Analysis of Patents in Cyber Security with Text Mining

Hatice Işık Özata, Önder Demir, and Buket Doğan

Abstract—**In the last decade, studies in the field of information security have progressed very rapidly. The advancement and development of new technologies in the field of information security have not only affected existing areas in the field but have also led to the emergence of new areas. Examining patent documents to monitor technological developments and track trends is an important way to make future predictions about technology. By doing so, companies know which areas they should invest in, and they can more easily predict in which direction research and development should proceed. To date, many studies have been conducted to analyze and share the patents obtained in different fields. This study will include advantages over current studies in the field of patent data and data mining. Although the patent analysis method used in the present study has recently become popular, it is expected to be a valuable resource for Turkey because the lack of technology prediction methods in the field of information security. In this study, patent documents obtained in the field of information security are analyzed. Data mining will be used to analyze patent documents retrieved from the field of information security. Furthermore, trends in the field of information security will be identified from the collected data.**

Index Terms—**Data mining, information security, patent analysis, text mining.**

I. Introduction

Recently, information security has become a complex construct that concerns software, hardware, and human components. Hardware comprises the computers that we use; software comprises the code, databases, and applications that we use; and the human component comprises laws, procedures, and education [1]. The methods of providing information security have progressed very rapidly over the last 10 years, and the amount of new technology related to information security has increased substantially. The advancement and development of these technologies have not only affected existing technology but have also led to the emergence of new areas of study.

Multiple data sources can be used to monitor technological advances. Publications, social media, and patent documents are just a few of these data sources. Patent documents are among the sources that provide the most comprehensive data. Therefore, examining patent documents to keep track of technological trends is an important way to make predictions. New generation technologies can be easily identified through information in a patent file, such as application date, applicant title, and international patent code (IPC).

In this study, the International Patent Documentation (INPADOC) database of the European Patent Office (EPO) was referenced to prepare the data set. The processing of patent data is a long and complex procedure that requires preparation to extract available data. Data sets can be prepared for different purposes by analyzing specific pieces of information, such as the application number, date, IPC, abstract, and claim.

Patent analysis is used to finalize lawsuits, compare patents, and make accurate research and development (R&D) investments. Thanks to patent analysis, companies can easily see their way. Patent analysis is essential for a healthy R&D.

This study aims to determine the trends in information security based on the patent data collected by using text mining algorithms and architectures. The manuscript begins with a study that was conducted to create the data set. Then a bibliography of the documents containing the obtained patents, are used for analyses in the field of information technology. This study is the initial stage of a long-term project. As such, this study is expected to reveal in which areas technology is advancing in information security. In the future project phase, artificial intelligence and various analysis methods will be applied, and future predictions will be made.

II. Literature Review

Data mining – also referred to as knowledge discovery from data – is the automated or convenient extraction of patterns representing knowledge from data that is implicitly stored or captured in large databases, data warehouses, the web, other massive information repositories, or data streams. In recent years, data mining techniques have been applied extensively in various fields, such as public administration, business, finance, education, and transportation [2].

The first step in data mining is to collect data from designated databases. The second step is to normalize the data to prepare it for data mining. Data mining is performed during the third step, and different models are tested on the data set. According to this study, data mining shows the relationships between data [3]. Data mining techniques are widely used in many areas for many purposes, such as predicting short-term adverse events in phototherapy treatments [4] or predicting index and non-index crimes in cities [5].

Text mining is used to extract meaningful information from unstructured or semi-structured texts. High-quality information is acquired by finding patterns and trends through means such as statistical pattern learning. Text mining is also often referred to as text data mining or text analytics. Text mining has 10 steps including text document gathering, text preprocessing, text cleanup, tokenization, removal of stop words, stemming, text transformation,

Manuscript received June 15, 2020; revised September 15, 2020.

Hatice Işık Özata, Önder Demir, and Buket Doğan are with Marmara University, Technology Faculty, Istanbul, Turkey (e-mail: hatice.isik@marun.edu.tr, {odemir, buketb}@marmara.edu.tr).

DOI: 10.7763/IJCTE.2021.V13.1285

feature selection, text mining techniques and evaluation [6].

Text mining techniques are combined with data mining techniques and applied to the text to discover patterns. There are various text mining techniques such as information extraction, information retrieval, summarization, clustering, classification. In summary, the text mining process is described in [7].

Patent analysis is the process of proposing and forwarding a new qualitative research method. This method focuses on the systematic analysis, description, and interpretation of a chosen patent in any area. Such an analysis process might lead to the development of new concepts or theories. A patent analysis involves an evaluation of a patent in terms of its advantages, benefits, constraints, disadvantages, effectiveness, and future value. Further, the use and application of patent analysis in strategic organizational decisions for foreseeing the new technologies are also discussed.

