THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

A TECHNOLOGY MINING SYSTEM BASED ON PATENT DOCUMENTS: A CASE STUDY OF CLOUD COMPUTING

Master's Thesis

BİLAL ALP

ISTANBUL, 2016

THE REPUBLIC OF TURKEY

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

COMPUTER ENGINEERING

A TECHNOLOGY MINING SYSTEM BASED ON PATENT DOCUMENTS: A CASE STUDY OF CLOUD COMPUTING

Master's Thesis

BİLAL ALP

Supervisor: YRD. DOÇ. DR. CEMAL OKAN ŞAKAR
YRD. DOÇ. DR. SERCAN ÖZCAN

ISTANBUL, 2016

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
COMPUTER ENGINEERING

Name of the thesis: A Technology Mining System Based On Patent Documents: A Case Study Of Cloud Computing
Name/Last Name of the Student: Bilal Alp
Date of the Defense of Thesis: 31/08/2016

The thesis has been approved by the Graduate School of Natural and Applied Sciences.

Assoc. Prof. Nafiz ARICA
Graduate School Director

I certify that this thesis meets all the requirements as a thesis for the degree of Master of Science.

Asst. Prof. Tarkan AYDIN
Program Coordinator

This is to certify that we have read this thesis and we find it fully adequate in scope, quality and content, as a thesis for the degree of Master of Science.

Examining Comitee Members Signatures

Thesis Supervisor ---------------------------------
Asst. Prof. Cemal Okan Şakar

Thesis Co-supervisor ---------------------------------
Asst. Prof. Sercan Özcan

Member ---------------------------------
Asst. Prof. Görkem SERBES

Member ---------------------------------
Assoc. Prof. Mehmet Alper TUNGA

Member ---------------------------------
Asst.Prof. Tarkan Aydın

ABSTRACT

A TECHNOLOGY MINING SYSTEM BASED ON PATENT DOCUMENTS: A CASE STUDY OF CLOUD COMPUTING

Bilal Alp

Computer Engineering

Thesis Supervisor: Asst. Prof. Cemal Okan Şakar

September 2016, 62 Pages

This study deals with the tech mining in cloud computing technology field. Cloud computing is a rising field in last ten years. Cloud computing was examined with its related technology fields. Although it has several subfields such as Saas, Paas, Iaas and MBaas, this study aims to find other related fields that use Cloud computing as a subfield.

Nowadays humanity has seen the limitations of a CPU, so they have started to look for other solutions such as virtualization and unlimited resources of computers. A single computer does not provide enough futures anymore. Then, people build systems serve as services because the hardware behind the services can be enormous. Thus, these reasons made cloud computing so popular and it has being used in many other fields such as instance, chemistry, astronomy, medical and etc.

This study aims to find related technologies with cloud computing by analyzing patent documents. A patent document may contain much information about a technology and its related technologies. By looking a patent document and its keywords, this study tries to predict the technology with using data mining techniques.

The findings of this study show that K-means clustering algorithm works well with the big datasets. For the similarity metric, cosine distance gives the most effective results. To score the words, TF-IDF and Term Variance algorithms compared and Term Variance gave the better results. At the end of this study, it can be seen that cloud computing technology fields and their related fields can be seen clearly.

Keywords: Cloud Computing, Tech Mining, Patent Mining

ÖZET

PATENT DÖKÜMANLARINA DAYALI BİR TEKNOLOJİ MADENCİLİĞİ SİSTEMİ: BULUT BİLİŞİM ÜZERİNE BİR VAKA ÇALIŞMASI

Bilal Alp

Bilgisayar Mühendisliği

Tez Danışmanı: Yrd. Doç Dr. Cemal Okan Şakar

Eylül 2016, 62 Sayfa

Günümüzde en hızlı gelişen teknolojilerden bir tanesi de işlemci teknolojisidir. Transistör boyutlarının atom seviyesine gelmesinden dolayı artık yeterince yol katedilememektedir. Bu noktada devreye bulut sistemler girmiştir. Kişisel bir bilgisayarın gücüyle yapılabilecekler sınırlı olduğundan insanlara bir çok iş için daha fazla güç gerekir hale gelmiştir. Bu sebeple insanlar işlerini paralel ya da bilgisayar topluluklarıyla halletmektedir.

Bulut bilişim son 10 yılın en popüler teknolojilerinden biridir ve bir çok alanda kullanılmaktadır. Örneğin; tıp, kimya, astronomi, astroloji ve daha bir çok alan. Bu çalışma bulut bilişimin alt teknolojileri yerine teknolojinin kendisinin kullanıldığı alanları tespit etmeyi amaçlamıştır. Çalışma yapılırken patentler ve patent metinleri kullanılmıştır. Bir patent amaçladığı ve kullandığı teknolojiyle ilgili bir çok bilgi vermektedir. Bu çalışma da bulut bilişim patentlerini ve metinlerini analiz edip kullanıldığı alanları bulmayı amaçlamaktadır.

Birinci bölümde patentlerin nasıl bulunduğu ve nasıl işlendiğinden detaylıca bahsedilmiştir. Bu kısımda yazı işleme teknikleri kullanılmıştır. İkinci bölümde patentlerin kümelenmesi ve görselleştirilmesi konu edilmiştir. Bu bölümde çeşitli kümeleme yöntemleri denenmiştir. Son bölümde ise hangi kümeleme yönteminin en iyi sonucu verdiği, hangi teknoloji alanlarının ilişkili olarak çıktığından bahsedilmiştir.

Çalışmada kelime skorlama için TF-IDF ve Term Variance teknikleri kullanıldı. Term Variance yönteminin daha iyi çalıştığı görüldü. Kümeleme için K-Means algoritması en iyi sonucu vermiştir. Benzerlik metriği olarak Cosine metriği kullanılmıştır. Çalışma sonucunda bulut bilişim teknolojinin ilgili teknolojileri ve ilişkili olduğu teknoloji alanları ortaya çıkmıştır.

Anahtar Kelimeler: Bulut Bilişim, Teknoloji Madenciliği, Patent Madenciliği

CONTENTS

TABLES

FIGURES

ABBREVIATIONS

AWS	: Amazon Web Services
CPU	: Central Processing Unit
DF	: Document Frequency
EPO	: European Patent Office
FPO	: Free Patent Online
HTML	: HyperText Markup Language
IAAS	: Infrastructure as a Service
IDF	: Inverse Document Frequency
IPC	: International Patent Code
JPO	: Japan Patent Office
MBAAS	: Mobile Backed as a Service
NLP	: Natural Language Processing
PAAS	: Platform as a Service
PCA	: Principal Component Analysis
R and D	: Research and Development
RAM	: Random Access Memory
SAAS	: Service as a Service
TF	: Term Frequency
TLC	: Technology Life Cycle
TPE	: Turkish Patent Office
TV	: Term Variance
USPTO	: United States Patent and Trademark Office
WIPO	: World Intellectual Property Organization

1. INTRODUCTION

Today, research and development (R and D) is the key to being a successful company. It requires high budget although it is necessary. Many companies have small budgets for investigations and innovations. Investment to the new field requires a serious budget moreover making an innovation or producing a product does not guarantee the place of a company in the market because being updated and keeping in step with new developments remains a challenge. Nowadays, the companies which make an investment, lead the technology world. It can be seen top 20 R and D spenders in figure 1.1

Figure 1.1: Top 20 R and D spenders in the world

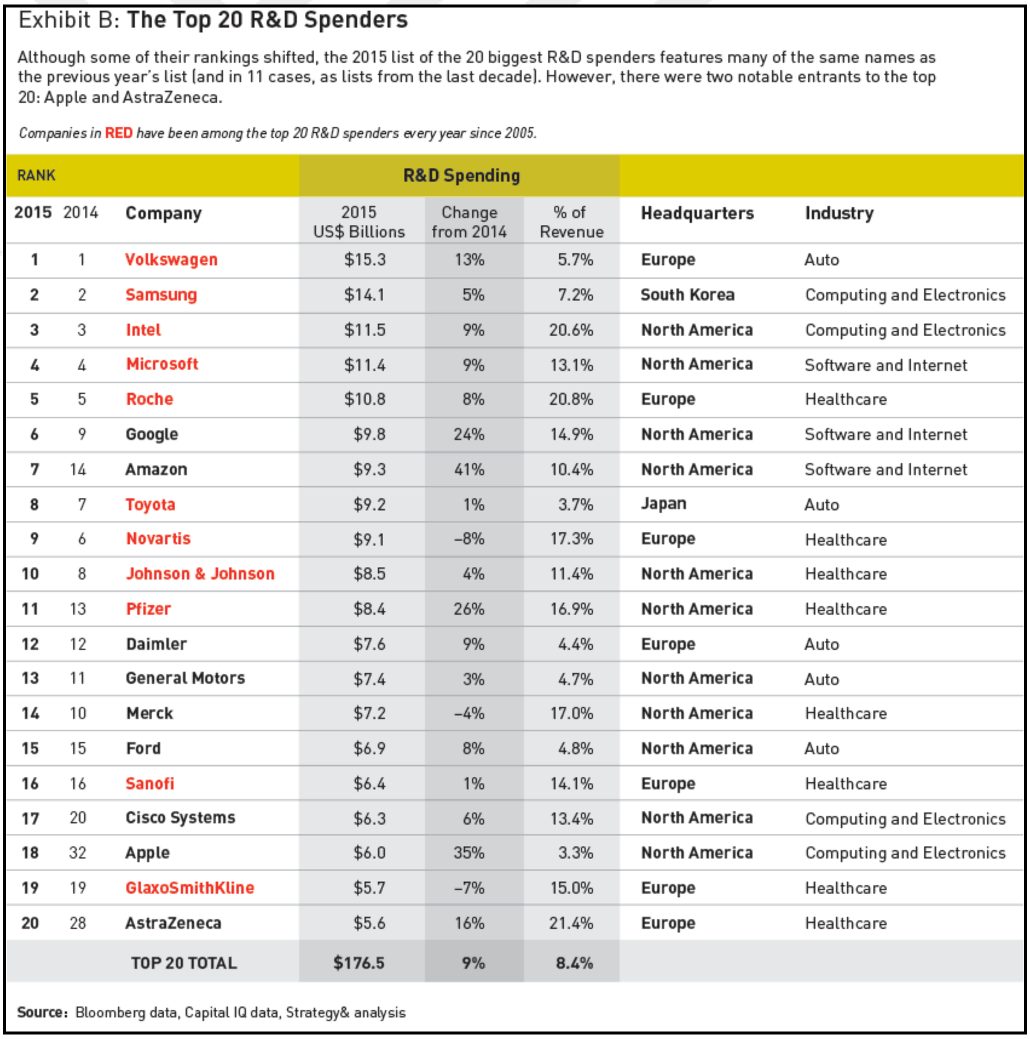


Exhibit B: **The Top 20 R&D Spenders**

Although some of their rankings shifted, the 2015 list of the 20 biggest R&D spenders features many of the same names as the previous year's list (and in 11 cases, as lists from the last decade). However, there were two notable entrants to the top 20: Apple and AstraZeneca.

*Companies in **RED** have been among the top 20 R&D spenders every year since 2005.*

RANK			R&D Spending				
2015	2014	**Company**	2015 US$ Billions	Change from 2014	% of Revenue	**Headquarters**	**Industry**
1	1	**Volkswagen**	$15.3	13%	5.7%	**Europe**	Auto
2	2	**Samsung**	$14.1	5%	7.2%	**South Korea**	Computing and Electronics
3	3	**Intel**	$11.5	9%	20.6%	**North America**	Computing and Electronics
4	4	**Microsoft**	$11.4	9%	13.1%	**North America**	Software and Internet
5	5	**Roche**	$10.8	8%	20.8%	**Europe**	Healthcare
6	9	**Google**	$9.8	24%	14.9%	**North America**	Software and Internet
7	14	**Amazon**	$9.3	41%	10.4%	**North America**	Software and Internet
8	7	**Toyota**	$9.2	1%	3.7%	**Japan**	Auto
9	6	**Novartis**	$9.1	–8%	17.3%	**Europe**	Healthcare
10	8	**Johnson & Johnson**	$8.5	4%	11.4%	**North America**	Healthcare
11	13	**Pfizer**	$8.4	26%	16.9%	**North America**	Healthcare
12	12	**Daimler**	$7.6	9%	4.4%	**Europe**	Auto
13	11	**General Motors**	$7.4	3%	4.7%	**North America**	Auto
14	10	**Merck**	$7.2	–4%	17.0%	**North America**	Healthcare
15	15	**Ford**	$6.9	8%	4.8%	**North America**	Auto
16	16	**Sanofi**	$6.4	1%	14.1%	**Europe**	Healthcare
17	20	**Cisco Systems**	$6.3	6%	13.4%	**North America**	Computing and Electronics
18	32	**Apple**	$6.0	35%	3.3%	**North America**	Computing and Electronics
19	19	**GlaxoSmithKline**	$5.7	–7%	15.0%	**Europe**	Healthcare
20	28	**AstraZeneca**	$5.6	16%	21.4%	**Europe**	Healthcare
		TOP 20 TOTAL	**$176.5**	**9%**	**8.4%**		

Source: Bloomberg data, Capital IQ data, Strategy& analysis

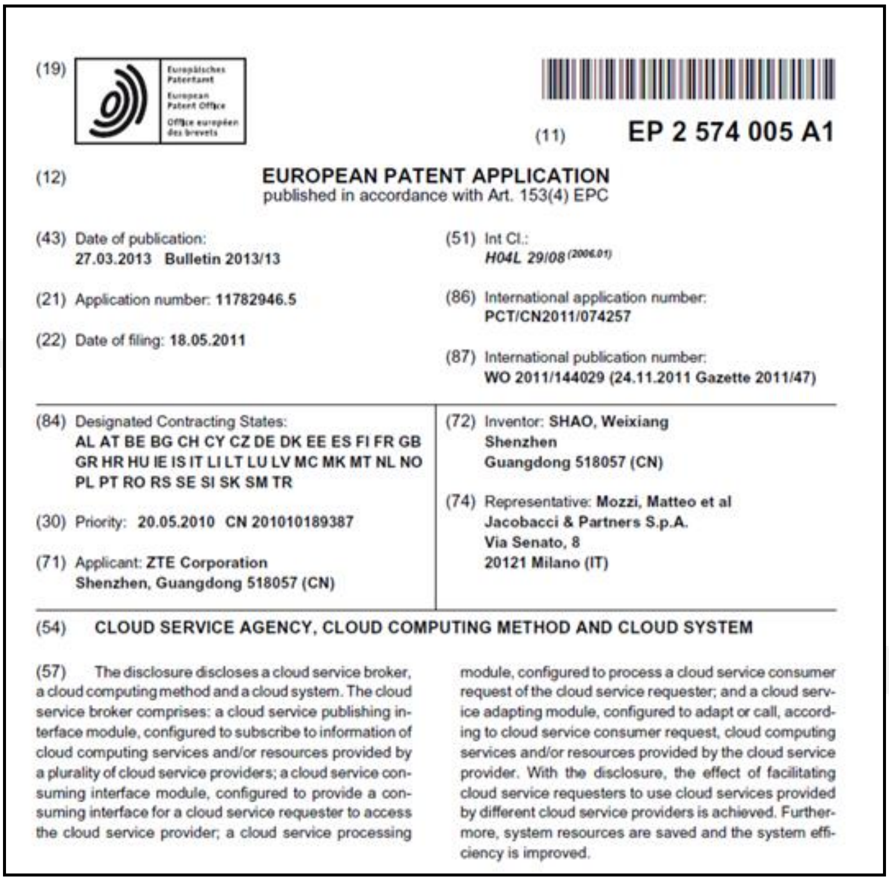
Source: *Car Companies and Their Countries*

Therefore, every company needs a system to monitor technological improvements on that product or innovation and what other competitors have implemented. Tech mining helps these companies in order to conduct a market survey perpetually by aiding them to analyze or monitor the business value in the market so they can see the rising trends of following innovations or alternative technologies potentially become popular in the future. It is very important to see the future in the market because the making right investment is very challenging which specifies the future of a company.

In tech mining, several sources can be used such as publications, social media, and patents to extract some information. Depending on the extracted information, tech mining can be used in many areas like technical intelligence, competitive intelligence, market intelligence, technology road mapping, technology forecasting, technology assessment, technology foresight (Porter and Cunningham 2004). Therefore, the source of data is so important to analyze the technology. Among the sources of tech mining, patents are very extensive inquiries since they contain much information to define and identify the technologies such as international patent code (IPC), application number, date of application, title, and abstract of a patent. Feasibility of the structural organization of patent documents evokes launching a new tech mining area called patent mining (Kasravi and Risov 2007). Moreover, the people who use patent mining method to process the patents named as patent experts. By definition, patent mining is a method used for processing, extracting and interpreting the information from patents to analyze the current and future technology tendencies. An example of the patent document can be seen in the following figure.

In this figure, it can be seen that there are many fields. A patent document contains all information about a product or an innovation. Each patent document has a patent number. 11th field refers to the patent number. As it can be seen, this patent document was taken from the European Patent Institution. 43rd field refers to the publication data. Each patent documents are located under a specific technology field. 51st field shows where this patent is placed. 21st field shows the application number. 86th field shows the international application number. 22nd field shows the filling data. 87th field shows the international publication number. 72nd field shows the inventors of the patent. 54th field shows the title of the patent document. 57th field shows the abstract of the patent document.

Figure 1.2: A sample patent document



(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 2 574 005 A1

(12) EUROPEAN PATENT APPLICATION
published in accordance with Art. 153(4) EPC

(43) Date of publication:
27.03.2013 Bulletin 2013/13

(21) Application number: 11782946.5

(22) Date of filing: 18.05.2011

(51) Int Cl.:
H04L 29/08 (2006.01)

(86) International application number:
PCT/CN2011/074257

(87) International publication number:
WO 2011/144029 (24.11.2011 Gazette 2011/47)

(84) Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB
GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO
PL PT RO RS SE SI SK SM TR

(30) Priority: 20.05.2010 CN 201010189387

(71) Applicant: ZTE Corporation
Shenzhen, Guangdong 518057 (CN)

(72) Inventor: SHAO, Weixiang
Shenzhen
Guangdong 518057 (CN)

(74) Representative: Mozzi, Matteo et al
Jacobacci & Partners S.p.A.
Via Senato, 8
20121 Milano (IT)

(54) CLOUD SERVICE AGENCY, CLOUD COMPUTING METHOD AND CLOUD SYSTEM

(57) The disclosure discloses a cloud service broker, a cloud computing method and a cloud system. The cloud service broker comprises: a cloud service publishing interface module, configured to subscribe to information of cloud computing services and/or resources provided by a plurality of cloud service providers; a cloud service consuming interface module, configured to provide a consuming interface for a cloud service requester to access the cloud service provider; a cloud service processing module, configured to process a cloud service consumer request of the cloud service requester; and a cloud service adapting module, configured to adapt or call, according to cloud service consumer request, cloud computing services and/or resources provided by the cloud service provider. With the disclosure, the effect of facilitating cloud service requesters to use cloud services provided by different cloud service providers is achieved. Furthermore, system resources are saved and the system efficiency is improved.

