and unstructured data are included. Therefore, in this paper, a big data-based patent analysis system is proposed to analyze patent data efficiently. The proposed system makes different formats of patent data into a common format by utilizing a big data technology and provides an environment to easily analyze the patent data.

Keywords: Big data, Patent processing, Patent Data Analysis

I. INTRODUCTION

As the era of the knowledge and information begins, the importance of intellectual property rights has been growing. Intellectual property rights are one of the means by which an exclusive status can be secured in the market and the company's or agency's technological development costs can be recovered [1][2]. In addition, Patent Trolls, companies that generate profits from royalties on patents or intellectual property rights without manufacturing or selling products, have appeared [3]. As the value of the intellectual property rights and patents has increased, there is the steady increase in patent application and registration worldwide.

IP5 is South Korea, the United States, Japan, Europe and China, which stands for countries with the largest number of patents applicated. Countries including South Korea, the United States, Japan are continuously applying for patents to secure their own rights. For South Korea and the United States, the number is steadily increasing year by year. For Japan, there has been a slight decrease, but a large number of patents are still applied for. China had a small number of patent applications in the past, but it has been increasing rapidly in recent years. Recently, IP5 agreed to disclose patent information through an agreement, making access to the patent information easier than before. Patents contain not only professional information such as thesis, but also a lot of information actually used in industries. Thus, analyzing patent data can provide the useful information to predict the prior art and the future technology trends [4]. It is also possible to derive new information through the analysis techniques such as data mining.

In this paper, we propose an analysis system to efficiently analyze patent data from IP5 countries and the WIPO (World Intellectual Property Organization) by using big data technology. However, there are some problems to be solved to do that. The first problem is to collect the existing large amounts of patent data and the continuously generated new patent data. Patent management institutions in each country provide patents and related data in the form of Open API of Bulk or SOAP/REST methods. Therefore, the proposed system periodically collects patent data through open API, which is continuously added. The second problem is the storage and processing of collected large amounts of patent data. In this paper, we use big data technologies such as *Hadoop* [5], database and search engine to store and process large amounts of patent data. The reason to use Hadoop is that it can distributed store large amounts of data and distributed process them by using MapReduce. The search engine supports searches of patent data such as bibliographics, abstracts, claims, and descriptions of patents, and the management of them is also easy when an index by country is generated. The third problem is that patent data stored by country by period have different schema. In this paper, we develop a common patent data model based on items important to the patent data analysis, and transform the data into one common format using this model. The last problem is how to analyze large amounts of patent data. In this study, patent data stored in Hadoop can be analyzed using MapReduce and all data analysis tools that can be linked with. The proposed system provides an environment that can standardize different patent data formats by country into one unified format and facilitate the patent data analysis. This paper expounds further the previous paper [6].

The composition of this paper is as follows. In Chapter 2, big data technologies and patents as related researches are looked at. It is followed by Chapter 3, where we describe the proposed system. In Chapter 4, experiments and evaluations using the proposed system are carried out. The conclusion is discussed in Chapter 5.

II. RELATED WORK

II.I Big Data Technologies

The existing technologies that analyze and process data have mostly stored and analyzed data into the memory of a single computer or a single server. The existing statistical tools used to process data also still stored data in the memory to execute statistical and analytical algorithms. Although the size of data that can be processed has increased with the development of database systems, data analysis programs are still optimized for a single computer or a single core. In fact, it is impossible with existing systems and software architectures to analyze data in from terabyte to petabyte and to deal with data such as search engines and social network services. Therefore, in the era of big data, big data analysis and infrastructure technologies must be used [7]. Big data analysis technologies are those already used in statistics and computer science, especially in machine learning/ data mining, and these analysis algorithms are improved for large amounts of data processing and applied to big data processing. Due to the growth of unstructured data, text mining, data mining, opinion mining, social network analysis and cluster analysis are drawing attention among the analysis methods [8]. Hadoop, a leading big data infrastructure technology, is a Java-based open-source framework that can handle large amounts of structured or unstructured data. It is the result of the implementation of Google File System and MapReduce Programming Model proposed by Google [9][10]. Hadoop can manage and analyze large volume of data, and flexibly deal with the increase of the amounts of data in the future. It consists of HDFS (Hadoop Distributed File System) and MapReduce frames [11].