Patents defining new technology and training are considered big data. In the last 10 years 2,500,000 patents have been filed all over the world [8]. This article also mentions that patent applications have increased year by year. For example, 1,457,000 patents were filed in 2001, and this number increased to 3,127,900 in 2016. Many patents have been published over many years on different topics, and so there is a great opportunity for researchers to analyze and interpret them so that they can access new information. The reasons for patent analysis are to anticipate new technologies, mapping technological changes, evaluating the risk of patent infringement, defining technology orientations, prediction of the effects of technology, prediction of the technological issues, monitoring trends, technology measurement, following new inventions, making business plans and following new ideas.

The rapid growth of patent documents has called for the development of sophisticated patent analysis tools. Various tools are currently being utilized by organizations to analyze patents. These tools can perform a wide range of tasks, such as analyzing and forecasting future technological trends, conducting strategic technology planning, detecting patent infringements, determining patent quality and establishing the most promising patents, and identifying technological hotspots and patent vacuums [9].

This study describes the state-of-the-art in patent analysis and presents a taxonomy of patent analysis techniques. Moreover, the key features and weaknesses of the discussed tools and techniques are presented, and several directions for future research are highlighted. In the study, the ScienceDirect, ACM Digital Library, IEEE Xplore, and CiteSeerX databases were used for patent analysis. “Patent analysis tools and techniques,” “visualization approaches in patent analysis,” and “text mining and patent analysis” were the keywords used. Patents obtained in recent years were selected, and 22 articles were chosen for the literature review. A text analysis was performed to obtain meaningful data from the patent articles, and visualization approaches were used for the results. In these approaches, the k-means algorithm was used to aggregate patents. Even though the techniques for patent analysis have matured, there are still areas that need improvement. Moreover, hybrid patent-retrieval approaches can also be utilized to search for documents other than patents, such as journal articles. Furthermore, the patent analysis approaches for strategic technology planning that have been developed can suggest only one strategy. It would be beneficial to managers if the approaches were more efficient and flexible so that multiple suggestions could be offered for devising strategies.

Several studies discuss examples of patent analysis. In 2018, a study on monitoring new developments in the field of data mining using data on Twitter and patent analysis in the field of perovskite solar cell technology was published. Firstly, patents related to perovskite solar cell technology were downloaded from the Derwent Innovations Index (DII) database, and a data set was created. Then, the data set was cleared, the unrelated data points were deleted, and the patents to be included in the study were converted into a text format for text mining. Then, the Lingo algorithm was run on the prepared data set to draw conclusions [7].

To work with Twitter, first, Twitter data were collected, and related data were downloaded. Unnecessary and repeated tweets were deleted, and the data were classified according to time. They were also used to analyze the professionalism and interests of Twitter users. In the study, they defined the perceptions, reactions, and expectations of Twitter users regarding emerging technology. The results obtained from the Twitter data of the patents were compared, and it was found that Twitter discussions started in 2013 compared to the patent data which was started in 2012 in the field of solar cell technology. From 2012 to 2016, the patents received were relevant to several areas related to solar cell technology over the years. For example, while patents were granted for the “organic materials” category in 2012, the name of this category changed to “organic-inorganic hybrid materials” in 2016 [10].

Another study published on patent analysis predicted promising technology to be used in the construction of high-rise buildings [11]. In this study, 2875 articles from the United States, Europe, Korea, China, and Japan published between 1995 and 2013 on the topic of high-rise buildings were examined. The object-solution matrix analysis was used for technology estimation. It was found that Korea and China have made progress in this field of technology. China has shown especially strong growth.

Kim and Bae proposed a new method for predicting new healthcare technologies by using patent analysis [12]. The general process of the proposed method consists of three steps. The first step was to classify technologies based on common patent classifications to form technology clusters. Secondly, the structures of the clusters created in the first step were examined. Finally, patent indicators, such as triadic patent families and independent patents, were analyzed to assess whether the established sets of patents were viable. They determined that telemedicine technology was the most promising technology set, as it has relatively high indicator values. They identified other promising technology clusters, such as technologies related to health care business systems, data management, and telemedicine.

An analysis of Apple’s patents was previously carried out to estimate the innovations made by Apple [13]. This method is seen as an objective approach to follow Apple's developments. To examine Apple's technological innovations, the authors examined all the patents applied and created models based on three different approaches. First, they used statistical models to map technology with time series regression and multiple linear regression methods. Second, they clustered all patents to explore Apple's idle technology areas. Finally, they used social network analysis to analyze technologies at the center of Apple's future plans.