Each country has its own patent institution. Also, there are some private patent institutions. A patent institution is used for controlling the rights of a patent in the country. It also helps people to take the patent of an innovation. All patents should be registered to patent institutions. Some of the patent institution which belongs the counties are listed below.

a) Patent and Trademark Office (USPTO) – USA
b) European Patent Office (EPO) – Europe
c) Canadian Intellectual Property Office (CIPO) – Canada
d) China Patent Office (SIPO) – China
e) Turkish Patent Office (TPE) – Turkey
f) Japan Patent Office (JPE) – Japan

These patent institutions open their databases free. They provide patents which can be belong to any country. Some of the private databases are listed below.

a) Google Patents
b) FreePatentsOnline
c) Medicines Patent Tool

In this study, FreePatentsOnline was used because it has access to many databases such as USPTO, EPO, JPO, German Patents, and WIPO. It does not provide any access to TPE.

Patent mining is a compelling procedure that requires many pre-processing steps to extract utilizable data. The components of patent documents like application number, IPC, date of filling are structured data which are more convenient to process comparing to unstructured components like abstract, claim. The structured parts of patent documents have been extensively used in the studies conducted so far, however, mining unstructured parts remain as a challenge. It even becomes more challenging since the unstructured data are so diverse because patents are collected from different resources such as USPTO (United States Patent and Trademark Office), Free Patent Online and WIPO (World Intellectual Property Organization). These are the well-known biggest patent websites available for patent studies.

On the other hand, since patent mining arises as a promising way to guide companies for their investments, several tools has emerged to serve this purpose. However, the tools that analyze the trending technologies are so expensive and many companies cannot afford them. Some of the patent analyzing tools are listed below. These patent automation tools are so expensive.

a) Thomson Routers
b) Patent iNSIGHT PRO
c) Patent INTEGRATION
d) PatBASE Analytics

In this study, we proposed a semi-automated patent based tech mining system. The system uses abstracts of the patents to extract features. The unstructured part of the patents such as abstracts contain more prosperous information to predict the technology trends and technology relations. Our system mainly differs from the others systems with the dataset. Until now, there is no such study that uses not only big dataset but also uses unstructured parts of the patents. In this study, we aimed to show that unstructured part of the patents can be converted to the structured ones and also to demonstrate the relations among technologies and analyzing the technology trends. At the end of the study, the results of clustering of extracted data from patent documents exhibit which technology is rising and which are not. The system aims to enclose a big gap in tech mining world. It is an alternative patent automation system via using tech mining proposed against traditional patent automation systems, which require way more costs. Even small companies can implement our system and integrate into their automation.

Cloud computing has become a significant technology trend, and many experts expect it to reshape information - technology processes and the IT marketplace within 5 years. People do not want to use their computer for long computation and storage because it is very memory and processor intensive. Cloud computing lowers the cost of application development and makes the process more scalable. There are some companies have invested in the past for cloud computing such as Amazon, Google, IBM and Dropbox. They lead the innovation world by publishing products and publications. For instance, Amazon has released AWS which is a cloud computing infrastructure, IBM has announced the quantum cloud computing that enables scientists to use quantum computers as a service. As concretely seen that cloud computing is the new age technology and it is an unavoidable field that almost every people use it somehow such as with wearable technologies, cell phones, navigations, games and etc. Therefore, we decided to select cloud computing technology to test the functionality of our system. Besides, cloud computing has great study fields, so it was thought that analyzing this technology and looking for its trends and technological relations can lead other studies.

We started with over sixty thousand patent document which includes cloud computing technology. There are data processing steps were used before preprocessing step. The data should have prepared to preprocess. After some data processing steps, some clustering algorithms used to cluster the data. The clustering step is as important as preprocessing steps although it is the shortest part of the algorithm. The clustering results show what kind of technologies are used with the given technology and which technologies are trends. The visualization and interpretation are another issues for clustering results. Generally, there are some tools to interpret the results though there was not used any tools for this step. It is not actually in this resource’s subject. There are some techniques used to analyze them.

2. LITERATURE REVIEW

Nowadays, it is hard to follow technological innovations and improvements since there is a huge growth in any technological fields. People often improve, create and innovate technologies. Tech mining helps people to follow technologies easily. It is a rising field in the computer science.

Alan and Cunningham (2004) divide tech mining into seven categories.

a) Technical intelligence – monitoring capabilities of a technological field.
b) Competitive intelligence – analyzing the fields to specify the important people and what their interests.
c) Market intelligence – specifying the technology fields that addresses the need of users or sectors in question.
d) Technology road mapping – showing the evaluation of a technology.
e) Technology forecasting – showing technological trends.
f) Technology assessment – helping to see potential technological fields and improvements.
g) Technology foresight – providing information about alternative technologies that can be popular in the future.

Trumbach (2006) points that all employees in the company including managers should be up to date and willing to follow the progress of the related fields. He proposes a model to follow and see new technology fields using tech mining. Tech mining helps companies to follow technology fields that are not found yet. Catching a technology is harder than the inventing it. Tech mining can make a big difference with other companies about the invention of unseen fields.

Trumbach (2006) mentions that small firms tend to make innovations more than large companies though they have small budgets. Tech mining helps small firms to find new innovative ideas as a helpful method. It is a primary area to let companies see or predict the technology evolutions. It also aids companies to see technological gaps. It also plays a key role in technology life cycle (TLC) analyzing. TLC analyzes whether a technology is worth to develop or not. Gao (2013) brought up a method that can evaluate TLC with using patents. He uses the nearest neighbor algorithm to decide if a technology can go further.

The companies also need to follow technological progress to be up to date. Analyzing the market is as important as inventing a product or technology. Before releasing a product or technology, it can be seen that people need that technology or not. A technology can be adapted to other fields. For example, sensor technology is used in medical field to analyze patients' behavior or to be informed urgent situations.

Technological improvements can be tracked using patents. For example, using patent citations, it can be decided that which technology is related to other technologies. Nowadays there are thousands of patents per technological field so it is hard to read and analyze all patents for patent experts. They need an automated system to analyze and interpret the patents with given criterions instead of them. Patent automation system plays a key role in the market. Early years, companies would hire patent experts to follow innovations and trends. Today, they just buy systems to see improvements. The best-known patent automation system in production is OWAKE that is used Japan government (Kakimoto, 2003). It forwards the patents to right human classifiers. It uses a Rocchio-like algorithm with the k-NN approach (Fall et al. 2003).

Patent automation is also important for governments to analyze the technological innovations in the countries. Each country has a patent industry like the United States Patent and Trademark Office (USPTO), Japan Patent Office (JPO), Turkish Patent Office (TPE), WIPO, European Patent Office (EPO). These corporations are responsible for categorizing and managing patents.

There is a number of online patent mining tool like Vantage Point and Patent Insight Pro which are so expensive to use. Rotoloa (2015) proposed a method to analyze emerging technology. All proposed methods create a baseline for the patent automation system. A classic patent analyzing system follows these steps.

a) Task identification: scope of the search
b) Searching: search, filter and download patents
c) Segmentation: pre-processing step before analyzing like using text mining techniques to convert unstructured data to structured one.
d) Abstracting: Analyzing structured data (claim, abstract, description) and summarize them
e) Clustering: Group the patents based on some extracted features.
f) Visualization: Create some meaningful images to monitor clustering results.
g) Interpretation: See technology trends and relations**.**

Lin F. (2008) built a framework to build technology monitoring application. He used second order cosine similarity, hierarchical aggregation clustering, and min-max cut. Larkey (1998) invented a tool for classifying US patent codes based on a k-Nearest Neighbors (k-NN) approach.

Processing patents is a big challenge for data scientists. There are some ways to use patents in tech mining. Since a patent can be taken in several countries and every patent industry gives different patent codes via the standard they rule. It causes diversity in data so International Patent Code provides unity among the diverse patent industries. (Jie et al., 2010) has developed a method to analyze patents using IPCs. For an example, relevant patents of cloud computing are classified under IPC G06 and/or H04 (Huang J. 2016).

A patent has Claims, Abstract, Description, References and many fields. These fields can be processed in machine learning and extracted. Qu P. (2014) did a term extraction using claim abstract and description. He used tf / idf and mutual information methods. Fall C. J. (2003) discovered a new method to classify patents using abstracts and verified the results using IPCs. He used knn, SVM, and bag of words.

In a patent automation system, there are some pre-processing steps which are generally text mining methods before analyzing the data. Text mining is a fundamental area to process unstructured parts of patents. By text mining, unstructured parts of the patents can get some meanings. To do that, generally, feature extraction method is used. There are some feature extraction methods like TF/IDF (Term Frequency-Inverse Document Frequency), the number of words occurring in the title, and a number of numerical data. It is very important to extract key features to process the patent. Another central problem in text mining and information retrieval area is clustering text. After word extraction step, there comes a very high dimensional data. High dimensional and dispersed data reduce~~s~~ the performance of clustering algorithm (LIU L. 2005). Generally, researchers reduce feature dimensions with dropping stop-words and words with high frequency. These words are given in a word list which is made of conjunctions or adverbs. They do not have any effects for clustering process (LIU L. 2005). Therefore, dropping them will increase performance and time. To cluster high dimensional data is a popular topic. A single personal computer fulfills this requirement. For Moore's Law (1975), people have seen the limits of CPUs. So they looked for another way to overcome this limitation. Parallelism comes to the scene with multiple core CPUs. Although there are multi-core CPUs, it is not the solution to the problem of proving or clustering big data. Here comes distributed computing after parallelism. There are also some concepts like distributed file systems which allow processing big data after feature extraction step.

In real world systems, every system has many users simultaneously, and each user requests a patent mining task in patent automation systems. Another concept comes to the area which is called cloud computing. The core of cloud computing comes from distributed computing and grid computing (Li et al.,2015). Chiou (2010) have done a bibliometric analysis about cloud computing life cycle and have seen that development cloud computing is still an emerging technology.

Huang (2016) interpreted cloud computing patents by using the frequency of keywords' occurrence and realized that cloud computing is the feature's technology. He just analyzed the companies which invest the cloud computing technology and the popularity of it by using patent codes and keywords. He also used the IPCs to explore the cloud computing technology and its sub-technologies. The missing part of this study is that it was only analyzed the subfields of the cloud computing technology. It does not concern any other related technologies or the technologies that are related to cloud computing.

The aim of this study is to analyze cloud computing field by using patent contents and explore the related technologies that use cloud computing technology in somehow. For example; at the end of this study, it can be found that robotic or medical field can be related to cloud computing technology. It is hard to see technology relations by just looking some patents. Many studies such as Huang (2016)'s study and Chiou (2010) just use a couple of patents to predict the cloud computing futures though it is not enough to use 100 or 1000 patents to explore it. This kind of study requires a serious number of patents. Here are the objects of this research.

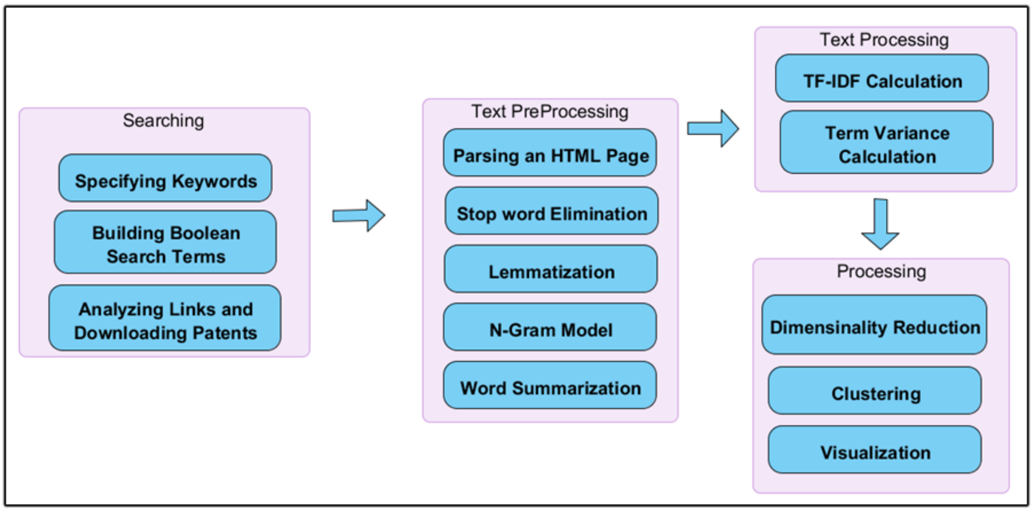
a) To explore technology relations between the cloud computing and other fields.
b) To examine a set of patents which belong different technology fields.
c) To build a technology mining system
d) To compare clustering methods by using the big data set.
e) To identify the problem between patent industry and computer science and build a patent automation system for that.

3. METHODOLOGY

Patents are a great source for analyzing technology related activities of cloud computing field. Many studies can be done via using patents. They do not only contain related technology information but also involve the information that refers another technology fields. So that, Tech mining is used to extract and analyze that information effectively. Before processing a patent, its internal structure should be understood well. Internal structure will give many information and clues about what kind of information can be extracted. Tech mining plays an important role to analyze this information. It does not only help the analyzing process but also aids on all over phases.

The tech-mining method is introduced in 2004 and still applied by various scholars in different fields such as medical, biology, nanotechnology, computer science and etc. It is a subdomain of R and D and also it will give a priceless opportunity to catch technological innovations before opponents in the market. Since it is introduced, it is used in many fields and also there are still many fields that are not examined in tech-mining. Most companies believe that it is a much expensive investment because it is one of R and D area. R and D usually sound expensive for companies though they know that if they do not invest enough in R and D, racing with other companies will be just a desire.

Figure 3.1: Representation of patent mining methodology



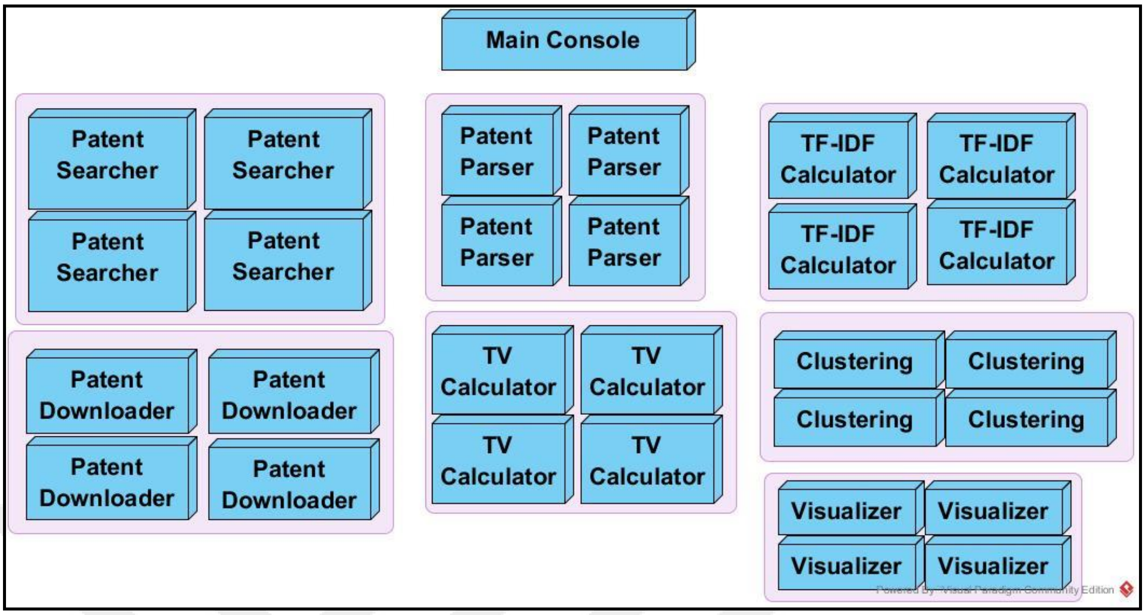
In this study, it has been analyzed the cloud computing subfields and their relations via using tech-mining methods with patents. Before processing patents, there should be specified a methodology. In the classic patent mining systems, there five steps which have already extended. The method that is used in this study includes six steps which are illustrated in figure 3.1.

Before starting all process, the technology field should be selected. Many types of research have done about technological fields such as cloud computing, robotics, nanotechnology and etc. before to choose cloud computing category. In this study, Technology forecasting study has been examined. Nowadays, cloud technologies are so popular and many companies are shifting to the cloud technologies. It is started with some websites such as rapidshare.com and megaupload.com. People started to share files each other on the internet such as a movie, picture, and songs. After that, Dropbox has come to the scene. It is both free and useful. With these technological improvements, vendors have decided to release their products such as Google Drive, Microsoft SkyDrive, and Apple iCloud. Today, people do not prefer to store personal files on their computer or external devices. They can be broken easily so that another solution has appeared as cloud storages and cloud computing area. They have guaranteed the accessibility, failover, and security. The volume of the data is not important anymore. Though, there is no such thing as disc size on the cloud, cloud storages charge the money with the volume of the data. For all these reasons, it has been selected that the cloud storage and hard drives topics. In this study, it can be seen that the relations, differences and technological transition between them. All process was started with collecting the patents.

3.1 MODULES OF THE SYSTEM

The system was built has mainly 8 parts. All system can work parallel and asynchronously. Each of the can run with many instances. In the first version of the program, it was seen that each step was waiting for the previous steps. So, all of the systems was designed again. It has one main console which can deliver the jobs to each module.

Figure 3.2: Modules of the system



The modules of the system can be seen in figure 3.2.

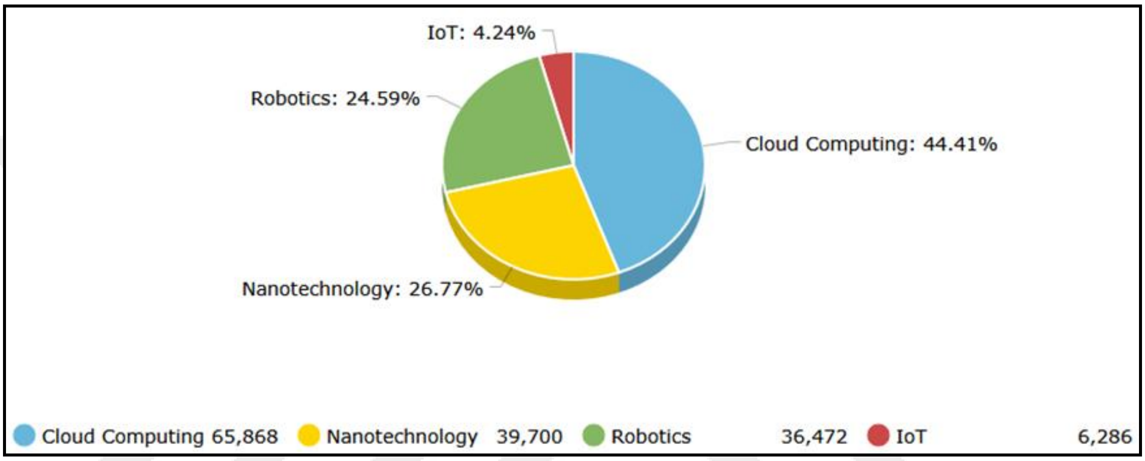
3.2 SEARCHING

First of all, patents or articles from specific and reliable websites were collected. However, many websites do it with specific prices, it has been created a patent search engine to collect all patents with the given keywords. There are many websites were used to collect patents such as Free Patents Online (FPO), WIPO, JPO, and German Patents. All patents can be searched on the websites freely. Although none of the sites provide a web service or an API to query the patents, there has been used URLs to search them effectively. In many countries, collecting data from the internet is not legal and there are some laws to prevent it. It is called data harvesting which means collecting data that requires fee charging. In this study, it has been collected the data does not require any fee-charging although the websites did not provide any sources. The another reason for selecting these websites is that these websites are the official patent industry websites of the countries. They are all reliable and good patent resources. There is also Google Patents site which allows searching patents but it does have any API and does not allow programmatic queries. That is why it was chosen FPO which allows seeing all counties' patents with a query.