II.II Patent Analysis

To analyze patent data, an accurate understanding of patent data is required. Patent data includes bibliographics such as application number, application date, classification information, citation information, name of invention, and personal information like the applicant, as well as image data such as abstract, claim, description, procedure diagrams and drawings. In addition, the administrative information including all procedures from the patent application to the patent registration and the follow-up management processing, can be regarded as a part of the patent data. As such, patent data contain many items and many forms of data, which can be analyzed using various analysis methods. In addition, it is necessary to utilize big data-based analysis technologies for large amounts of patent data analysis [12]. Because of these characteristics of patent data, the research on the patents is currently being conducted by using various methods in different fields.

Patent data includes specialized information, enabling to predict future technology trends and plan for technology development in specific areas [13][14][15]. For example, by analyzing patents filed by one firm, the direction of products to be released can be predicted, helping other companies develop strategies for technology development. The patent analysis has been basically conducted using data analysis methods such as text mining, data mining, and opinion mining [15][16][17].

Through the analysis using characteristics of patent data such as IPC (International Patent Classification) code information and cited relationship information, research on patent categorization is also being carried out[18][19][20]. In addition, the research on the technology management using patent management and the research on the patent analysis technology using ontology and semantic technology are also being carried out [21][22].

Currently, domestic and foreign organizations that provide information on patent search and patent analysis online are as follows. In Korea, there are KIPRIS [23] provided by KIPO (Korean Intellectual Property Office), and Wert Intelligence [24], WIPS ON [25], WINTELIPS [26], and WISDOMAIN [27] provided by private companies. Domestic patent information service providers provide basic patent search services and simple statistical and analysis services. Foreign patent information service companies include Total Patent [28] and Thomson Innovation [29]. Total Patent is a U.S. company that provides services to search for patents and utility models in 100 countries, and help check the litigation information by establishing a database on U.S. litigation information. Thomson Innovation provides services to search for Canadian companies' patent data and uses its own patent-related indicators and tools called DWPI (Derwent World Patents Index).

As the significance and the value of patents have increased, there are many analysis and search systems. Still, analysis system studies are needed to analyze patent data around the world from a more diverse point of view.

III. PROPOSED SYSTEM

In this paper, we propose a big data-based patent analysis system that can efficiently process and analyze large amounts of patent data. Fig. 1 shows the overall configuration of the system proposed in this paper. The proposed system consists of an application service module, a data processing module, a data aggregation module and a patent data storage (HDFS, DB), and each module is used to aggregate, store, process and analyze patent data.