In this research, it was also found that Apple uses technologies that focus on unexplored technology. The results indicate that Apple's most basic technology is G06F; electrical digital data processing. Patents are divided into five clusters, and since two of the five clusters have few elements, empty technology areas have been selected. The generated clusters were examined by the researchers, and G06F, H04B, and H04N IPC codes were found in all clusters. G06F, H04B, and H04N technologies are related to "electrical digital data processing," "transmission," and "pictorial communication." When the patent areas of the cluster that can be considered empty are examined, electronic music, and speech communication and processing technologies were found to be focus areas. Therefore, Apple needs to develop these technologies more than others.

III. Methodology

Text mining is the process of searching for and analyzing large amounts of unstructured text data supported by software that can identify concepts, patterns, topics, keywords, and other attributes of the data [14]. The specific technology field must be selected before starting operations. The field of information security affects many technology fields and is growing rapidly. This study is intended to uncover the relationships, differences, and technological transitions between technologies in the field of information security. This process starts with the collection of patents.

A. Searching Patent Databases

To access patent documents online, the Espacenet database was used. Espacenet offers free access to information on inventions and technical developments from 1782 to the present, with worldwide coverage. This database is updated daily by experts and contains data on more than 110 million patents from around the world. To collect data, the International Patent Documentation Center (INPADOC) database (OPS) was used. The INPADOC database is a collection of bibliographic data for patents. OPS is an international patent office database search service. The search services of the international patent office include Espacenet, OPS, and the Global Patent Index (GPI).

B. Getting Advice from Patent Experts

Patent experts were asked to propose different methods for identifying keywords because searching patent documents requires expertise. Getting help from experienced people is very important to create an appropriate data set. For example, when we searched for the word "security" in the titles of patent documents, we found not only patents related to information security but also patents covering different security areas such as hats and keys. For this reason, in this study, a patent search was done as recommended by patent experts, and a detailed data set was developed by following the strategies provided by the experts. The strategies we used are explained in detail in the following sections.

C. Identifying Leading Companies in Information Security

To access patents obtained in the field of information security, it is essential to use suitable keywords in the patent search. The most appropriate way to choose the right words is to take advantage of the patents previously obtained in this field. To gain access to the patents previously obtained in the field of information security, patents obtained by the leading companies in this field were examined. In this way, the words that these companies frequently use in their patent documents can be determined. An internet search revealed that the companies that invested the most in the field of information security in recent years are Cisco, IBM, Microsoft, Amazon, and Symantec. To identify the key words that will be used in the patent search phase of our study to form the data set, the patents obtained by these five companies in the last three years were examined.

D. Identifying the Most Popular IPC Classes in the Field of Information Security

IPC classes needed to be added to the query to access the patents obtained by the companies mentioned above. In this way, patents directly related to the relevant field could be accessed. In this study, H04L9 and G06F21 IPC codes were used in the queries created to determine keywords. These classes were found by reading the definitions of IPC classes and selecting the class closest to the area to be analyzed.

For class H04L9, the class "H" is the electrical division, the sub-class "H04" represents the electrical communication technique, and the field of transmission of digital information is indicated by the further subclass "H04L." The patents related to the regulations required for confidential or secure communication are denoted by "H04L9." G06F21 encompasses patents for security regulations to protect computers, their components, programs, or data from unauthorized activities. Such patents are classified as being in the electrical digital data processing area (G06F) of the class of calculations (G06), which is a subgroup of the physics (G) class.

E. Creating Queries Required to Identify Keywords

To determine commonly used words and phrases in the patents obtained in the last three years in the field of information security and popular IPC classes determined by the methods mentioned above, the following query sentences were created.

```
'pa="CISCO TECH INC" and pd within "2017 2019"
and (ipc="H04L9")'
'pa="CISCO TECH INC" and pd within "2017 2019"
and (ipc="G06F21")'
```

In the sample queries, "pa" filters the applicant parameter, "pd" represents the publication date, and "ipc" represents the IPC class.

F. Identifying Keywords

After running the queries through the web service to determine the keywords, 3283 patents related to the two IPC classes obtained by the five companies in the last three years were found. In the titles of 3283 unique patents, frequently used words and phrases (and combinations of these) were found. To determine these combinations, the 34,025 words in the titles of the documents that made up the preliminary data set were analyzed. As a result of the analysis, frequently used word groups were determined. These words were searched for in the titles of patents in the queries to be used to form the data set.