3.2.1 Specifying the Technology field

Before starting to build the system, a technology should be selected. So, mainly 4 technology field was selected such as Cloud Computing, IoT, Nanotechnology, and Robotics. To select the technology field, last ten years' patents collected for each field. For patent sources, USPTO and EPO were used.

Figure 3.3: Selecting the technology field



After looking up the number of patents for each technology field. It was seen that most of the patent documents belong to the Cloud Computing field. The results can be seen in figure 3.3. The cloud computing is an important field which has risen in last decade and also it seems it is the future's technology. Many technologies are adapting to the cloud computing such as nanotechnology, big calculations, hosting systems, and biology.

3.2.2 Selecting the Keyword(s)

After selecting the technology field, keywords must be chosen wisely. It will decide the characteristics of the data. It has been also followed a methodology to select the keywords which can be seen below.

3.2.2.1 Internet search

It has been started to look for keywords via Wikipedia which is a good source to collect some information for a given technology in this resource.

Table 3.1: List of keywords related to cloud computing

File hosting services	Cooperative storage cloud
Outsourcing data storage	Block storage
Cloud database	Personal cloud content management service
Object storage	Cloud storage gateway
Cloud collaboration	Cloud desktop storage
Cloud data management ınterface	Mobile cloud storage
Cloud computing	Data-intensive storage services
Scalable data sharing	Cloud storage service

The keywords that have been found on the internet can be found in table 3.1.

3.2.2.2 Searching each keyword on patent sites

After finding the keywords, they have been queried on patent websites as it was mentioned before. This step is important because this step will play an important role to select the keywords. The frequency of a word will specify that how popularly is in patents. This criterion is not only enough to decide whether a given keyword is the word that it has being looked for and it will use the given technology. As the last step, taking a patent expert's opinion is very useful to select them.

3.2.2.3 Asking for advice from patent or technology experts

After finding words and their trends, patent experts can give specific information about the technology or can advise new words which are not on the list it has been proposed. Usually, a single word is not enough to find related patents so that selecting two or more keywords. For example, the word of computing can be used in many fields such as computer science, banking or transportation. Patent experts can advise keywords or mix them to use in searching step. For instance, cloud computing + storage + virtual memory. The final keyword list will be used while building Boolean search terms.

3.2.3 Building Boolean Search Terms

The Boolean search term is a set of expression is created by a query language. A Boolean search term is composed of several queries.

Table 3.2: Example of a boolean search term

TTL/”coffee maker” AND ABST/”heating element”

An example Boolean search term can be seen in table 3.2. This example illustrates a search query that looks for the patents with the word “coffee maker” in Title field and the word “heating element” in Abstract field.

Table 3.3: Parts of patents and their corresponding boolean search terms

Sections of Patents	**Search Terms**
Claim	ACLM
Abstract	ABST
Description	SPEC
Filing Date	APD

There are many shortenings are used in search terms. Some of them listed in table 3.3.A search term is made of a set of keywords and 3 Boolean operators which are AND, OR and NOT. Any field can be queried by using them.

Two search terms were used to collect two different patents to compare them.

Table 3.4: Search term of ‘Cloud Computing’ keyword

(((ABST/"cloud computing" OR ACLM/" cloud computing" OR SPEC/" cloud computing”) AND APD/1/1/2006->12/31/2016

First Boolean search term has been used, is shown, table 3.4. In this Boolean search term, only one word has been used, which is “cloud computing”. The word that it has been looked for must be located in Abstract, Claim, and Description. Each word must appear only once one of the fields abstract, Claim or Description, also patents that were queried must be between 2006 and 2016 years. Only last ten years’ patents have been taken.

3.2.4 Analyzing Links and Downloading Patents

Searching in a website is easy for a human but not for a machine. To make searches automatic it has been needed to study the structure of a link.

Table 3.5: The resulted link of ‘cloud computing’ boolean search term

http://www.freepatentsonline.com/result.html?**p=1**&edit_alert=&srch=xprtsrch&query_txt=%28%28ABST%2F%22cloud+computing%22+OR+ACLM%2F%22cloud+computing%22+OR+SPEC%2F%22cloud+computing%22%29%29%0D%0A+AND+APD%2F1%2F1%2F2006-%3E12%2F31%2F2016&uspat=on&usapp=on&eupat=on&jp=on&pct=on&date_range=all&stemming=on&sort=relevance&search=Search

The example link that was analyzed is shown in table 3.5.

It had to be solved how it was working. After a while, it was realized that after searching, pages changing with a sequence which correspond to **“p=1”.** The number is changing with every page. This is how it has been traversed all patent pages.

A web page is made of HTML codes. If a machine looks at a web page, the only thing it sees codes like HTML, CSS, and JavaScript. After finding all patents that were looked for, they were downloaded all HTML pages in parallel to extract the patent information. An example patent link is shown in table 3.6.

Table 3.6: An example of a patent link

http://www.freepatentsonline.com/y2013/0097275.html

Once the patents are downloaded, they are ready to text pre-processing step.

3.3 TEXT PRE-PROCESSING

The plain HTML document does not make sense for a human without processing it. FreePatentsOnline.com gives many fields of a patent.

Table 3.7: Patent fields

Title	Assignee
Abstract	Primary Class
Inventors	International Classes
Application Number	Related US Applications
Publication Date	Claims
Filing Date	Description
Export Citation	View Patent Images

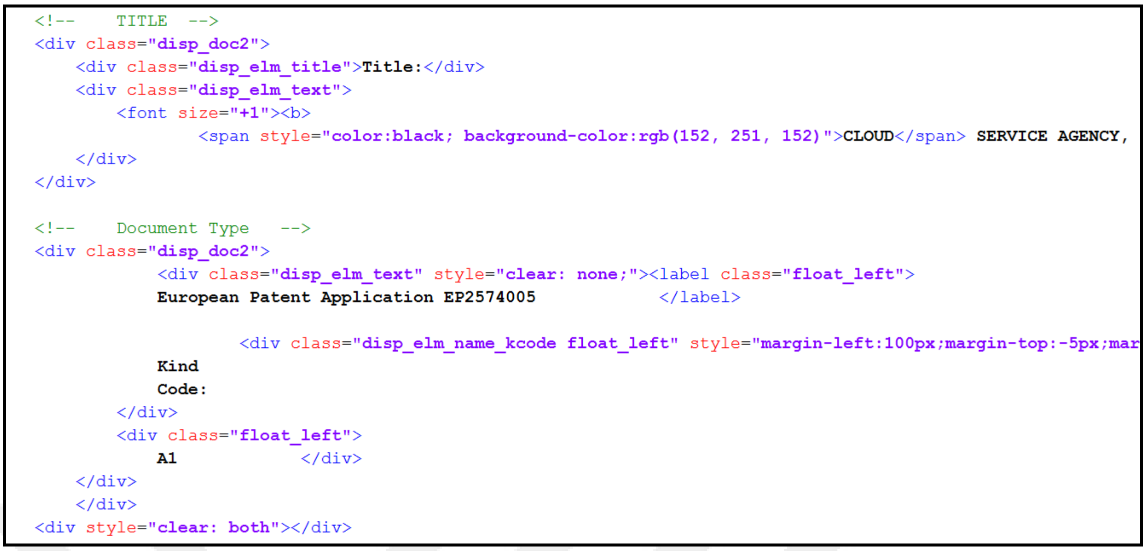
Fields are shown in table 3.7.

3.3.1 Parsing an HTML Page

Processing an HTML page is as hard as understanding its own structure because a web page is an unstructured data source, so it had to be converted a structured one. There is a library which is JSoup to process an HTML page. It gives all nodes with a meaningful sequence, so we can extract the parts of a patent easily.

In this study, there are only two fields used such as Abstract and Filling Date. The Abstract field is a plain text that contains the summary of a patent. Also, the another field was used, is Filling Date which is to make some filtering. For example; it can be seen that the patents which are released in last 5 years.

Figure 3.4: A sample HTML page



```
<!--    TITLE  -->
<div class="disp_doc2">
    <div class="disp_elm_title">Title:</div>
    <div class="disp_elm_text">
        <font size="+1"><b>
                <span style="color:black; background-color:rgb(152, 251, 152)">CLOUD</span> SERVICE AGENCY,
    </div>
</div>

<!--    Document Type   -->
<div class="disp_doc2">
            <div class="disp_elm_text" style="clear: none;"><label class="float_left">
            European Patent Application EP2574005            </label>

                    <div class="disp_elm_name_kcode float_left" style="margin-left:100px;margin-top:-5px;mar
            Kind
            Code:
        </div>
        <div class="float_left">
            A1            </div>
    </div>
    </div>
<div style="clear: both"></div>
```

In figure 3.4, a sample HTML page can be seen.

3.3.2 Stop Word Elimination

Stop words are the words that refer common words in a language. These words usually removed from texts before or after processing them in natural language processing. It is a very common technique that is done almost every study. Although they appear much in a text, they do not give any extra meaning.

Table 3.8: An example of stop words list

a's	able	about	Above	according
accordingly	across	actually	after	afterwards
again	against	ain't	all	allow
allows	almost	alone	along	already
also	although	always	am	among
amongst	an	and	another	any
anybody	anyhow	anyone	anything	anyway
anyways	anywhere	apart	appear	appreciate

Here are the some of the stop words that are used is shown in table 3.8. All stop words were just removed from the Abstract field by using this stop words list. Elimination of stop words will give the content it was expected.

In this study, stop words were selected with using the patents. To select the stop words, patents which belong to Nanotechnology, IoT, Cloud Computing, and Robotics were used. Term frequency was calculated for all patents and have found the all mutual words for selected technological fields. These mutual words composed the stop words list.

The stop words elimination step is the difficult part of the pre-processing step. It requires much attentions because it specifies the quality of the data. If this step is not done with good quality, the words will be used, becomes less qualified. For example; suppose that there is a sentence like “The cloud computing is the future’s technology.”. If this step is not applied, the result becomes “the cloud”, “cloud computing”, “computing is”, “the future”, “future technology” after N-Gram modeling step. If it is applied, the result of the data becomes “cloud computing”, “computing future”, “future technology”. So, there would be useless words without this step.

3.3.3 Lemmatization

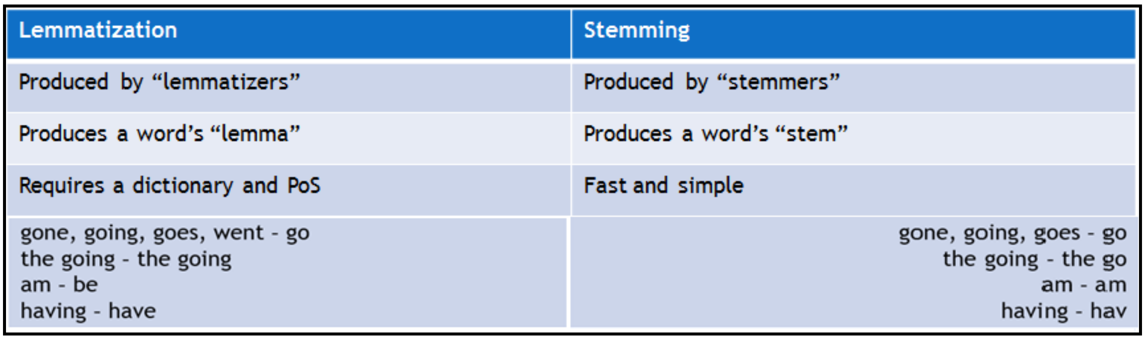
After stopping words elimination process, there are the pure data but there is one more problem. There are plural words, that are also verbs in any forms. Also, they should be eliminated, although all of them looks different. This process is called lemmatization that gives the base form of a verb or gives the single form of a word. Before applying lemmatization, it was compared to stemming algorithms. Stemming algorithms only give the single form of the words.

Table 3.9: An example of lemmatization

Before	After
Children	Child
Came	Come
Father’s shoes	Father shoe
A single patent contains much information	A single patent contain much information

They do not reduce the verbs to the base form. As it can be seen in table 3.9, although they look different, they are the same for this study. They had to be lemmatized because the distinct words will be used in processing step. In this step, Stanford NLP library was used. It gives many options such as detecting the words, reducing to the base form and removing them. Before applying lemmatization step, it was also analyzed the stemming algorithms. There are many stemming algorithms that allow the clean unknown forms of the words. Jivani (2011) also compared the stemming algorithms and suggests Porter’s Stemmer though it is time-consuming. Although stemming process works well, it does not give what is needed in this research. Here is an example of comparison stemming and lemmatization. For example, the words gone goes map to the “go”. The word “went” will not map to the same word. However, lemmatization will map “went” to “go”.

Figure 3.5: Lemmatization and stemming



Lemmatization	Stemming
Produced by “lemmatizers”	Produced by “stemmers”
Produces a word’s “lemma”	Produces a word’s “stem”
Requires a dictionary and PoS	Fast and simple
gone, going, goes, went - go the going - the going am - be having - have	gone, going, goes - go the going - the go am - am having - hav

The comparison of stemming and lemmatization can be found in figure 3.5.

3.3.4 N-Gram Model

An n-gram is a contiguous sequence of a sentence. After stop words, elimination and lemmatization, the content of the abstract field become less, so n-gram modeling step helps to enrich the data. N-gram modeling which is a general method is used in Natural Language Processing (NLP), is an essential step in our study. The other application can be seen below.

a) Probability
b) Natural Language Processing
c) Computational Biology
d) Data Compression

N-Gram modeling doubles the content so the data becomes so diverse. In the study, Unigram and Bigram modeling are generated from Abstract content.

Table 3.10: An example of n-gram modeling

Sentence	UniGram	BiGrams
Patent documents are so important	Patent	Patent documents
	Documents	Documents are
	Are	Are so
	So	So important
	important	

An example of n-gram modeling is shown in table 3.10. As can be seen in table 3.10, when there are five words, after n-gram modeling there become nine words.

3.3.5 Word Summarization

Word summarization is the last step of the pre-processing. After n-gram modeling, there become 3.614.602 n-gram words for “cloud computing” patents. It is hard to process all of the words because there are many repetitive words even if some elimination processes have been done. In this step, all of the words have been summarized to a smaller data set, because it is easy to process smaller data set. The new data sets size are 486.735 for “cloud storage”. These words will represent the data sets so, the study will continue with these data sets.

3.4 TEXT PROCESSING

Text processing step is one of the important processes in this study. In cluster process, the dimension of a patent is very high. If there are 216.452 words, that means there are 216.452 features for each patent so, a high number of features makes clustering process longer. So, text processing step will score the each feature of patents. In text processing step, two methods were used such as TF-IDF and Term Variance. TF-IDF (Term Frequency - Inverse Document Frequency) is not strong by itself in a sparse data set so Term Variance method has supported by rescoring.

3.4.1 TF-IDF Calculation

Term frequency is a highly used process in text retrieval, is a statistical method to find the count of the each word in a given text.

Figure 3.6: The formula of TF

$$\boxed{\mathrm{tf}(t,d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}}$$

The formula is shown in figure 3.6.

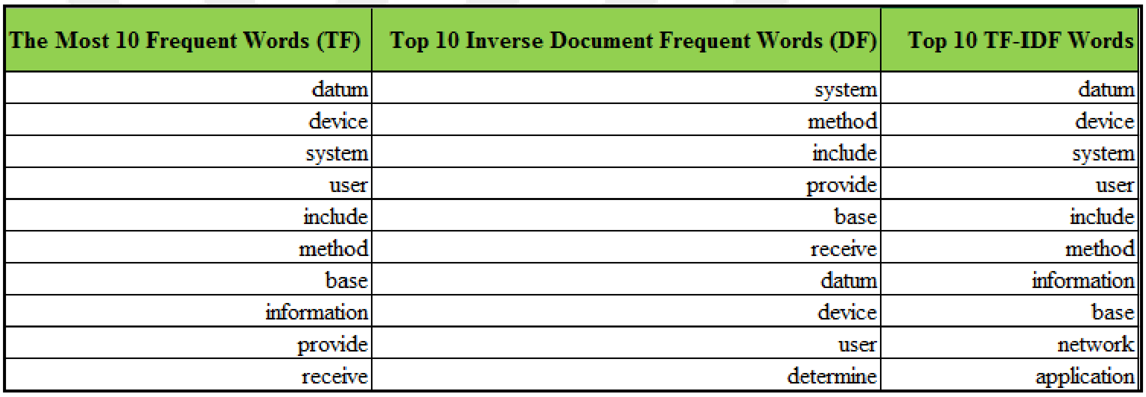
TF just gives the frequency of a word in a text. Generally, TF method is not enough to give meaningful results to score the words. There is also another problem in TF which is , in some cases the words that are not important seem important such as “the”, “are”, “this”. Although they have higher occurrences, these words should have the lower scores, so Inverse Document Frequency (IDF) method is used to eliminate unnecessary words. This method measures how important a word is. While TF computation, each word is considered equal, though they are not. IDF is used for weighting down unnecessary words.

Figure 3.7: The formula of IDF

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

TF-IDF method is applied for both “cloud computing” data sets. The formula is shown in figure 3.7. After calculation TF-IDF the results are not promising.

Figure 3.8: The results of TF, DF, and TF-IDF



The Most 10 Frequent Words (TF)	Top 10 Inverse Document Frequent Words (DF)	Top 10 TF-IDF Words
datum	system	datum
device	method	device
system	include	system
user	provide	user
include	base	include
method	receive	method
base	datum	information
information	device	base
provide	user	network
receive	determine	application

The TF and the TF-IDF results are almost the same. It can be seen in figure 3.6. TF-IDF is not successful in this study because of the sparse data. After text elimination step, the data became so sparse. TF-IDF could not eliminate the words like datum and device.