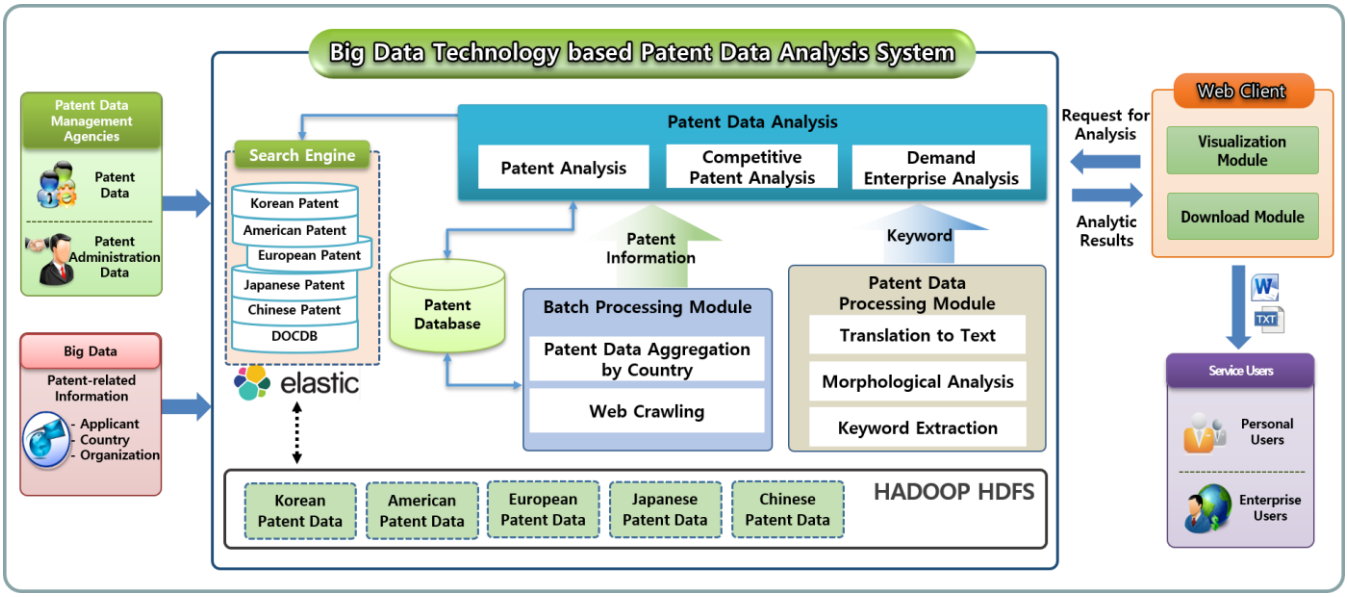

Fig. 1. Big data technology based patent data analysis system architecture

In the proposed system, patent data is stored using HDFS (Hadoop Distributed File System), databases, and search engines. HDFS stores the original and converted patent data. The reason to use HDFS in this paper is to store and process large amounts of patent data including original texts and drawings. Basically, the original patent data aggregated through the data aggregation module will be stored in HDFS and in the database, and the stored information is converted using the data processing module and then stored in the database or search engine.

Elasticsearch based on Lucene was used as the search engine. The search engines are effective in patent analysis as well as in search for patent data. Search engines support search of patent full text such as bibliographic, abstract, claim, and description of patents, and make indexes separately by country. In Elasticsearch, the index serves as a table for RDBMS. The patent database stores patent bibliographic extracted through data processing and other metadata needed for the patent analysis. It also stores the information from the patent data analysis.

III.I Data Aggregation Module

The first thing to analyze patent data is to solve the problem of aggregating existing patent data and new patent data generated continuously. In this paper, we aggregate patent data periodically depending on the data providing methods of patent data management institutions such as WIPO, KIPO, USPTO, JPO, EPO and SIPO through the data aggregation module. Because patents continue to be filed and registered, patent management institutions update patent data periodically to provide them in many ways. Although each institution has different methods of providing patent data, it can be summarized in three ways: to provide it using the Web, and to provide it as Open API, and to provide it in the form of Bulk data. For example, KIPO provides patent data in the form of Bulk data and as Open API, and USPTO provides patent data in the form of Bulk data through the Web. Fig. 2 shows the structure of the data aggregation module of the proposed system. As shown in Fig. 2, the data aggregation module consists of batch processing module and data aggregation tasks.

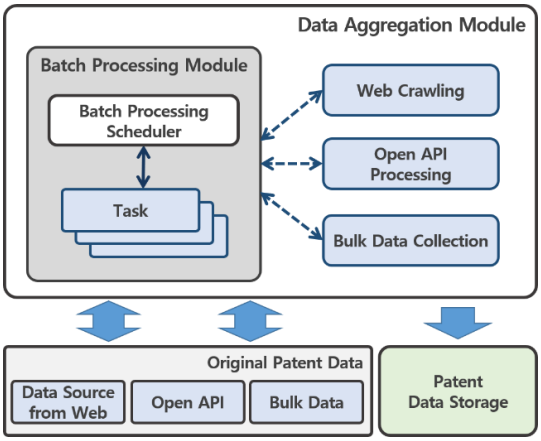

Fig. 2. Data aggregation module

The data aggregation module periodically aggregate data using a batch processing method in order to collect patent data from patent data providers. The batch processing module manages batch processing tasks that are executed periodically by using the batch scheduler. Major batch processing tasks include Web crawling, Open API processing, and Bulk data processing, and other tasks that need to be processed periodically, can be registered with the scheduler to perform the tasks. Patent data aggregated through the data aggregation module are stored in the patent data storage of the proposed system.