According to the results of the analysis, the word groups are as follows;

GROUP A -1: hardware*, software*, malware*

GROUP A-2: serv*, comput*, execut*

GROUP A-3: virtual*, digital*, data*, system*

GROUP A-4: inform*, network*, device*, method*, applicat*

GROUP B-1: verificat*, authenticat*, authoriz*, block* chain*, blockchain*

GROUP B-2: crypto*, encrypt*, password*, encod*, key*

GROUP B-3: protect*, secre*, secur*

GROUP B-4: priva*, safe*, sure*, policy*

Groups A and B were crossed, and queries were formed using data from the last 10 years.

G. Data Collection

Using the keywords, a data set consisting of patents obtained from patent titles in the last 10 years (specifically, from January 2009 to May 2019) was created. In this study, a total of 27,290 queries were run to form the data set. Prepared queries were run automatically through a web service, and XML files containing the bibliographic data of the patents were downloaded.

A sample query is (ti any "hardware* software* malware*") and (ti any "verificat* authenticat* authoriz* block* chain* blockchain*") and (pd = "20090101 20090104 ").

As a result of utilizing the queries, 28,4972 patent files were downloaded. The downloaded patent bibliographic files were analyzed using XML file parsing methods.

IV. Results

When the downloaded XML files were separated based on the parts of the patents, it was seen that different results were obtained according to years, countries, firms and codes.

A. Analysis of the Data Set by Year

When we examined the patents obtained in the field of information security between January 2009 and May 2019, we saw that the number of results has increased gradually from 2009 to 2019. While less than 20,000 patents were procured in 2009 in the field of information security, this number increased to over 40000 in 2019.

B. Analysis of the Data Set by Country

When the data set was analyzed by country, it was seen that 56 countries have obtained patents in the field of information security in the last 10 years. Of these countries, China, the US, and Korea are the top three. Number of patents in the dataset according to countries is presented in Table I.

TABLE I: Data Set by Country

Country	Number of Patents	Rank
China	116907	1
USA	66397	2
South Korea	19245	3
European Union	16650	4
Japan	13149	5
Taiwan	5954	6
Canada	4196	7
Australia	3411	8
Germany	2243	9
Russia	2208	10
UK	1794	11

C. Analysis of the Data Set by Firm

When the data set collected was examined based on the applicant's firm, it was seen that some firms have filed substantially more applications than others. Number of patents in the dataset according to firms is presented in Table II.

TABLE II: Data Set by Firm

Firm	Number of Patents	Rank
Huawei	7739	1
Samsung	7532	2
IBM	7233	3
ZTE	6522	4
Intel	4649	5

D. Analysis of the Data Set CPC

The analysis results, according to the cooperative patent classification (CPC) subcodes are presented in Table III.

TABLE III: Data Set by CPC

CPC	Definition	Number of Patents	Rank
H04L	Transmission of Digital Information	306581	1
G06F	Electric Digital Data Processing	195895	2
H04W	Wireless Communication Networks	81370	3
G06Q	Data Processing Systems or Methods	76363	4
H04N	Pictorial Communication	72796	5
G06K	Recognition of Data	25931	6
G08B	Signaling or Calling Systems	15321	7
H04M	Telephonic Communication	13642	8
A61B	Diagnosis, Surgery, Identification	13137	9
G07C	Time or Attendance Registers	12492	10

E. Analysis of the Data Set IPC

The analysis results, according to the IPC subcodes are presented in Table IV.

TABLE IV: Data Set by IPC

IPC	Definition	Number of Patents	Rank
H04L	Transmission of Digital Information	5257	1
G06F	Electric Digital Data Processing	4230	2
H04W	Wireless Communication Networks	2323	3
G06Q	Data Processing Systems or Methods	2051	4
H04N	Pictorial Communication	1616	5
G08B	Signaling or Calling Systems	551	6
G06K	Recognition of Data	536	7
C12N	Microorganisms or Enzymes	403	8
G10L	Speech Analysis or Synthesis	399	9

V. Conclusion

This study analyzed patents obtained in the field of information security to show how technology is developing in this field. The first stage of the study was conducted to collect the data set. The details of how the keywords required to generate the data set were determined are described above. At the end of the first stage, a data set of 284,972 patents was obtained. In the second stage, the data set was analyzed, and results were obtained based on year, country, firm, and patent class.

This study is the initial stage of a long-term project that will show in which areas information security technology is advancing. Artificial intelligence and various analysis methods will be explored, and predictions will be made during the later stages of the project.