3.4.2 Term Variance Calculation

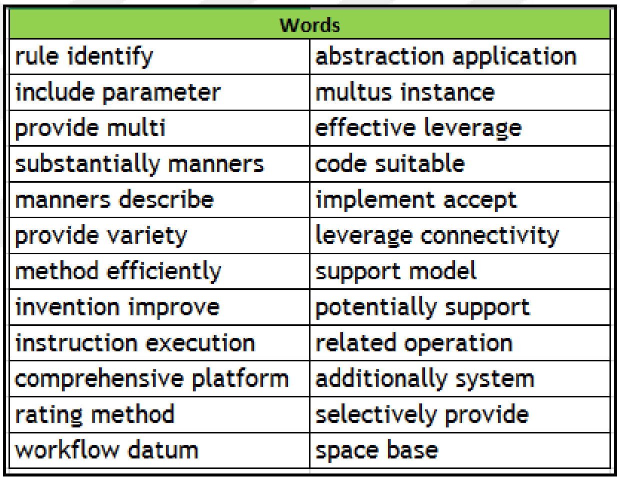
TF-IDF assume that each term has equal priority in each document, so it can be easily manipulated by common words that have high document frequency but equally distributed. Therefore, LIU (2015) proposed a new method which is called Term Variance (TV). It has the same idea of Document Frequency that the words with lower document frequency is not important, can figure out the problem that solves TF-IDF. TV method requires the TF-IDF values to give the final score each term.

Figure 3.9: Formula of TV

$$v(t_i) = \sum_{j=1}^{N} \left[f_{ij} - \overline{f_i} \right]^2$$

Figure 3.9 shows the formula of TV which is derived from TF formula shown in figure 3.6.

Figure 3.10: The results of TV



Words	
rule identify	abstraction application
include parameter	multus instance
provide multi	effective leverage
substantially manners	code suitable
manners describe	implement accept
provide variety	leverage connectivity
method efficiently	support model
invention improve	potentially support
instruction execution	related operation
comprehensive platform	additionally system
rating method	selectively provide
workflow datum	space base

The results of the TV can be seen in figure 3.10. In almost every patent, the words like datum, system, device occur a lot. So they corrupt the term extraction results. TF-IDF could not eliminate them because of the sparse data matrix. After applying TV, it can be seen that the results are interesting and better than the TF-IDF.

3.5 PROCESSING

3.5.1 Dimensionality Reduction

The dataset has the high dimensionality, so the clustering step takes too much time. Dimensionality reduction methods convert a dataset into two pieces, so the dimension of the data decreases. After dimensionality reduction step, the dataset can be recovered again calculating the new two datasets.

Figure 3.11: The formula of PCA

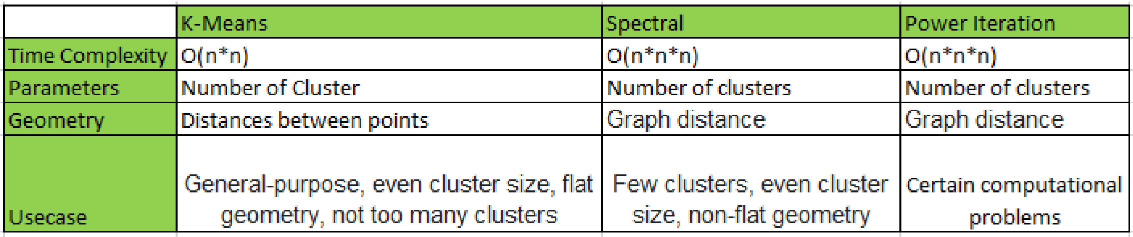
$$\boxed{\mathbf{w}_{(1)} = \underset{\Vert\mathbf{w}\Vert=1}{\arg\max} \left\{ \sum_i (t_1)^2_{(i)} \right\} = \underset{\Vert\mathbf{w}\Vert=1}{\arg\max} \left\{ \sum_i \left(\mathbf{x}_{(i)} \cdot \mathbf{w}\right)^2 \right\}}$$

For the dimensionality reduction step, Principal Component Analysis (PCA) that is one of the dimensionality reduction methods is used.The formula of PCA can be found in figure 3.11. In this study, %80 and %90 variances applied to the main dataset. Some of the clustering methods are not working with big datasets because of the exiguous RAM and CPU.

3.5.2 Clustering

Clustering has the same importance as the other steps. It also determines the study’s fate,so it should be done carefully. This step will show the technologies that are related to the cloud computing and also what kind of technologies use the cloud computing. Clustering methods help words that look similar to get in the same category. K-Means, Spectral, and Power Iteration clustering methods have been tried in this study.

Figure 3.12: Comparison of clustering algorithms



	K-Means	Spectral	Power Iteration
Time Complexity	O(n*n)	O(n*n*n)	O(n*n*n)
Parameters	Number of Cluster	Number of clusters	Number of clusters
Geometry	Distances between points	Graph distance	Graph distance
Usecase	General-purpose, even cluster size, flat geometry, not too many clusters	Few clusters, even cluster size, non-flat geometry	Certain computational problems

The comparison of clustering algorithms can be seen in figure 3.12. As it can be seen that K-Means is faster than the others. The convergence time of Power Iteration is so slow although it detects the clusters better than the Spectral. The Spectral can give better results to detect the nongeometric clusters.

3.5.2.1 K - means

K - means is one of the most popular cluster methods. It can be applied to many situations and datasets. It aims to split n observations into k clusters that each observation belongs to the cluster with the nearest mean.

Figure 3.13: The formula of K - means

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2$$

The formula of K-means can be found in figure 3.13. There are some methods used for the similarity calculations that are called distance metrics. In this study, two distance metrics have been used.

3.5.2.1.1 *Euclidian distance*

Euclidian distance calculates the distance between two points in Euclidian space.

Figure 3.14: The formula of euclidian distance

$$\begin{aligned} \mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) &= \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \\ &= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}. \end{aligned}$$

Figure 3.14 shows the formula of Euclidian distance.

3.5.2.1.2 *Cosine distance*

Cosine distance calculates the distance between two non-zero vectors that calculates the cosine angle between them.

Figure 3.15: The formula of cosine distance

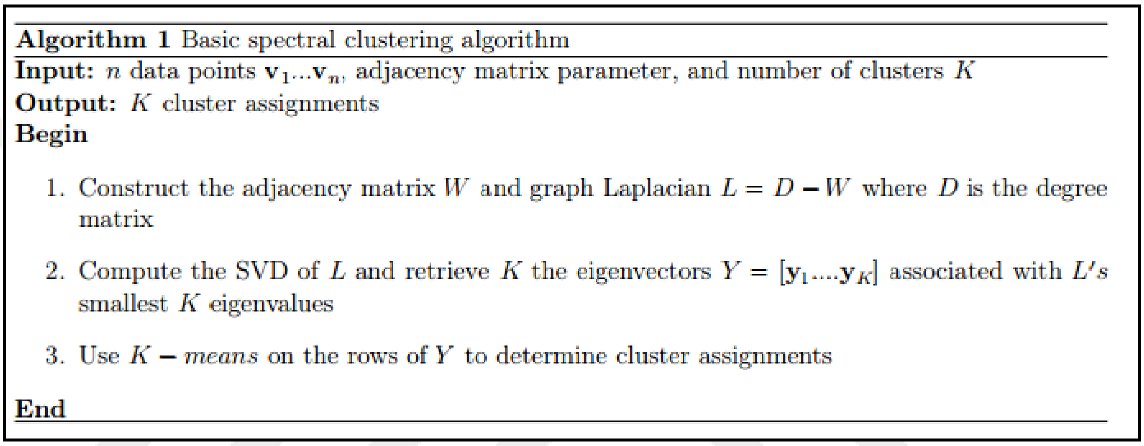
$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2} \sqrt{\sum\limits_{i=1}^{n} B_i^2}}$$

The formula of cosine distance can be seen in figure 3.15.

3.5.2.2 Spectral

Spectral is one of the methods is used. Spectral makes use of eigenvalues of similarity matrix of the data to reduce the dimension before clustering. This similarity matrix is given as an input and consists of a quantitive assessment of the relative similarity of each pair of points in the dataset.

Figure 3.16: The algorithm of spectral clustering



```
Algorithm 1 Basic spectral clustering algorithm
Input: $n$ data points $\mathbf{v}_1 ... \mathbf{v}_n$, adjacency matrix parameter, and number of clusters $K$
Output: $K$ cluster assignments
Begin
    1. Construct the adjacency matrix $W$ and graph Laplacian $L = D - W$ where $D$ is the degree matrix
    2. Compute the SVD of $L$ and retrieve $K$ the eigenvectors $Y = [\mathbf{y}_1 .... \mathbf{y}_K]$ associated with $L's$ smallest $K$ eigenvalues
    3. Use $K - means$ on the rows of $Y$ to determine cluster assignments
End
```

The algorithm of spectral clustering can be found in the figure 3.12.

3.5.2.3 Power Iteration

Power iteration is a clustering method that produces the eigenvalue number and a non-zero vector. It does not compute a matrix decomposition. It is used when there is a huge sparse matrix. It can find one eigenvalue number though it converges slowly.

Figure 3.17: The algorithm of power iteration

$$b_{k+1} = \frac{Ab_k}{\|Ab_k\|}.$$

$$b_k = e^{i\phi_k} v_1 + r_k.$$

$$\mu_k = \frac{b_k^* A b_k}{b_k^* b_k}$$

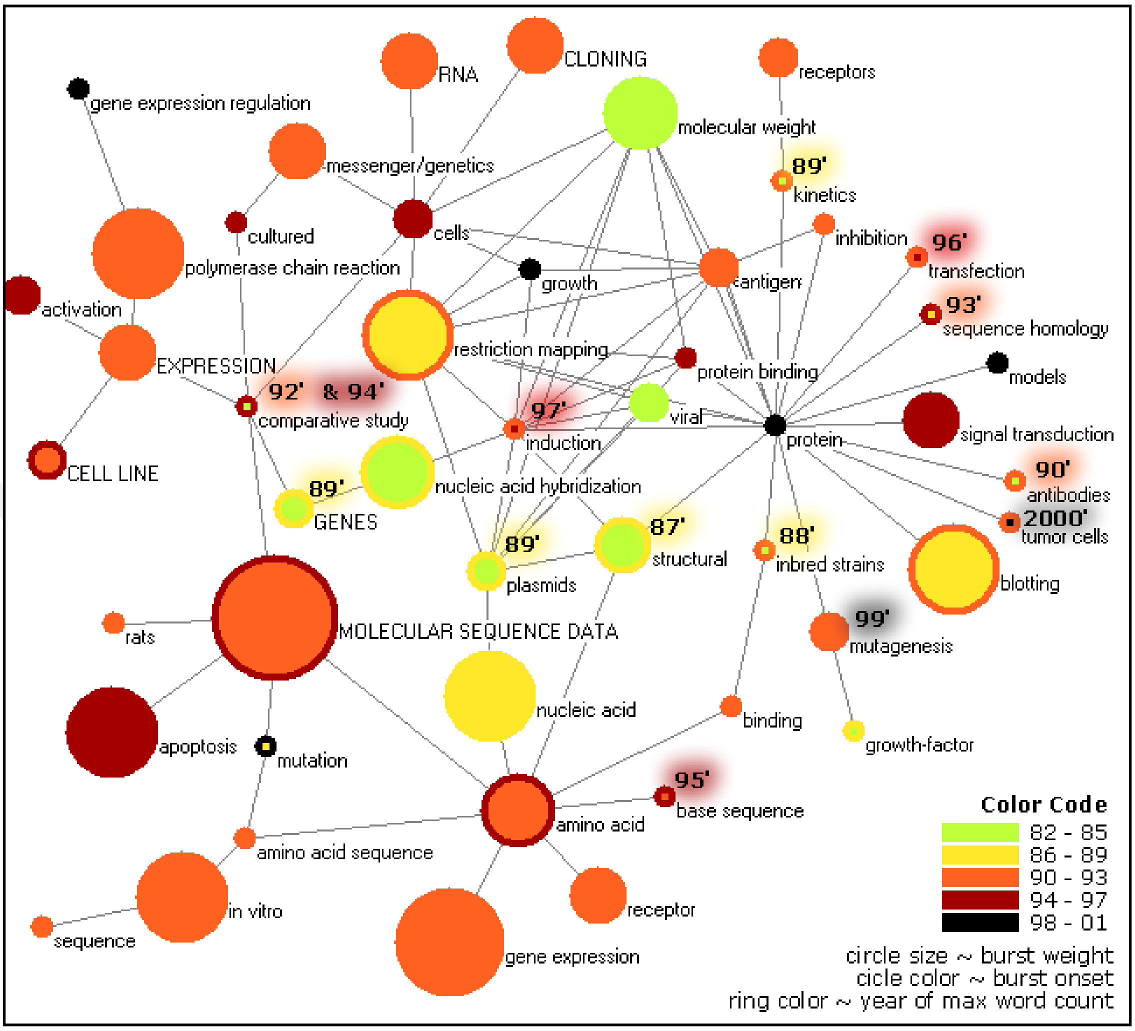
The formula of power iteration can be found in the figure 3.17.

3.5.3 Visualization

There are many visualization tools in the market such as Pajek, VosViewer, Gephi and etc.The VosViewer is such a successful tool and it is free. Firstly, it clusters the data and gives some visualization options. It does have any features like visualizing the results without clustering. The another powerful tool is Gephi. It is also free and useful although it does not have many visualization options. Pajek is widely used in bibliometrics and can handle large datasets. It is a matter of personal choice but tools such as Gephi may be superseding Pajek because they are more flexible. However, Pajek is limited with an edge in precision, ease of reproducibility and ability to easily save the study that Gephi does not have any talent as a Beta version.

For this step, Pajek visualization tool was used. Pajek is one of the most popular visualization tools. It’s all free and easy to use. This is why it was used in this research. In this study, network visualization was used.

Figure 3.18: An example of PAJEK visualization



Source: Maps and Science

An example of network visualization can be found in the figure 3.18.

4. RESULTS

In this study, the dataset consists of two different patents. The first one belongs to last two years' patents and the other one contains last three years' patents. There are several clustering methods have been tried in this study. The results are extremely exciting and efficient. The first clustering method is studied is K-Means. For the K-Means, two similarity metrics have been used such as Euclidian Distance and Cosine Distance.

Table 4.1: Characteristics of all patents

Attribute	Count
Total Patent Count	65868
Word Count	1000
Total Data Size	65868000

The characteristics of the dataset are listed in table 4.1.

Table 4.2: Characteristics of the patents covering last three years

Attribute	Count
Total Patent Count	36076
Word Count	1000
Total Data Size	36076000

Table 4.2 shows the last three years patents

Table 4.3: Characteristics of the patents covering last two years

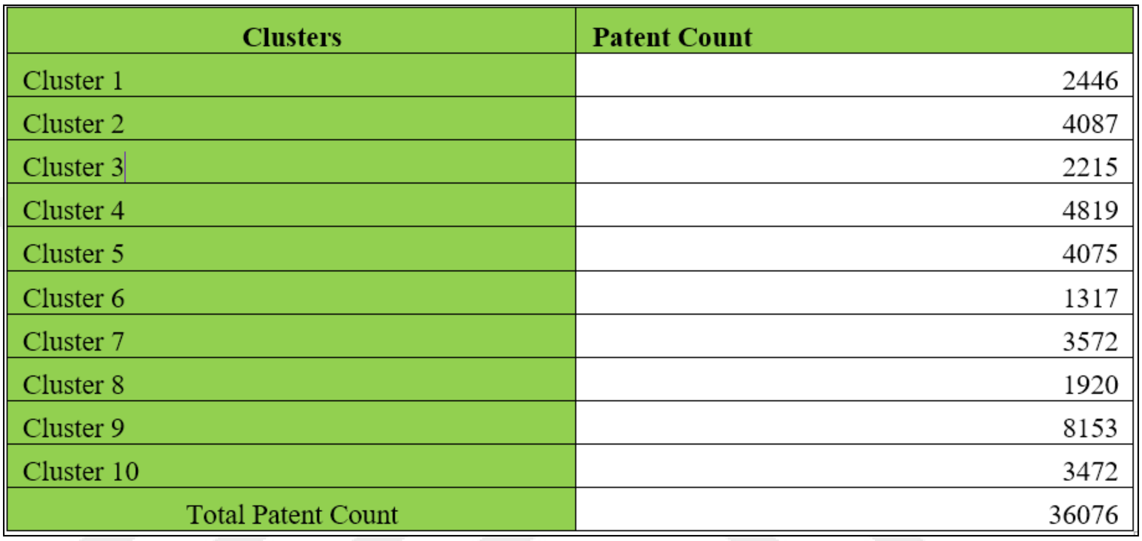
Attribute	Count
Total Patent Count	19989
Word Count	1000
Total Data Size	19989000

Table 4.3 shows the last two years patents.

4.1 K-MEANS COSINE DISTANCE 1000 WORDS 10 CLUSTERS (95 PERCENT VARIANCE)

In this experiment, a thousand words used and ten clusters were generated. The number of principal components used are determined so that 95% of the variance is preserved.

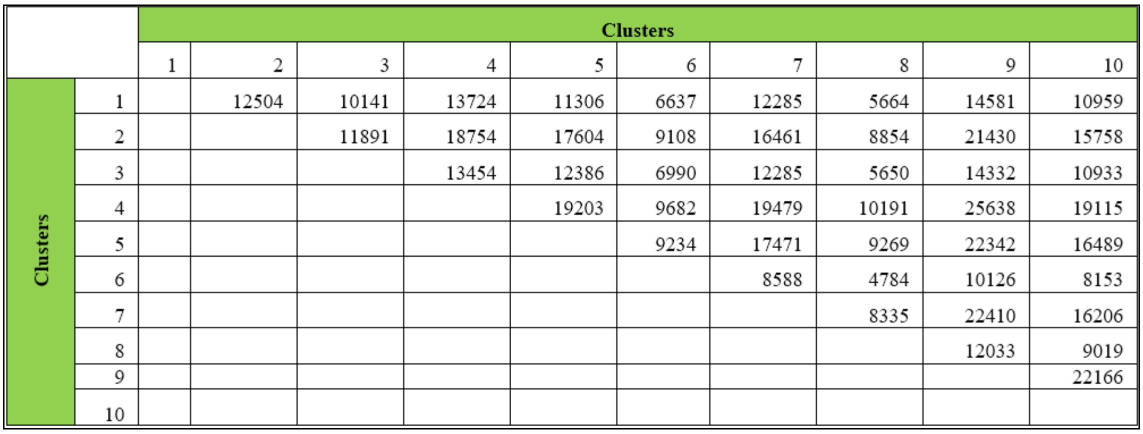
Figure 4.1: Distribution of patents according to defined clustering algorithm



Clusters	Patent Count
Cluster 1	2446
Cluster 2	4087
Cluster 3	2215
Cluster 4	4819
Cluster 5	4075
Cluster 6	1317
Cluster 7	3572
Cluster 8	1920
Cluster 9	8153
Cluster 10	3472
Total Patent Count	36076

The patent cluster distribution obtained after applying the k-means clustering algorithm with cosine distance can be seen in figure 4.1 .As it is shown, patents are distributed well to each cluster.