III.II Patent Data Module

Currently, patents worldwide are based on WIPO's patent standard format. However, different management organizations manage patent data in different formats. In addition, the formats of the past patent data and the current data are often different. Therefore, in this paper, a common patent data model is created based on data important for analysis

among patent data around the world, and by using this model, the patent data in different formats are converted into one format. In this paper, WIPO's patent standard format and country-specific data schema are analyzed in order to create a common patent data model. In this study, a considerable amount of time has been invested in this study for the analysis of the patent standard format and country-specific data schema. Patent data include a number of types of data, including the bibliographic. In addition, different types of schema are used by country, and for the past patent data, patent data were analyzed directly because the information about the schema was not available. All the patents were analyzed to extract common items, and a common patent data model for the patent analysis was developed from the extracted common items.

III.III Data Processing Module

The data processing module is a module for refining and process the original patent data stored in HDFS. Fig. 3 shows the structure of the data processing module of the proposed system. In this paper, the data processing module was designed and implemented using MapReduce to process large amounts of patent data.

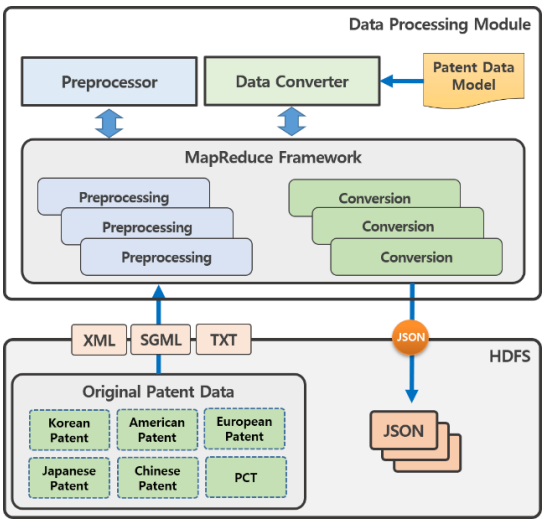

Fig. 3. Data processing module

The number of patents applied and registered worldwide amounts to about 200 million. For accurate analysis of patent data around the world, data refining and data conversion are essential. It takes an average of about 0.01 seconds to preprocess and to convert a patent data on a single server. So, it takes about 23.1 days only to process 200 million patents worldwide. In fact, it may take several months to a year or more to process patent data worldwide on a single server. This is because, as mentioned above, patent data are kept in SGML (Standard Generalized Markup Language), XML (Extended Markup Language) and TXT format by using different schema by country, and patents managed by an institution also use different versions of schema. In addition, there are many errors and exceptions to the patent data because it is directly input by a person. Therefore, the data processing is required at several times in order to address all of these data errors and exceptions. And patent data may also need to be reprocessed for the patent analysis. Therefore, it is essential to use the distributed processing frameworks such as MapReduce in order to reduce patent data processing time.

In this paper, a patent data parser was developed to process data formats and data versions by country to analyze the patent data around the world. By using this parser, the data processing module extracts metadata from patent data and converts the original data into a common patent data model format.