In the final stage of the project, social network analysis will be applied. The relationship between patents via a social network analysis will be presented using the network visualization method. Network visualization can be used to illustrate the relationship between different aspects of a patent, such as IPC, country, year, applicant, and CPC. Information about the relationships between patents gained through a social network analysis will provide insights into future technologies.

Although the patent analysis method to be used in the study is popular today, the limited number of related publications in Turkey and the lack of technology estimations in the field of information security make the study of patent analysis crucial.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

ÖD conducted the research. HIÖ prepared the dataset and analyzed the data. BD evaluated the results. All authors wrote the paper together and approved the final version.

Acknowledgment

This work was supported in part by Marmara University BAPKO FEN-C-YLP-130319-0063.

References

[1] J. L. Heimerl. (2012). The evolution of information security. [Online]. Available: https://www.securityweek.com/evolution-information-security

[2] J. Liu, X. Kong, X. Zhou, L. Wang, D. Zhang, I. Lee, B. Xu, and F. Xia, "Data mining and information retrieval in the 21st century: A bibliographic review," *Computer Science Review*, vol. 34, p. 100193, 2019.

[3] J. Chen, W. Wei, C. Guo, L. Tang, and L. Sun, "Textual analysis and visualization of research trends in data mining for electronic health records," *Health Policy and Technology*, vol. 6, no. 4, pp. 389-400, 2017.

[4] S. Mohamed, A-M. Tobin, A. D. Irvine, D. R. Wall, N. J. O'Hare, and M-T. Kechadi, "The application of data mining to predict the occurrence of short-term adverse events in NB-UVB phototherapy treatments," *International Journal of Machine Learning and Computing*, vol. 8, no. 2, pp. 104-111, 2018.

[5] A. J. P. Delima, "Applying data mining techniques in predicting index and non-index crimes," *International Journal of Machine Learning and Computing*, vol. 9, no. 4, pp. 533-538, 2019.

[6] G. Miner, J. Elder, A. Fast, T. Hill, R. Nisbet, and D. Delen, *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*, Academic Press, 2012.

[7] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," *Expert Systems with Applications*, vol. 41, no. 16, pp. 7653-7670, 2014.

[8] P. S. Aithal and S. Aithal, "Patent Analysis as a new scholarly research method," *International Journal of Case Studies in Business, IT, and Education (IJCSBE)*, vol. 2, no. 2, pp. 33-47, 2018.

[9] A. Abbas, L. Zhang, and S. U. Khan, "A literature review on the state-of-the-art in patent analysis," *World Patent Information*, vol. 37, pp. 3-13, 2014.

[10] X. Li, Q. Xie, J. Jiang, Y. Zhou, and L. Huang, "Identifying and monitoring the development trends of emerging technologies using patent analysis and Twitter data mining: The case of perovskite solar cell technology," *Technological Forecasting and Social Change*, vol. 146, pp. 687-705, 2019.

[11] H. P. Cho, H. Lim, D. Lee, H. Cho, and K. I. Kang, "Patent analysis for forecasting promising technology in high-rise building construction," *Technological Forecasting and Social Change*, vol. 128, pp. 144-153, 2018.

[12] G. Kim and J. Bae, "A novel approach to forecast promising technology through patent analysis," *Technological Forecasting and Social Change*, vol. 117, pp. 228-237, 2017.

[13] S. Jun and S. Sung Park, "Examining technological innovation of Apple using patent analysis," *Industrial Management & Data Systems*, vol. 113, no. 6, pp. 890-907, 2013.

[14] S. E. Seker. (2011). Apriori Algoritması. [Online]. Available: http://bilgisayarkavramlari.sadievrenseker.com/2011/09/07/apriori-algoritmasi/

Hatice Işık Özata was born in Kütahya, Turkey, in 1988. She got the bachelor's degree in computer engineering from İzmir University of Economics, İzmir, Turkey from 2006 to 2011.

She is currently a student for a master's degree in computer engineering in the Institute of Pure and Applied Sciences, Marmara University. Her research interests include data mining, machine learning and deep learning.

Önder Demir received the MS and PhD degrees in electronics and computer education from Marmara University in 2006 and 2013, respectively. From 2003 to 2013 he worked as a research assistant and lecturer. He has been working as an assistant professor in Computer Engineering Department of Technology Faculty. His research interests are digital image processing, biomedical image processing and algorithms.

Buket Dogan received the MS and PhD degrees in computer-control education from Marmara University in 2001 and 2006, respectively. From 1999 to 2007 she worked as a research assistant. She has been working as an assistant professor in Computer Engineering Department of Technology Faculty. Her research interests include adaptive intelligent web based educational systems, data mining and digital image processing.