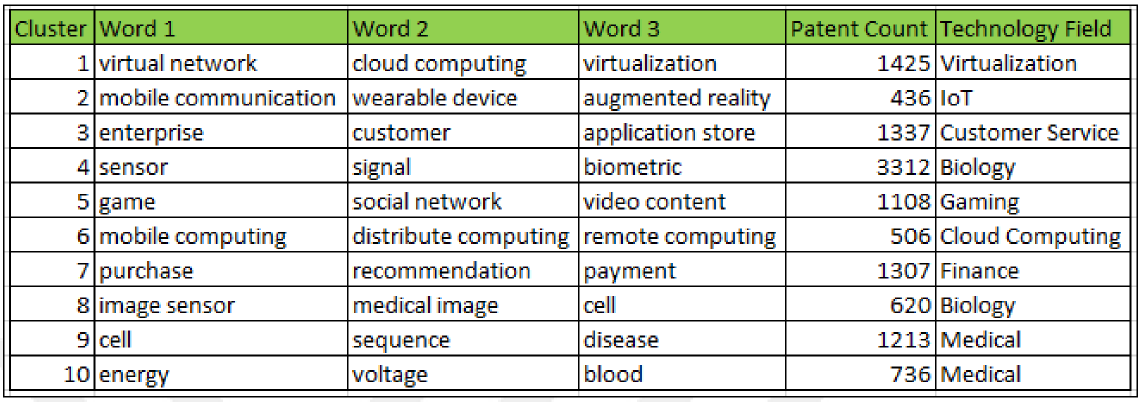
Figure 4.2: Cluster-relation table according to defined algorithm



		Clusters									
		1	2	3	4	5	6	7	8	9	10
Clusters	1		12504	10141	13724	11306	6637	12285	5664	14581	10959
	2			11891	18754	17604	9108	16461	8854	21430	15758
	3				13454	12386	6990	12285	5650	14332	10933
	4					19203	9682	19479	10191	25638	19115
	5						9234	17471	9269	22342	16489
	6							8588	4784	10126	8153
	7								8335	22410	16206
	8									12033	9019
	9										22166
	10										

Figure 4.2 shows the relation among the clusters. Each number represents the number of words that appear in the patents fall into the corresponding pair of clusters.

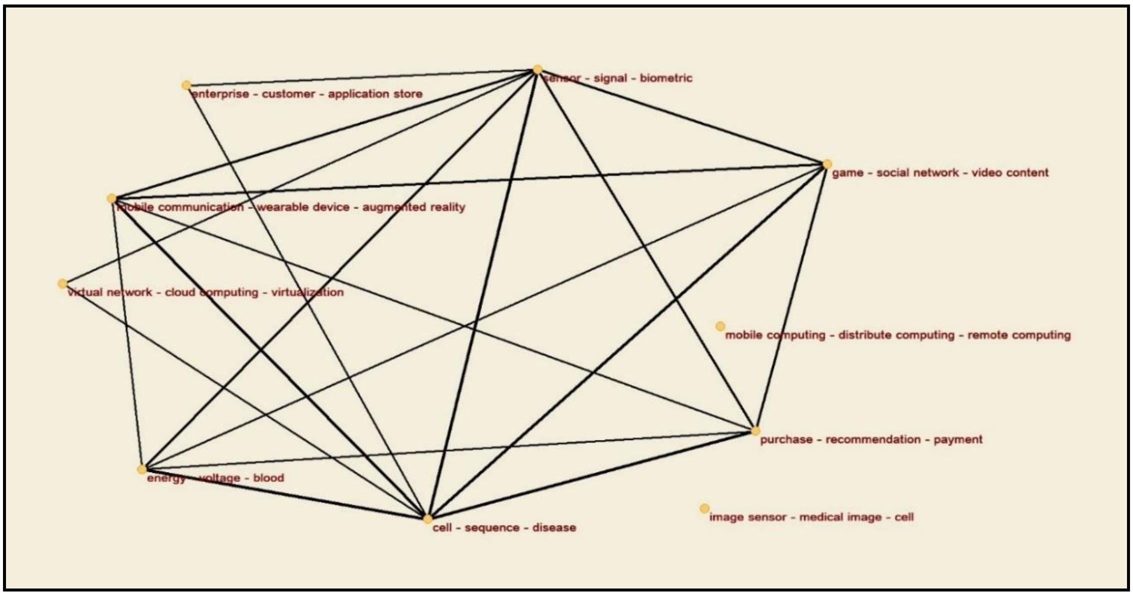
Figure 4.3: Top 3 words of the result of defined algorithm



Cluster	Word 1	Word 2	Word 3	Patent Count	Technology Field
1	virtual network	cloud computing	virtualization	1425	Virtualization
2	mobile communication	wearable device	augmented reality	436	IoT
3	enterprise	customer	application store	1337	Customer Service
4	sensor	signal	biometric	3312	Biology
5	game	social network	video content	1108	Gaming
6	mobile computing	distribute computing	remote computing	506	Cloud Computing
7	purchase	recommendation	payment	1307	Finance
8	image sensor	medical image	cell	620	Biology
9	cell	sequence	disease	1213	Medical
10	energy	voltage	blood	736	Medical

The most frequent 3 words of each cluster are listed in figure 4.3. By looking at the clusters, technological fields that the cloud computing applied or used were aimed to be identified. The results show that there are 8 main technology fields which are Virtualization, IoT, Customer Service, Biology, Gaming, Cloud Computing, Finance, and Medical. The number of patents in which all of the 3 most frequent words appear together for each cluster is also shown in figure 4.3. It is seen that the most of the patent documents belong to the Biology fields.

Figure 4.4: Visualization of relations among clusters


enterprise - customer - application store
sensor - signal - biometric
game - social network - video content
mobile communication - wearable device - augmented reality
virtual network - cloud computing - virtualization
mobile computing - distribute computing - remote computing
purchase - recommendation - payment
energy - voltage - blood
image sensor - medical image - cell
cell - sequence - disease

Clustering results were also visualized in figure 4.4. Each point represents a cluster. Each line shows the relation among clusters. The strength of a relation among clusters can be seen by the thickness of a line. This diagram was produced by using the cluster relations. It can be seen that the IoT and the Biology fields have a strong relation. Also, IoT field is related to Gaming field. To verify these results, some of the patents documents are listed below, which are related to cloud computing field. These documents were found by using the game and augmented reality keywords together as a boolean search term.

a) Method and system for playing an augmented reality in a motor vehicle display
b) Head wearable electronic device for augmented reality and method for generating augmented reality using the same
c) Remote control game system based on augmented reality

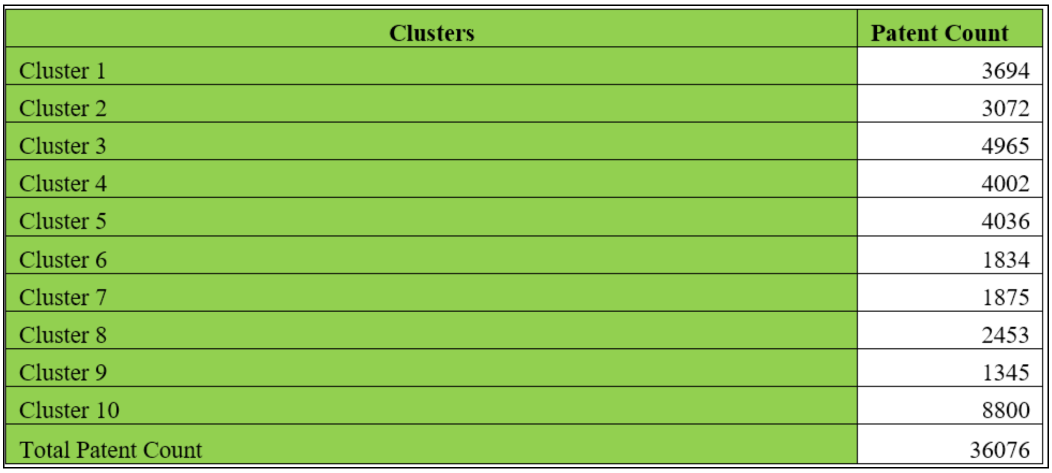
Also, there is an interesting relation between the gaming and the biology fields. The patents which verify the relation can be found below.

a) Game of the immune system
b) Biologically fit wearable electronics apparatus and methods
c) Educational game

4.2 K-MEANS COSINE DISTANCE 1000 WORDS 10 CLUSTERS (80 PERCENT VARIANCE)

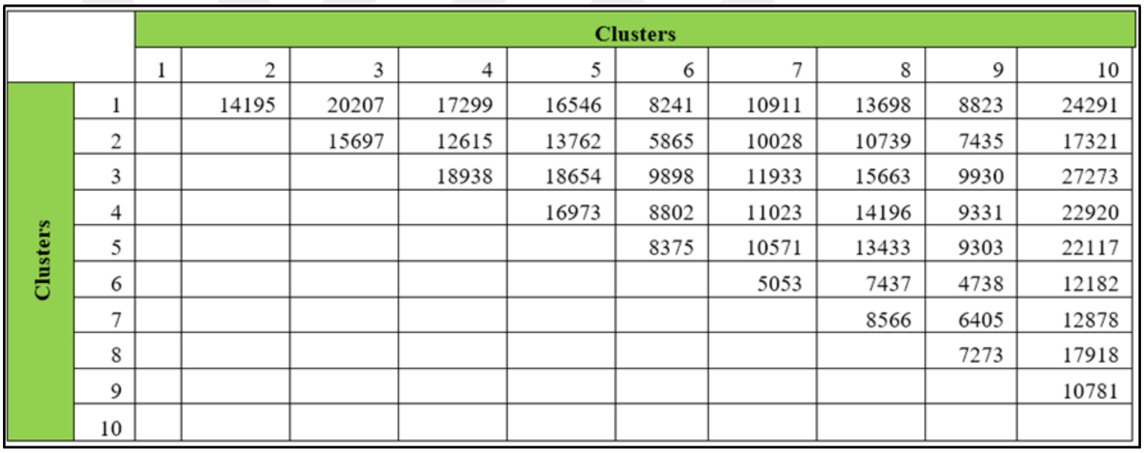
In this experiment, 1000 words used and 10 clusters were generated by k-means clustering algorithm with cosine distance. PCA was applied by preserving 80% of the variance. The patent cluster distribution can be seen in figure 4.5. These results show that patents distributed well. Figure 4.6 shows us the relation among each cluster. If the relations of patents are weak, the visualization results become irrelevant.

Figure 4.5: Distribution of patents according to defined clustering algorithm



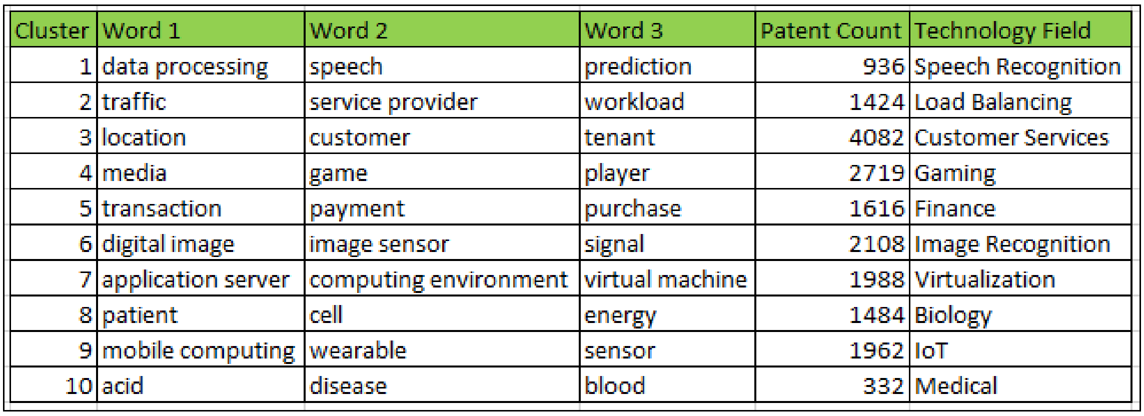
Clusters	Patent Count
Cluster 1	3694
Cluster 2	3072
Cluster 3	4965
Cluster 4	4002
Cluster 5	4036
Cluster 6	1834
Cluster 7	1875
Cluster 8	2453
Cluster 9	1345
Cluster 10	8800
Total Patent Count	36076

Figure 4.6: Cluster-relation table according to defined algorithm



		Clusters									
		1	2	3	4	5	6	7	8	9	10
Clusters	1		14195	20207	17299	16546	8241	10911	13698	8823	24291
	2			15697	12615	13762	5865	10028	10739	7435	17321
	3				18938	18654	9898	11933	15663	9930	27273
	4					16973	8802	11023	14196	9331	22920
	5						8375	10571	13433	9303	22117
	6							5053	7437	4738	12182
	7								8566	6405	12878
	8									7273	17918
	9										10781
	10										

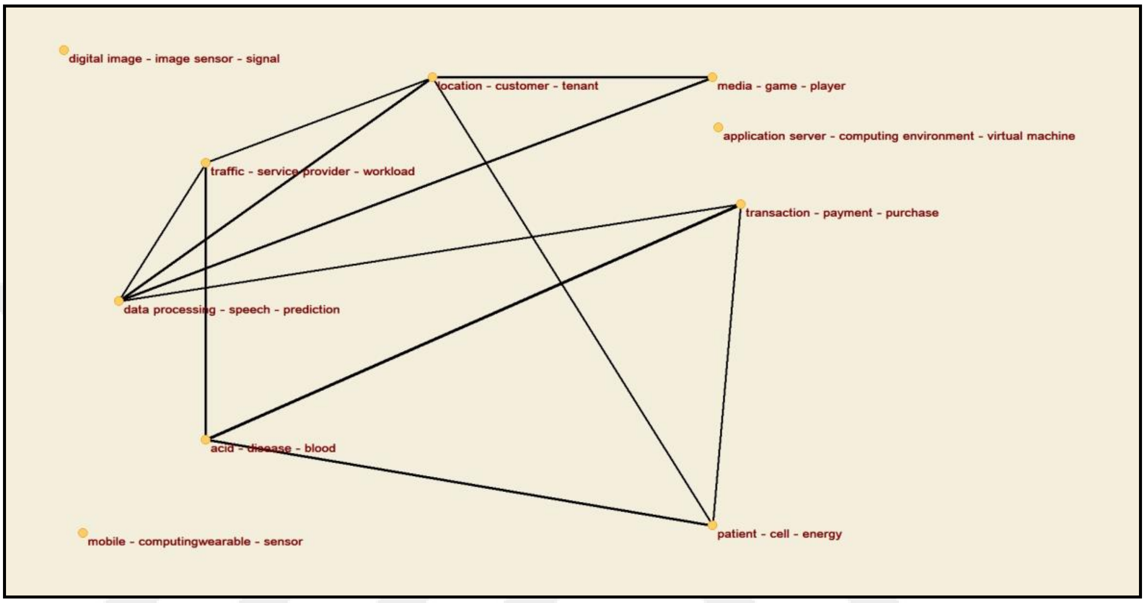
Figure 4.7: Top 3 words of the result of defined algorithm



Cluster	Word 1	Word 2	Word 3	Patent Count	Technology Field
1	data processing	speech	prediction	936	Speech Recognition
2	traffic	service provider	workload	1424	Load Balancing
3	location	customer	tenant	4082	Customer Services
4	media	game	player	2719	Gaming
5	transaction	payment	purchase	1616	Finance
6	digital image	image sensor	signal	2108	Image Recognition
7	application server	computing environment	virtual machine	1988	Virtualization
8	patient	cell	energy	1484	Biology
9	mobile computing	wearable	sensor	1962	IoT
10	acid	disease	blood	332	Medical

As it can be seen in figure 4.7, there is 10 technology fields have found. The 3rd cluster includes the highes number of patent documents which is Customer Services field. Similar technology fields found in figure 4.3 are also seen in figure 4.7.

Figure 4.8: Visualization of relations among clusters


digital image - image sensor - signal
location - customer - tenant
media - game - player
application server - computing environment - virtual machine
traffic - service provider - workload
transaction - payment - purchase
data processing - speech - prediction
acid - disease - blood
mobile - computingwearable - sensor
patient - cell - energy

As it can be seen in this diagram, biology and finance field have a strong relation. Also, speech recognition and gaming fields have closeness. To verify the relation of speech recognition and gaming fields, the found patent documents can be seen below.

a) Menu-driven voice control of characters in a game environment
b) Interactive game playing preferences
c) Subscriber system and method for lotto and lottery games

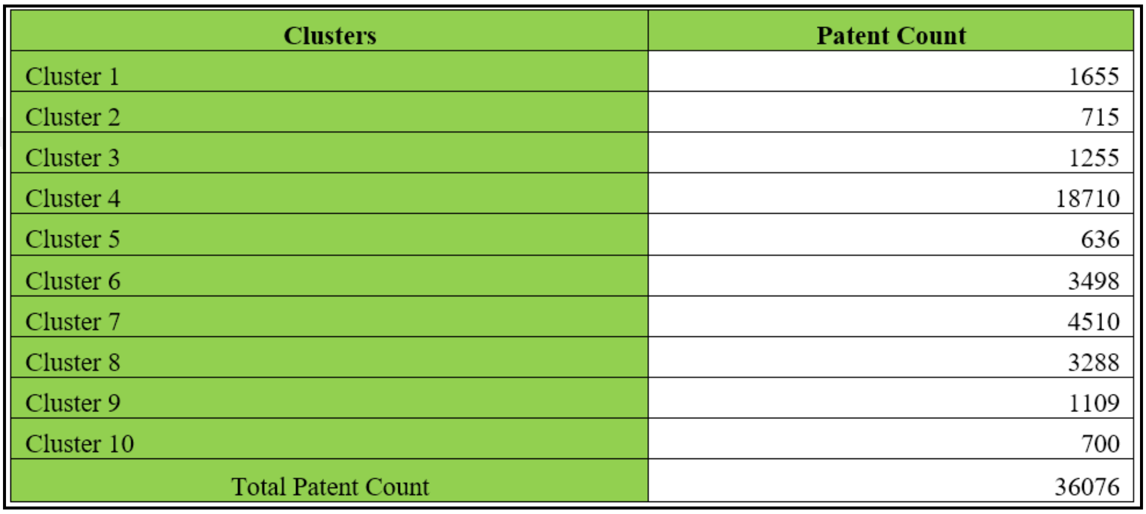
These patent documents prove the relations between the gaming and technology fields. Also, the patents are below, verify the relation between the finance and the biology field.

a) Method and system for clustering optimization and applications
b) Systems and methods for modeling and analyzing networks
c) Methods and Systems of Automatic Ontology Population

4.3 K-MEANS EUCLIDIAN DISTANCE 1000 WORDS 10 CLUSTERS (95 PERCENT VARIANCE)

In this experiment, a thousand words used and ten clusters were generated. The Euclidian similarity metric was used to identify the nearest neighbors. After applying PCA, the 95% of variance is preserved.