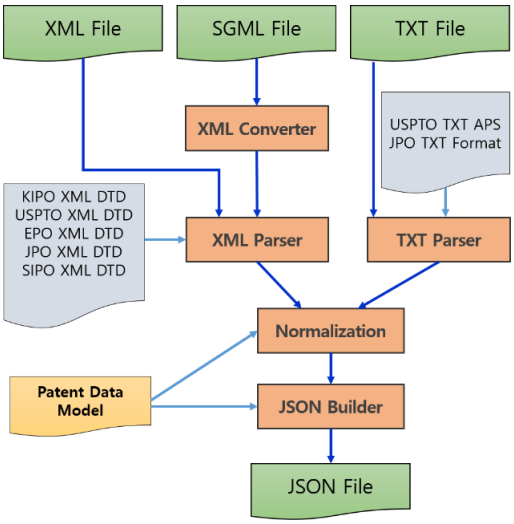

Fig. 4. Patent data conversion processing

The process of the patent data conversion is shown in Fig. 4. In the patent data converter, three types of patent data files are input: XML, SGML, and TXT. The input data are processed differently depending on the format.

The processing for each file is described as follows. For the XML file, an XML parser is directly input and processed. The XML parser then parses and extracts each item included in the XML using the DTD of input XML. The extracted items are subjected to the normalization process depending on the patent data model. After completing the normalization process, JSON (Javascript Object Notation) file is created through the JSON builder.

In this paper, a separate parser for SGML file is not developed. The SGML file is a form for patent management organizations to store patent data prior to the beginning of 2000, and is not currently used. SGML is a broad concept that includes XML, and the structure of SGML and XML is almost the same. However, XML is more constrained than SGML, so the conversion of SGML into XML is required to process SGML files through an XML parser. Thus, instead of developing a SGML parser, the proposed system developed an XML converter that converts SGML into XML. The XML converter is responsible for converting SGML in accordance with XML constraints. The main tasks of the XML converter are as follows: The XML converter matches the start-tag with the end-tag. For SGML, only a start tag is required, but for XML, an end tag is needed if a start-tag is. Next, the format of the attributes of elements in SGML need to be adapted. Fig. 5 shows an example of converting SGML into XML using an XML parser.

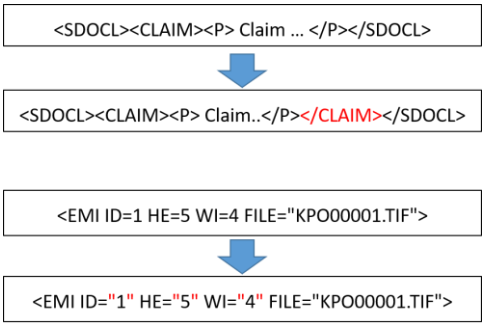

Fig. 5. An example of converting SGML to XML

This converted SGML is converted into a JSON file through the same process as an XML file. Finally, the TXT file is processed by directly typing a TXT parser. However, the TXT file is the past patent data of USPTO and JPO and is not well structured like XML and SGML, so we developed the TXT parser using the manuals or the definitions provided by organizations. When parsed through the TXT parser, it is converted into a JSON file through the data normalization and a JSON builder like the XML file conversion process. As described above, the proposed system generates the converted patent data in JSON format. The reason for the conversion into JSON format in this paper is, for JSON, the size of the data is smaller than XML and it is easier to use in various analysis libraries or analysis applications due to accessibility to various programming languages. In addition, Elasticsearch, which is used as a search engine in the proposed system, supports JSON type as the input and output, making it easier to work. Fig. 6 shows a sample of the JSON file.

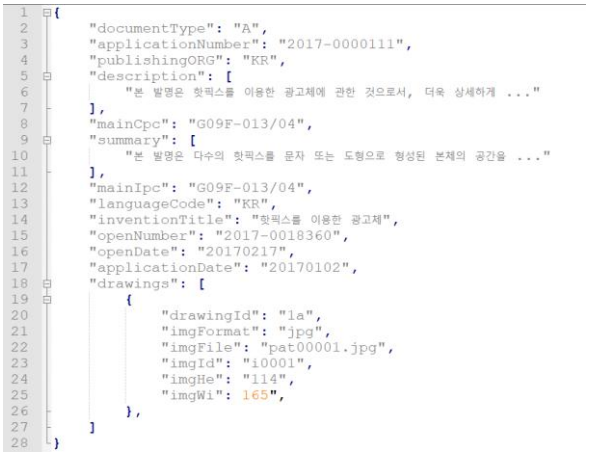

Fig. 6. Example of patent data formatted to JSON

III.IV Application Service Module

The application service module provided in this paper serves to provide users with services such as data processing and analysis using the information stored in the data storage. Fig. 7 shows the structure of the application service module. The patent data analyzer analyzes data using the patent data stored in the data storage. The patent data analyzer can be used along with the MapReduce framework, and with all analysis applications including R, SPARK, Mahout, TensorFlow, etc. which can work with HDFS. The search engine data processor handles all the tasks requested for the search engine, and HDFS allows to insert or update patent data converted into the JSON format through the data processing module by considerable. In addition, these inserted data can be searched through the index data of the search engine. The morpheme analyzer for indexing and searching can also be used and tested. The morpheme analyzer plays a very important role in processing of natural language and in generating the indexes of search.

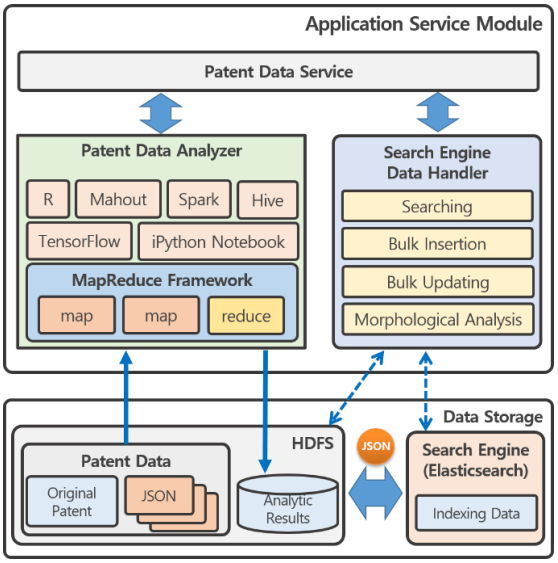

Fig. 7. Application service module

IV. EXPERIMENTS AND RESULTS

The experiment and evaluation of the big data-based analysis system proposed in this paper show the range of patent data provided by the proposed system and the analysis results obtained by using several proposed systems. The total number of patents that have been converted into a proposed standard model to date by using the proposed system is about 150 million. Table 1 shows the number of patent data converted by institution. DOCDB in Table 1 is the data of the bibliographics and English abstracts of patent data all over the world aggregated and provided by WIPO, whose number of the data exceeds 110 million.

Table 1. Summary of the performance of ML algorithms

Patent Management Institution	Number of Patents
KIPO	5,064,174
JPO	10,496,330
USPTO	10,868,952
EPO	2,716,474
SIPO	5,391,294
PCT	3,355,901
DOCDB	113,228,328
Total	151,121,453

A simple analysis was conducted using the proposed system. The experiment was conducted on the domestic patents whose name includes 'display'. Based on the data processed in the proposed system, the number of patents containing 'display' in the name of the invention is 65,571. First, we extracted only the top five nouns by counting the number of words from the names of inventions including 'display'. Table 2 shows the frequency of appearance of nouns included in the name of the inventions and in abstracts among patents containing 'display' in the name of the invention. Words without meaning, such as 'inclusion', 'above', and 'invention', were not dealt with.

Table 2. Appearance Frequency of a Specific Keyword in the Invention Name or Abstract

Keyword(noun)	Appearance Frequency	
	Invention Name	Abstract
display	65,571	59,490
device	33,664	33,197
method	27,395	25,254
panel	17,934	23,110
produce	7,863	15,843

Next, we graphed the results of analysis on the recent 10,000 patents based on the date of application among patents containing 'display' in the name of the invention.

Fig. 8 shows the number of the application of patents containing 'display' in the name of the invention by year. As shown in Fig. 8, there are many patent applications related to display in 2015. In fact, the reason why the number of recent applications is small is that patents are basically disclosed one year after the application.