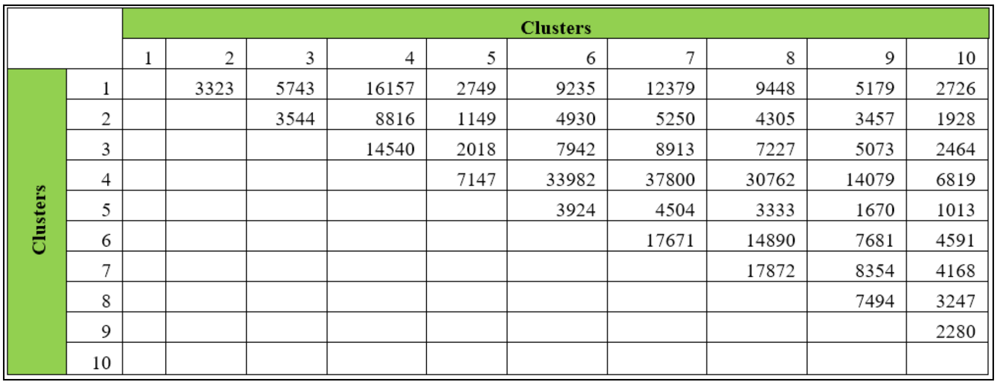
Figure 4.9: Distribution of patents according to defined clustering algorithm



Clusters	Patent Count
Cluster 1	1655
Cluster 2	715
Cluster 3	1255
Cluster 4	18710
Cluster 5	636
Cluster 6	3498
Cluster 7	4510
Cluster 8	3288
Cluster 9	1109
Cluster 10	700
Total Patent Count	36076

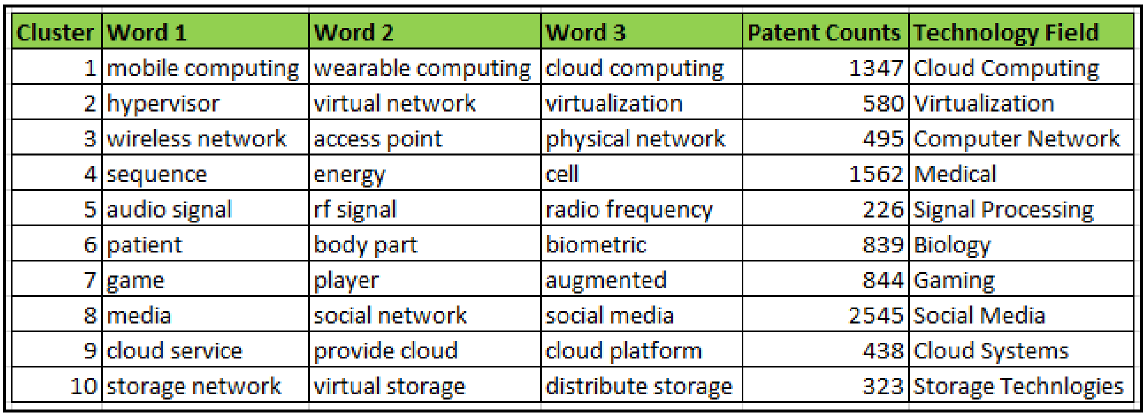
After applying the method, the patent cluster distribution shown in table 4.9. As can be seen in figure 4.9, patents did not distribute well to each cluster. Almost half of the patent documents fall into the 4th cluster. The figure 4.10 shows the relation among each cluster.

Figure 4.10: Cluster-relation table according to defined algorithm



		Clusters									
		1	2	3	4	5	6	7	8	9	10
Clusters	1		3323	5743	16157	2749	9235	12379	9448	5179	2726
	2			3544	8816	1149	4930	5250	4305	3457	1928
	3				14540	2018	7942	8913	7227	5073	2464
	4					7147	33982	37800	30762	14079	6819
	5						3924	4504	3333	1670	1013
	6							17671	14890	7681	4591
	7								17872	8354	4168
	8									7494	3247
	9										2280
	10										

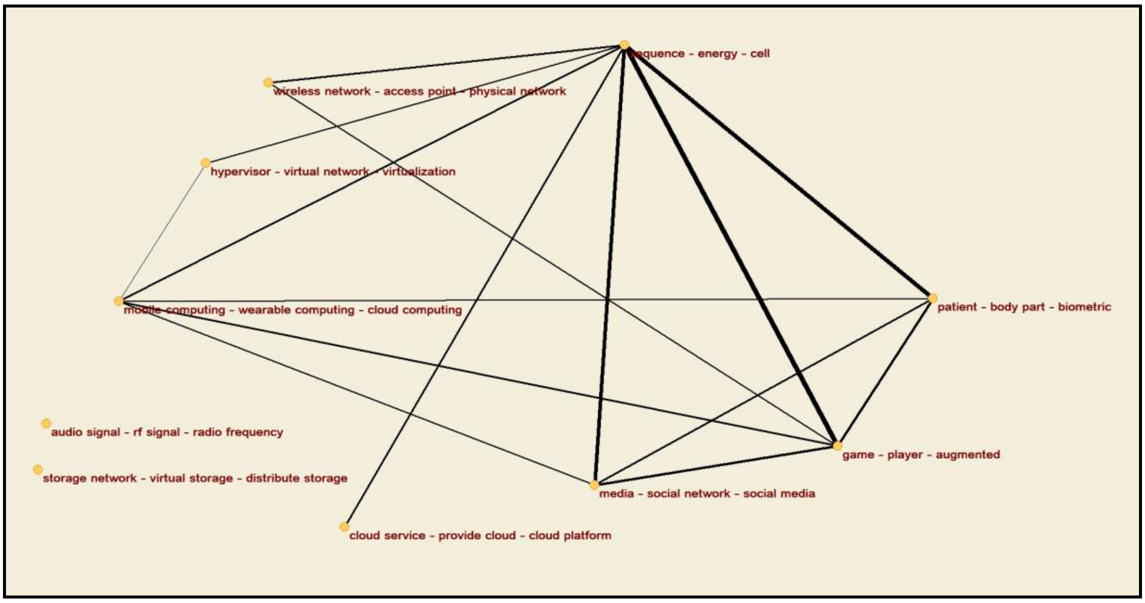
Figure 4.11: Top 3 words of the result of defined algorithm



Cluster	Word 1	Word 2	Word 3	Patent Counts	Technology Field
1	mobile computing	wearable computing	cloud computing	1347	Cloud Computing
2	hypervisor	virtual network	virtualization	580	Virtualization
3	wireless network	access point	physical network	495	Computer Network
4	sequence	energy	cell	1562	Medical
5	audio signal	rf signal	radio frequency	226	Signal Processing
6	patient	body part	biometric	839	Biology
7	game	player	augmented	844	Gaming
8	media	social network	social media	2545	Social Media
9	cloud service	provide cloud	cloud platform	438	Cloud Systems
10	storage network	virtual storage	distribute storage	323	Storage Technlogies

As it can be seen in figure 4.11, there are 10 technology fields have been found in this method. The most of the patents belong to the Social Media field. The least number of patents belong to the Signal Processing. Although the words like cloud service, provide cloud, and cloud platform seems are near to the Cloud computing, they were categorized under different technology field.

Figure 4.12: Visualization of relations among clusters


sequence - energy - cell
wireless network - access point - physical network
hypervisor - virtual network - virtualization
mobile computing - wearable computing - cloud computing
patient - body part - biometric
audio signal - rf signal - radio frequency
storage network - virtual storage - distribute storage
game - player - augmented
media - social network - social media
cloud service - provide cloud - cloud platform

Each point represents a cluster. Each line shows the relation among clusters. The strength of a relation among clusters can be seen by the thickness of a line. The relations of the clusters are not good because of the unfair distribution of the patents. In this diagram, there are some relations between gaming and social media. To verify these relations, the patents can be seen below.

a) Social media applications for a wager-based gaming system
b) Integrating social communities and wagering games
c) System and method for social networking in a gaming environment

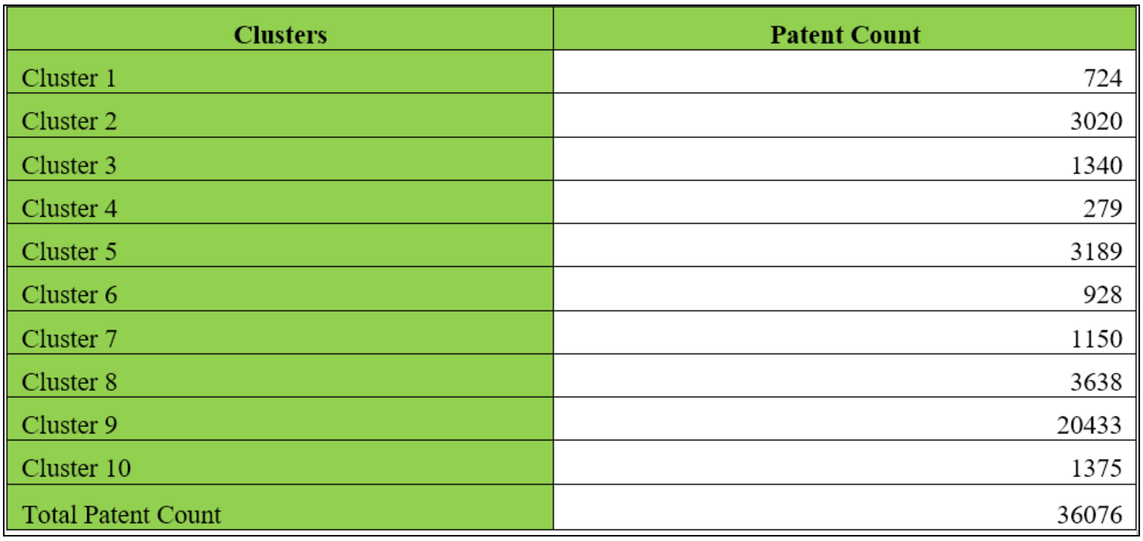
Therefore, there is also a relation between cloud computing and the biology field. To verify this relation, patent documents can be found below.

a) Apparatus and method for providing media commerce platform
b) Portable wireless personal head impacts reporting system
c) Systems and methods for sample processing and analysis

4.4 K-MEANS EUCLIDIAN DISTANCE 1000 WORDS 10 CLUSTERS (80 PERCENT VARIANCE)

In this experiment, 1000 words used and 10 clusters were generated. The Euclidian similarity metric was used. After applying PCA, 80% of variance was preserved.

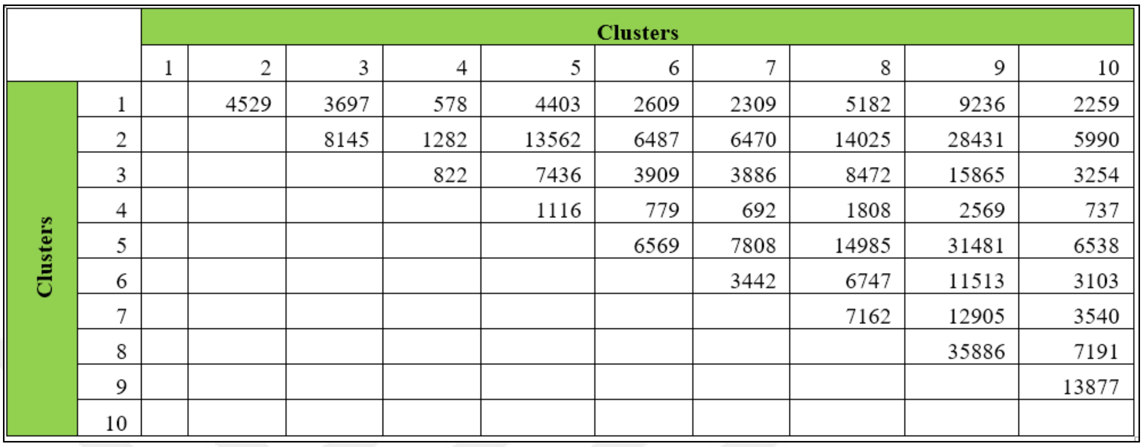
Figure 4.13: Distribution of patents according to defined clustering algorithm



Clusters	Patent Count
Cluster 1	724
Cluster 2	3020
Cluster 3	1340
Cluster 4	279
Cluster 5	3189
Cluster 6	928
Cluster 7	1150
Cluster 8	3638
Cluster 9	20433
Cluster 10	1375
Total Patent Count	36076

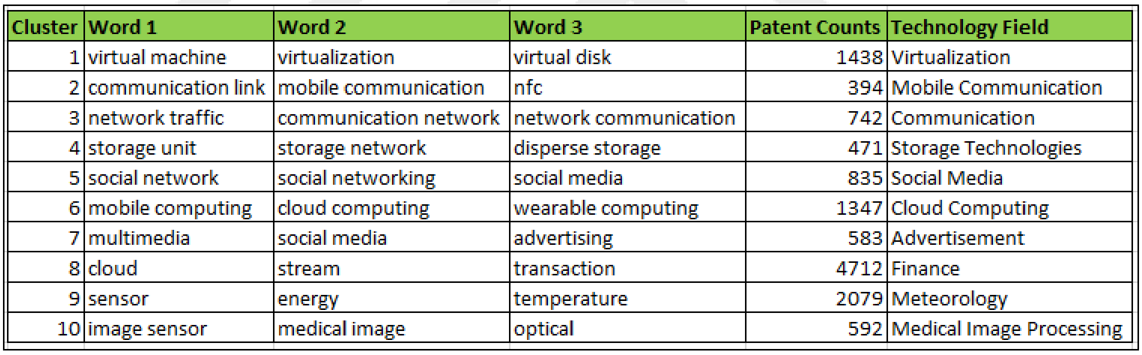
After applying the method, the patent cluster distribution can be seen in figure 4.13. As can be seen in figure 4.13, patents did not distribute well to each cluster. This unfair distribution will affect the visualization part. The figure 4.14 shows the relation among each cluster.

Figure 4.14: Cluster-relation table according to defined algorithm



		Clusters									
		1	2	3	4	5	6	7	8	9	10
Clusters	1		4529	3697	578	4403	2609	2309	5182	9236	2259
	2			8145	1282	13562	6487	6470	14025	28431	5990
	3				822	7436	3909	3886	8472	15865	3254
	4					1116	779	692	1808	2569	737
	5						6569	7808	14985	31481	6538
	6							3442	6747	11513	3103
	7								7162	12905	3540
	8									35886	7191
	9										13877
	10										

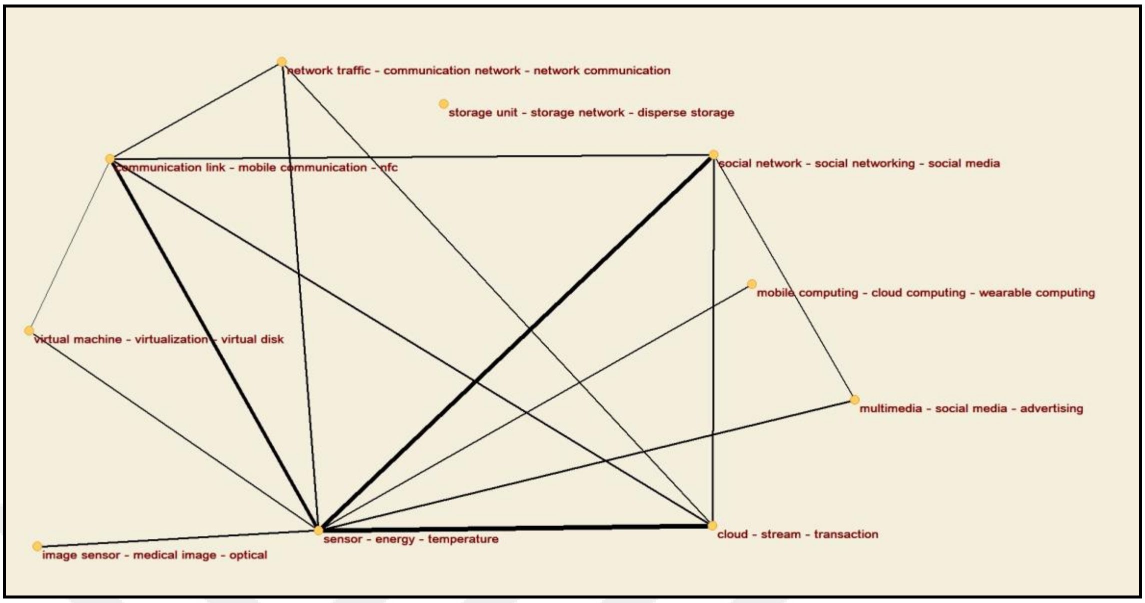
Figure 4.15: Top 3 words of the result of defined algorithm



Cluster	Word 1	Word 2	Word 3	Patent Counts	Technology Field
1	virtual machine	virtualization	virtual disk	1438	Virtualization
2	communication link	mobile communication	nfc	394	Mobile Communication
3	network traffic	communication network	network communication	742	Communication
4	storage unit	storage network	disperse storage	471	Storage Technologies
5	social network	social networking	social media	835	Social Media
6	mobile computing	cloud computing	wearable computing	1347	Cloud Computing
7	multimedia	social media	advertising	583	Advertisement
8	cloud	stream	transaction	4712	Finance
9	sensor	energy	temperature	2079	Meteorology
10	image sensor	medical image	optical	592	Medical Image Processing

As it can be seen figure 4.15, there are 10 technologies in this experiment. The most patents belong to the Finance field. The least number of patent documents belongs to the Mobile Communication field. The interesting field which does not occur in other experiments is Medical Image Processing, though it has a limit number of patents.

Figure 4.16: Visualization of relations among clusters


network traffic - communication network - network communication
storage unit - storage network - disperse storage
communication link - mobile communication - nfc
social network - social networking - social media
mobile computing - cloud computing - wearable computing
virtual machine - virtualization - virtual disk
multimedia - social media - advertising
sensor - energy - temperature
cloud - stream - transaction
image sensor - medical image - optical

As it can be seen in figure 4.16, the image is not so meaningful because of weak patent distributions. There are some technologies which have strong relations such as meteorology, social media, and mobile communication. There is also the relation between mobile communication and finance. To verify this relation, patents can be seen below.

a) Mobile electronic wallet
b) Mobile device credit account
c) Mobile communication facility usage and social network creation

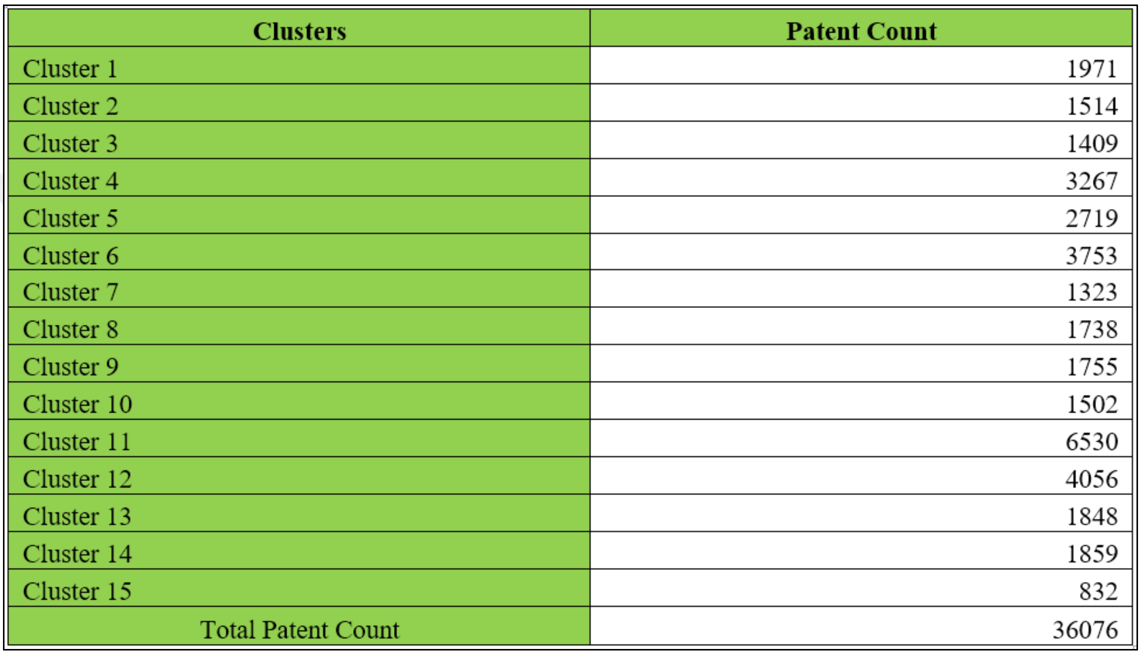
These patents both related to mobile communication and finance fields. Also, there is another promising relation between medical image and meteorology. To verify this relation, patents can be seen below.

a) Image data navigation method and apparatus
b) Method and apparatus for performing linear filtering in wavelet based domain
c) Systems and methods for image resolution enhancement

4.5 K-MEANS COSINE DISTANCE 1000 WORDS 15 CLUSTERS (95 PERCENT VARIANCE)

In this experiment, 1000 words used and 15 clusters were generated. The Cosine similarity metric was used and PCA was applied so that 95% of variance is preserved.