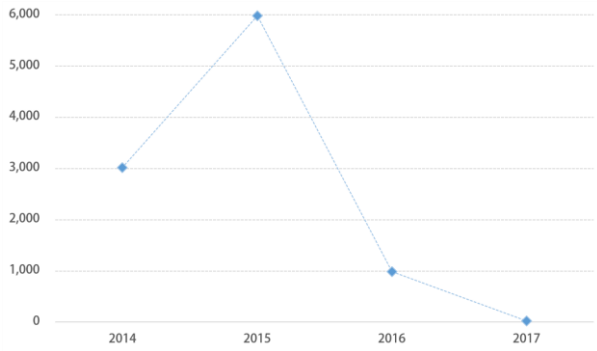

Fig. 8. The number of the application of patents containing 'display' in the name of the invention by year

Fig. 9 shows the number of the application of patents including 'display' in the name of the invention depending on the applicant. As shown in Fig. 9, Samsung Electronics Co. has the largest number of patent applications related to displays.

Fig. 10 shows the number of the application of patents including 'display' in the name of the invention depending on IPC. IPC (International Pattern Classification) means an international classification code that classifies patents by sector.

As shown in Fig. 10, G06F sector has the largest number of the application of patents including 'display' in the name of invention. G06F refers to the sector of digital data processing by electricity.

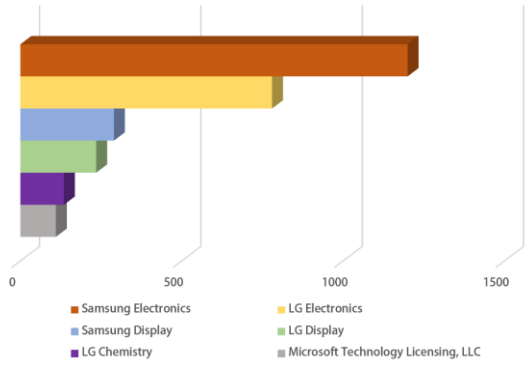

Fig. 9. The number of the application of patents including 'display' in the name of the invention depending on the applicant

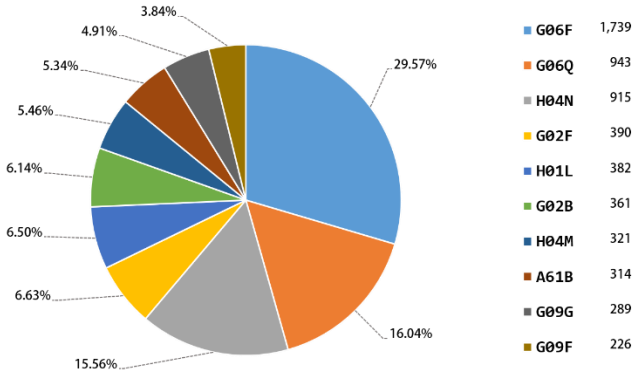

Fig. 10. The number of the application of patents including 'display' in the name of the invention depending on IPC

Experiments show that the proposed system makes it easier to analyze the information of existing institutions that provide the patent information service.

V. Conclusions

In this paper, we propose a big data-based patent analysis system. The proposed system aggregates continuously generated patent data using big data-based technologies such as Hadoop and converts various types of patent data into the same type of data using a common patent data model. It also provides an environment for efficient patent data analysis by using the converted patent data. The proposed system is expected to be of great help to the patent data research and analysis. Using the proposed system, we have converted and processed about 150 million patent data around the world. In the future, we will aggregate additional data and convert them, and conduct the analysis research based on the results by using various approaches.

REFERENCES

[1] Griliches, Z., 1990, "Patent statistics as economic indicators: a survey," National Bureau of Economic Research, No. w3301

[2] Jaffe, A. B., Trajtenberg, M., and Henderson, R., 1993, "Geographic localization of knowledge spillovers as evidenced by patent citations," National Bureau of Economic Research, 108(3), pp. 577-598

[3] Young, G. P., 2010, "The Study of Patent Misuse in respect of Patent Troll," 59(7), pp. 166-205

[4] Daim, T. U., Rueda, G., Martin, H., and Gerdsri, P., 2006, "Forecasting emerging technologies: Use of bibliometrics and patent analysis," Technological Forecasting and Social Change, 73(8), pp. 981-1012

[5] Hadoop. http://hadoop.apache.org

[6] Shin, J., Lee, S., and Wang, T., 2017, "Semantic Patent Analysis System Based on Big Data," In Proceedings of Semantic Computing (ICSC), pp. 284-285

[7] Paul, Z., and Chris, E., 2011, "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data," IBM

[8] Ho, J. P., 2012, "Big Data and In-Memory Computing Technologies," Korea Information Processing Society Review, 19, pp. 45-53

[9] Ghemawat, S., Gobioff, H., and Leung, S. T., 2003, "The Google File System," In Proceedings of the nineteenth ACM symposium on Operating Systems principles, 37(5), pp. 29-43

[10] Dean, J., and Ghemawat, S., 2008, "MapReduce: Simplified Data Processing on Large Clusters," Communications of the ACM , 51(1), pp. 107-113

[11] Borthakur, D., 2007, "The Hadoop Distributed File System : Architecture and Design," The Apache Software Foundation

[12] Hunt, D., Nguyen, L. D., and Rodgers, M., 2007, "Patent Searching Tools & Techniques," Wiley

[13] Junegak, J., and Kim, K., 2017, "Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data," Technological Forecasting and Social Change 114, pp. 281-292

[14] Guan, J., and Liu, N., 2016, "Exploitative and exploratory innovations in knowledge network and collaboration network: A patent analysis in the technological field of nano-energy," Research policy, 45(1), pp. 97-112

[15] Tang, J., 2012, "PatentMiner: topic-driven patent analysis and mining," In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1366-1374

[16] Tseng, Y. H., Lin, C. J., and Lin, Y. I., 2007, "Text mining techniques for patent analysis," Information Processing & Management, 43(5), pp. 1216‑1247

[17] Altuntas, S., Dereli, T., and Kusiak, A., 2015, "Analysis of patent documents with weighted association rules," Technological Forecasting and Social Change 92, pp. 249-262

[18] Jeong, E., Kim, Y., Lee, S., Kim, Y., and Cxhang, Y., 2014, "Identifying Emerging Free Technologies by PCT Patent Analysis," The Journal of the Korea Institute of Electronic Communication Sciences, 9(1), pp. 111-122

[19] Anthony, F.J., 2017, "Patent Citations Analysis and Its Value in Research Evaluation: A Review and a New Approach to Map Technology-relevant Research," Journal of Data and Information Science, 2(1), pp. 13-50

[20] Sun, L., and Song, Y., 2008, "Research on clustered patent mapping visualization and interaction," In Proceedings of 9th International Conference on Computer-Aided Industrial Design and Conceptual Design, pp. 1130-1133

[21] Lai, Y., Che, H., and Wang, S., 2008, "Managing Patent Legal Value via Fuzzy Neural Network Incorporated with Factor Analysis Based on Patent Infringement Lawsuits," In Proceedings of International Conference on Wireless Communications, Networking and Mobile Computing, pp.1-6

[22] Indukuri, K., Mirajkar, P., and Sureka, A., 2008, "An Algorithm for Classifying Articles and Patent Documents Using Link Structure," In Proceedings of International Conference on Web-Age Information Management, pp. 203-210

[23] KIPRIS, http://www.kipris.or.kr/

[24] Keywert, http://www.keywert.com

[25] Wipson, http://www.wipson.com

[26] Wintellips, http://www.wintelips.com

[27] Wisdomain, http://www.wisdomain.com

[28] TotalPatent, http://www.lexisnexis.com/totalpatent

[29] Thomson innovation, http://info.thomsoninnovation.com