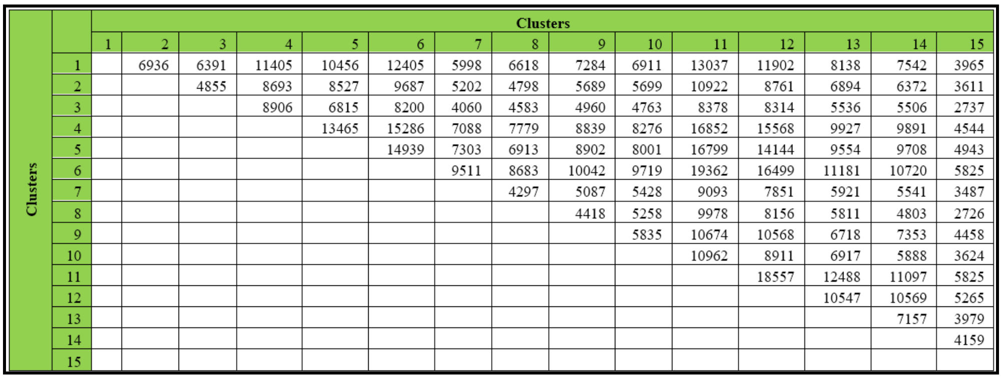
Figure 4.17: Distribution of patents according to defined clustering algorithm



Clusters	Patent Count
Cluster 1	1971
Cluster 2	1514
Cluster 3	1409
Cluster 4	3267
Cluster 5	2719
Cluster 6	3753
Cluster 7	1323
Cluster 8	1738
Cluster 9	1755
Cluster 10	1502
Cluster 11	6530
Cluster 12	4056
Cluster 13	1848
Cluster 14	1859
Cluster 15	832
Total Patent Count	36076

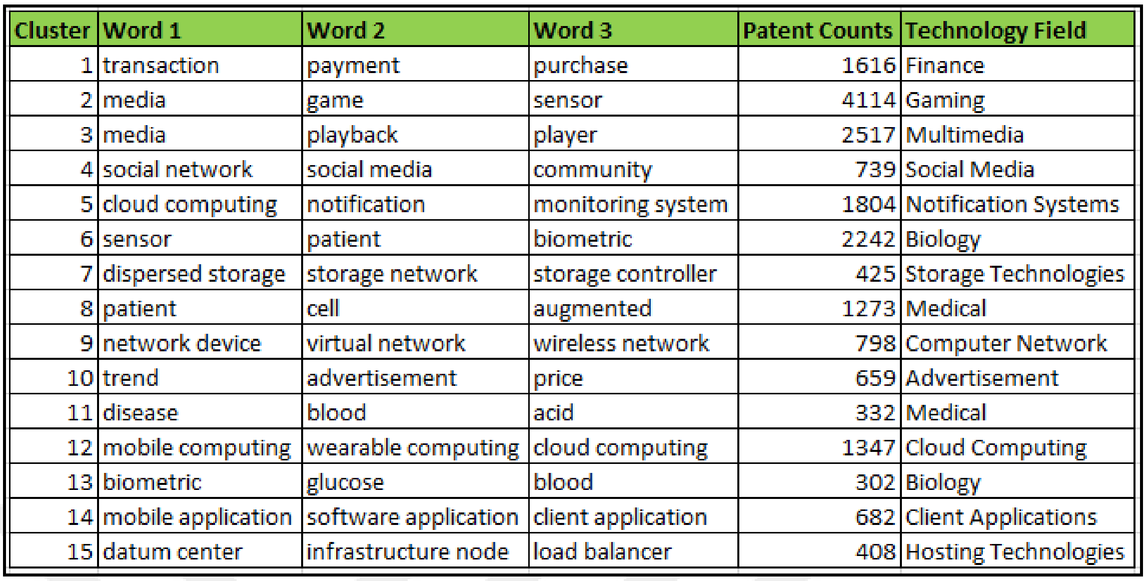
After applying the method, the patent cluster distribution can be seen in figure 4.17. As can be seen in figure 4.17, patents distributed well to each cluster. The figure 4.18 shows the relation among each cluster.

Figure 4.18: Cluster-relation table according to defined algorithm



		Clusters														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Clusters	1		6936	6391	11405	10456	12405	5998	6618	7284	6911	13037	11902	8138	7542	3965
	2			4855	8693	8527	9687	5202	4798	5689	5699	10922	8761	6894	6372	3611
	3				8906	6815	8200	4060	4583	4960	4763	8378	8314	5536	5506	2737
	4					13465	15286	7088	7779	8839	8276	16852	15568	9927	9891	4544
	5						14939	7303	6913	8902	8001	16799	14144	9554	9708	4943
	6							9511	8683	10042	9719	19362	16499	11181	10720	5825
	7								4297	5087	5428	9093	7851	5921	5541	3487
	8									4418	5258	9978	8156	5811	4803	2726
	9										5835	10674	10568	6718	7353	4458
	10											10962	8911	6917	5888	3624
	11												18557	12488	11097	5825
	12													10547	10569	5265
	13														7157	3979
	14															4159
	15															

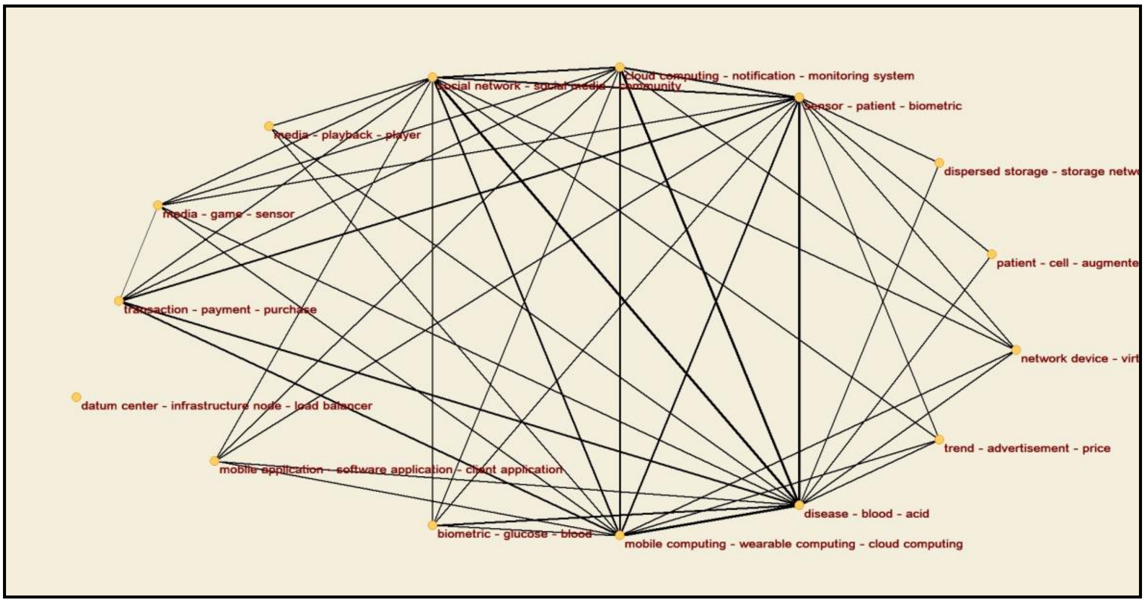
Figure 4.19: Top 3 words of the result of defined algorithm



Cluster	Word 1	Word 2	Word 3	Patent Counts	Technology Field
1	transaction	payment	purchase	1616	Finance
2	media	game	sensor	4114	Gaming
3	media	playback	player	2517	Multimedia
4	social network	social media	community	739	Social Media
5	cloud computing	notification	monitoring system	1804	Notification Systems
6	sensor	patient	biometric	2242	Biology
7	dispersed storage	storage network	storage controller	425	Storage Technologies
8	patient	cell	augmented	1273	Medical
9	network device	virtual network	wireless network	798	Computer Network
10	trend	advertisement	price	659	Advertisement
11	disease	blood	acid	332	Medical
12	mobile computing	wearable computing	cloud computing	1347	Cloud Computing
13	biometric	glucose	blood	302	Biology
14	mobile application	software application	client application	682	Client Applications
15	datum center	infrastructure node	load balancer	408	Hosting Technologies

In the figure 4.19, there are 13 different technologies. Some of the technologies are repeated such as Biology and Medical. This shows that the number of clusters should be decreased. The most number of patents belongs to Gaming field. Also, the least number of patents belongs to Biology field.

Figure 4.20: Visualization of relations among clusters


cloud computing - notification - monitoring system
social network - social media - community
sensor - patient - biometric
media - playback - player
dispersed storage - storage netw
media - game - sensor
patient - cell - augmente
transaction - payment - purchase
network device - virt
datum center - infrastructure node - load balancer
trend - advertisement - price
mobile application - software application - client application
disease - blood - acid
biometric - glucose - blood
mobile computing - wearable computing - cloud computing

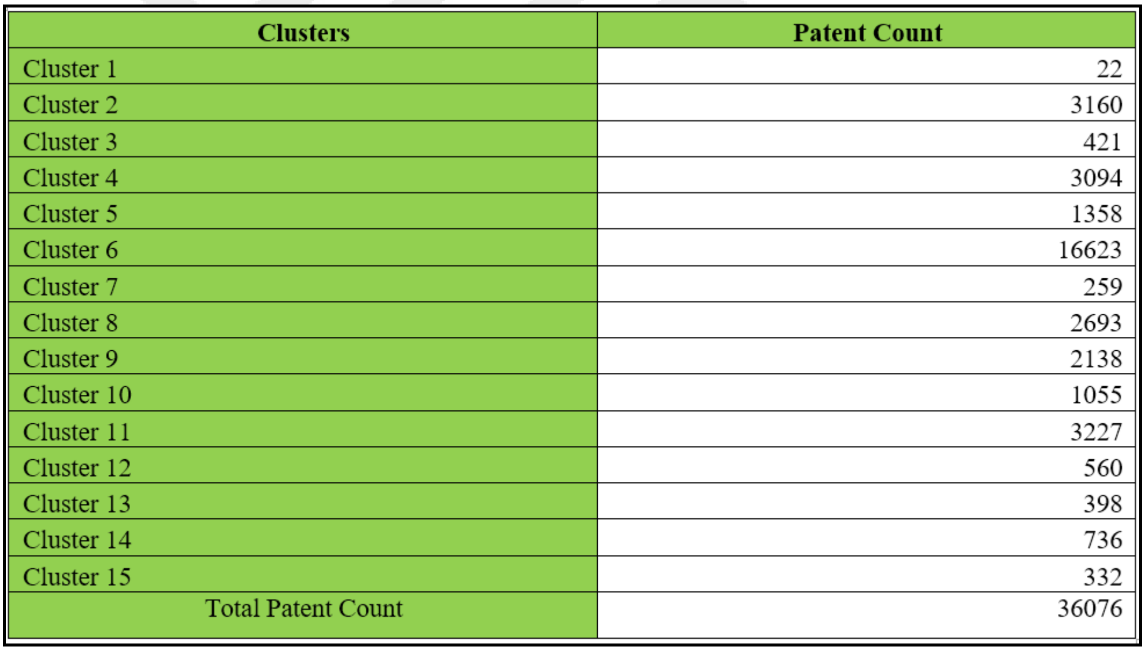
As it can be seen in figure 4.20, the diagram is so complex because of the cluster number but there are some good technology relations such as medical and biology, social media and medical. To verify the relation between social media and medical field, patents can be seen below.

a) Connecting users based on medical experiences
b) Computational system and method for memory modification
c) Ingestible event marker data framework

4.6 K-MEANS EUCLIDIAN DISTANCE 1000 WORDS 15 CLUSTERS (95 PERCENT VARIANCE)

In this experiment, 1000 words used and 15 clusters were generated. The Euclidian similarity metric was used.

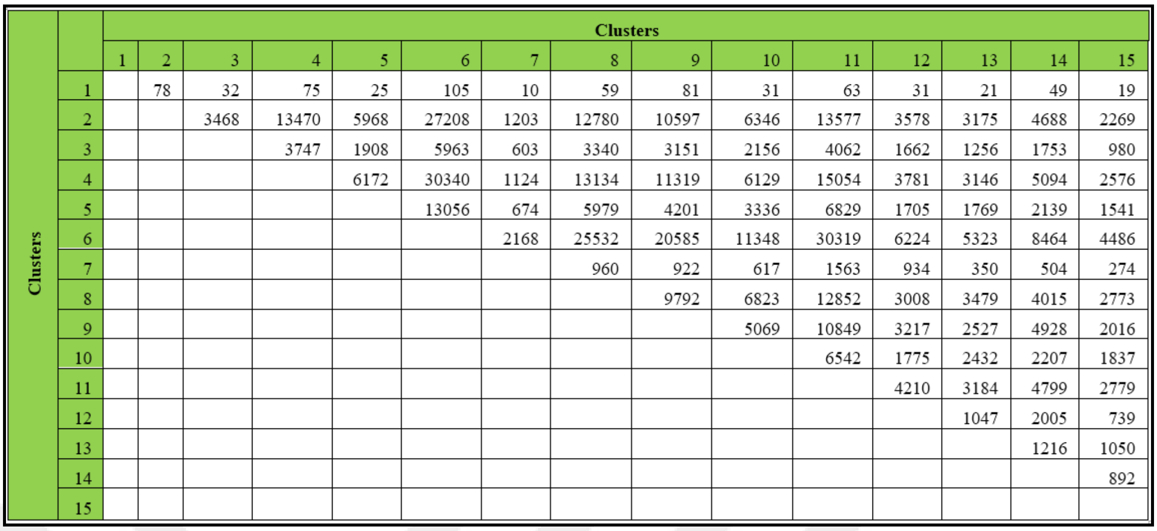
Figure 4.21: Distribution of patents according to defined clustering algorithm



Clusters	Patent Count
Cluster 1	22
Cluster 2	3160
Cluster 3	421
Cluster 4	3094
Cluster 5	1358
Cluster 6	16623
Cluster 7	259
Cluster 8	2693
Cluster 9	2138
Cluster 10	1055
Cluster 11	3227
Cluster 12	560
Cluster 13	398
Cluster 14	736
Cluster 15	332
Total Patent Count	36076

After applying the method, the patent cluster distribution can be seen in table 4.21. As can be seen in Figure 4.21, patents did not distribute well to each cluster.

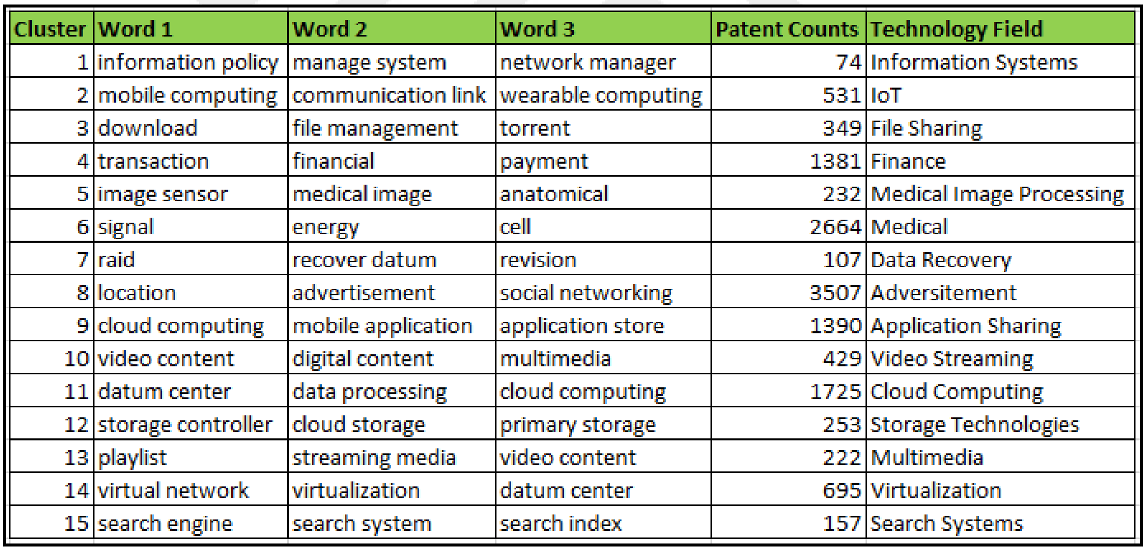
Figure 4.22: Cluster-relation table according to defined algorithm



		Clusters														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Clusters	1		78	32	75	25	105	10	59	81	31	63	31	21	49	19
	2			3468	13470	5968	27208	1203	12780	10597	6346	13577	3578	3175	4688	2269
	3				3747	1908	5963	603	3340	3151	2156	4062	1662	1256	1753	980
	4					6172	30340	1124	13134	11319	6129	15054	3781	3146	5094	2576
	5						13056	674	5979	4201	3336	6829	1705	1769	2139	1541
	6							2168	25532	20585	11348	30319	6224	5323	8464	4486
	7								960	922	617	1563	934	350	504	274
	8									9792	6823	12852	3008	3479	4015	2773
	9										5069	10849	3217	2527	4928	2016
	10											6542	1775	2432	2207	1837
	11												4210	3184	4799	2779
	12													1047	2005	739
	13														1216	1050
	14															892
	15															

The figure 4.22 shows the relation among each cluster.

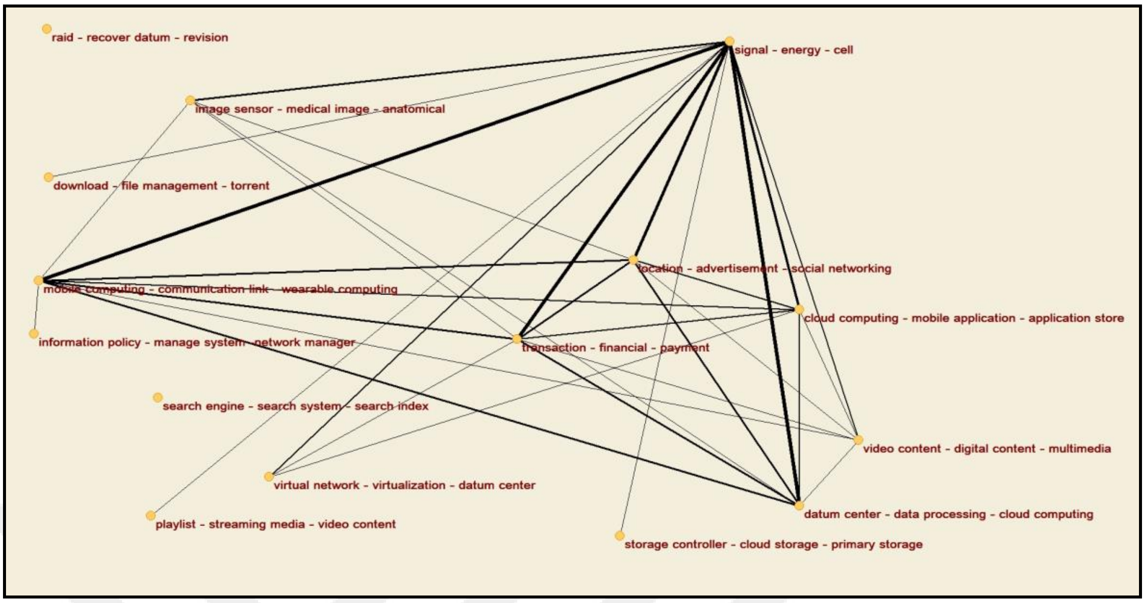
Figure 4.23: Top 3 words of the result of defined algorithm



Cluster	Word 1	Word 2	Word 3	Patent Counts	Technology Field
1	information policy	manage system	network manager	74	Information Systems
2	mobile computing	communication link	wearable computing	531	IoT
3	download	file management	torrent	349	File Sharing
4	transaction	financial	payment	1381	Finance
5	image sensor	medical image	anatomical	232	Medical Image Processing
6	signal	energy	cell	2664	Medical
7	raid	recover datum	revision	107	Data Recovery
8	location	advertisement	social networking	3507	Adversitement
9	cloud computing	mobile application	application store	1390	Application Sharing
10	video content	digital content	multimedia	429	Video Streaming
11	datum center	data processing	cloud computing	1725	Cloud Computing
12	storage controller	cloud storage	primary storage	253	Storage Technologies
13	playlist	streaming media	video content	222	Multimedia
14	virtual network	virtualization	datum center	695	Virtualization
15	search engine	search system	search index	157	Search Systems

As it can be seen in figure 4.23, there are 15 different technology fields have been occurred. The least number of patents belongs to the Information Systems. Also, the most number of patents belongs to the Advertisement field. The 5th and 6th clusters seem similar although they are not. They differ each other with the image words.

Figure 4.24: Visualization of relations among clusters


raid - recover datum - revision
signal - energy - cell
image sensor - medical image - anatomical
download - file management - torrent
location - advertisement - social networking
mobile computing - communication link - wearable computing
cloud computing - mobile application - application store
information policy - manage system - network manager
transaction - financial - payment
search engine - search system - search index
video content - digital content - multimedia
virtual network - virtualization - datum center
datum center - data processing - cloud computing
playlist - streaming media - video content
storage controller - cloud storage - primary storage

As it can be in the previous image, this one also looks complicated because of weak distribution. There are some technologies which have strong relations such as cloud computing, finance, hosting, and advertisement. To verify the relation between Hosting and Medical fields, the patents can be seen below.

a) System and method for a personal computer medical device based away from a hospital
b) System and method for handling the acquisition and analysis of medical data over a network
c) Medical device systems implemented network scheme for remote patient management

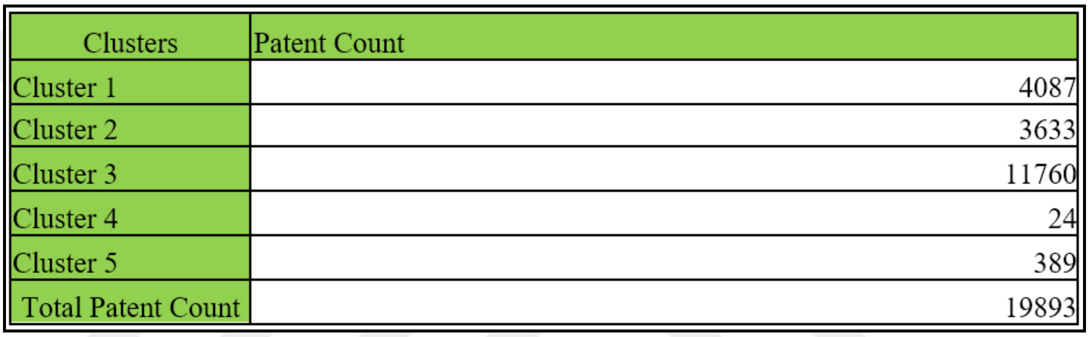
There is also another strong relation between Finance and Hosting. To verify this relation, the patents can be seen below.

a) Multilayer policy language structure
b) Portfolio synchronizing between different interfaces
c) Method and system for modifying host application functionality based upon downloaded content

4.7 SPECTRAL CLUSTERING 1000 WORDS 5 CLUSTERS (95 PERCENT VARIANCE)

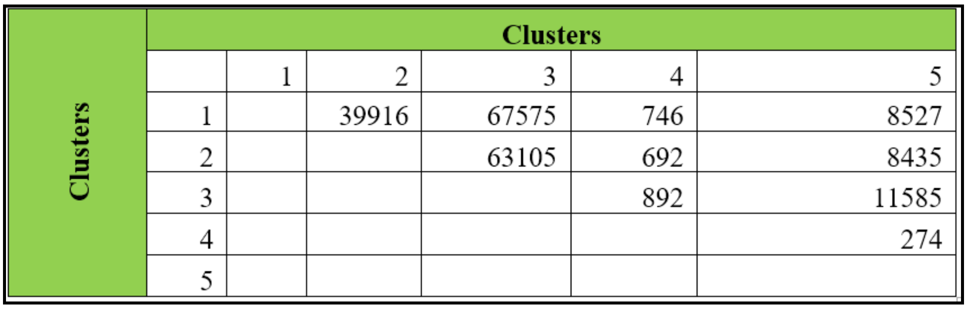
In this experiment, 1000 words used and 5 clusters were generated. After applying the method, the patent cluster distribution can be seen in the figure 4.25. As can be seen in figure 4.25, most of the patents fall into the 3rd cluster. The figure 4.26 shows the number of common words which appear in the patent documents that fall into the corresponding pair of clusters.

Figure 4.25: Distribution of patents according to defined clustering algorithm



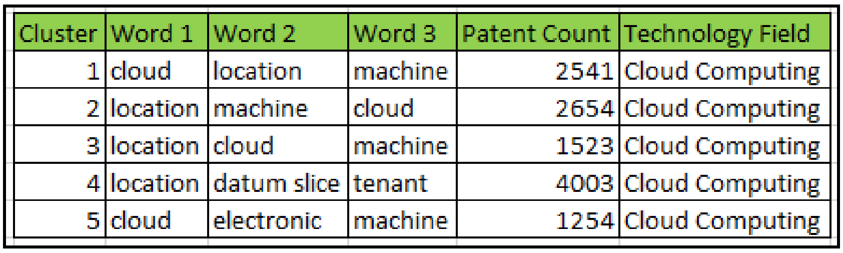
Clusters	Patent Count
Cluster 1	4087
Cluster 2	3633
Cluster 3	11760
Cluster 4	24
Cluster 5	389
Total Patent Count	19893

Figure 4.26: Cluster-relation table according to defined algorithm



		Clusters				
		1	2	3	4	5
Clusters	1		39916	67575	746	8527
	2			63105	692	8435
	3				892	11585
	4					274
	5					

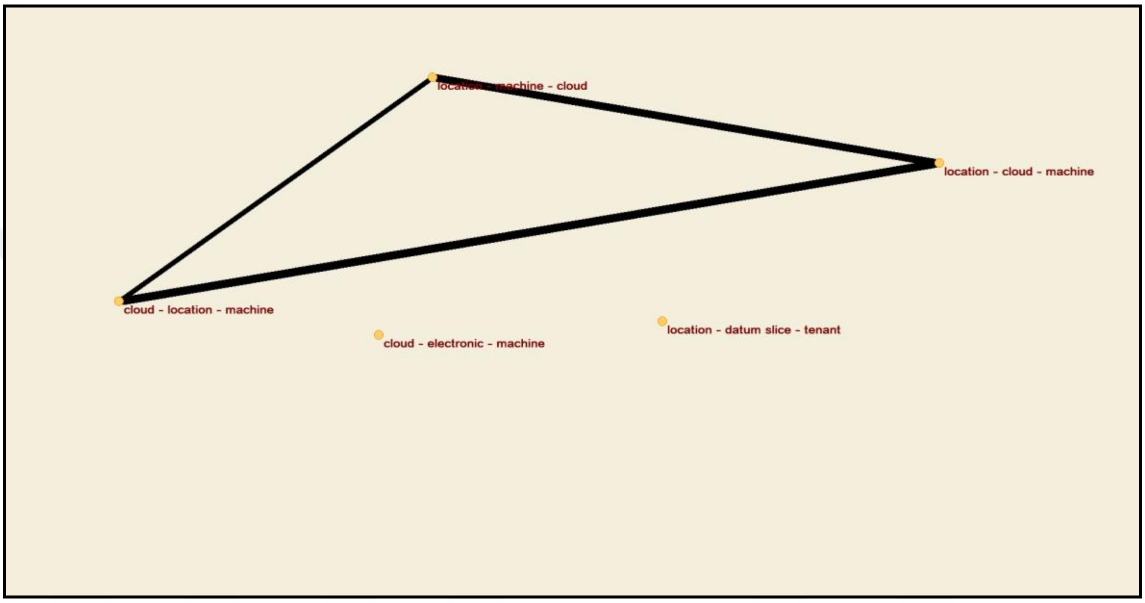
Figure 4.27: Top 3 words of the result of defined algorithm



Cluster	Word 1	Word 2	Word 3	Patent Count	Technology Field
1	cloud	location	machine	2541	Cloud Computing
2	location	machine	cloud	2654	Cloud Computing
3	location	cloud	machine	1523	Cloud Computing
4	location	datum slice	tenant	4003	Cloud Computing
5	cloud	electronic	machine	1254	Cloud Computing

As it can be seen in the figure 4.27, technology fields could not be determined. Spectral clustering did not work well, although the patent distributions are good enough. In the figure 4.28, it can be seen the relations. The relations are meaningless because of the undetermined technologies.

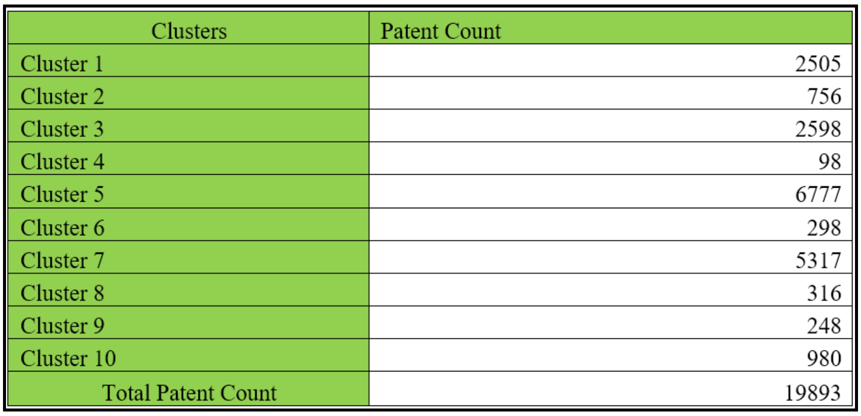
Figure 4.28: Visualization of relations among clusters


location - machine - cloud
location - cloud - machine
cloud - location - machine
location - datum slice - tenant
cloud - electronic - machine

4.8 SPECTRAL CLUSTERING 1000 WORDS 10 CLUSTERS (95 PERCENT VARIANCE)

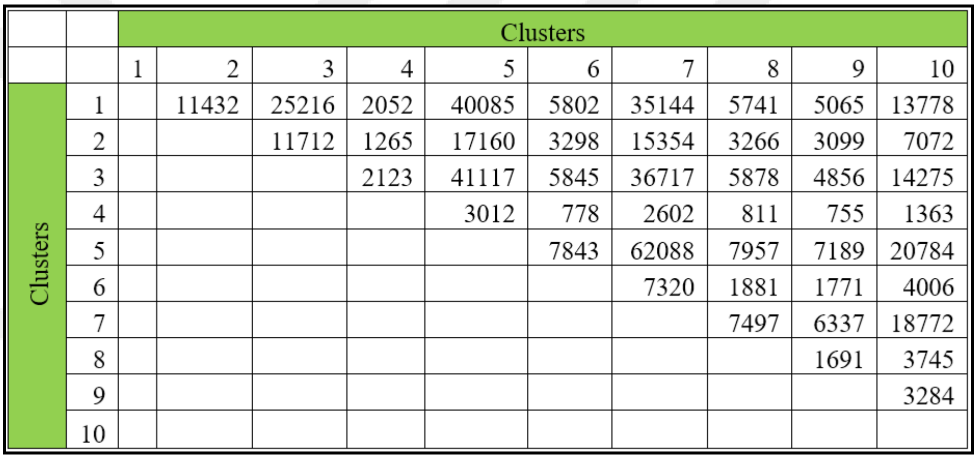
In this experiment, 1000 words used and 10 clusters were generated. After applying the method, the patent cluster distribution can be seen in figure 4.29. The figure 4.30 shows the relation among each cluster

Figure 4.29: Distribution of patents according to defined clustering algorithm



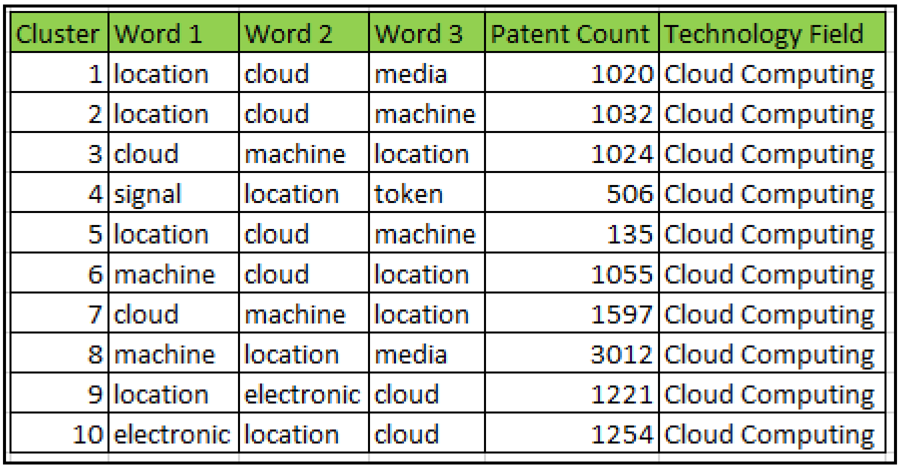
Clusters	Patent Count
Cluster 1	2505
Cluster 2	756
Cluster 3	2598
Cluster 4	98
Cluster 5	6777
Cluster 6	298
Cluster 7	5317
Cluster 8	316
Cluster 9	248
Cluster 10	980
Total Patent Count	19893

Figure 4.30: Cluster-relation table according to defined algorithm



		Clusters									
		1	2	3	4	5	6	7	8	9	10
Clusters	1		11432	25216	2052	40085	5802	35144	5741	5065	13778
	2			11712	1265	17160	3298	15354	3266	3099	7072
	3				2123	41117	5845	36717	5878	4856	14275
	4					3012	778	2602	811	755	1363
	5						7843	62088	7957	7189	20784
	6							7320	1881	1771	4006
	7								7497	6337	18772
	8									1691	3745
	9										3284
	10										

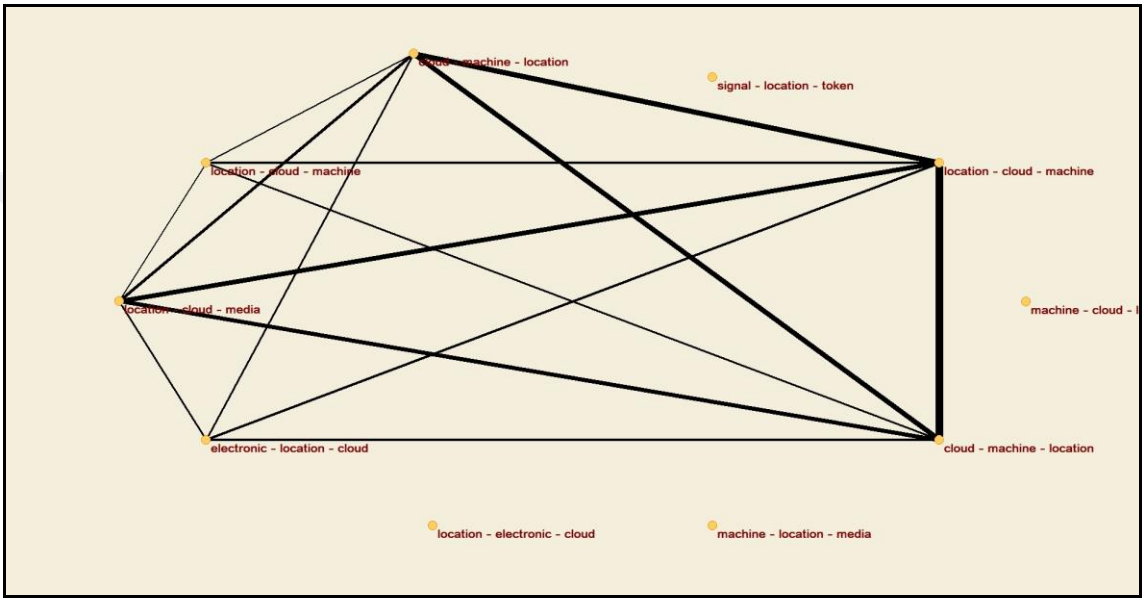
Figure 4.31: Top 3 words of the result of defined algorithm



Cluster	Word 1	Word 2	Word 3	Patent Count	Technology Field
1	location	cloud	media	1020	Cloud Computing
2	location	cloud	machine	1032	Cloud Computing
3	cloud	machine	location	1024	Cloud Computing
4	signal	location	token	506	Cloud Computing
5	location	cloud	machine	135	Cloud Computing
6	machine	cloud	location	1055	Cloud Computing
7	cloud	machine	location	1597	Cloud Computing
8	machine	location	media	3012	Cloud Computing
9	location	electronic	cloud	1221	Cloud Computing
10	electronic	location	cloud	1254	Cloud Computing

This method is not promising as the previous one. The clusters do not include patent documents that refer to a specific field. So, the spectral clustering with the parameters used in our experiments did not result in acceptable results. In the figure 4.32, visualization of relations among clusters is presented. As it is seen, the most frequent words in different clusters are almost identical.

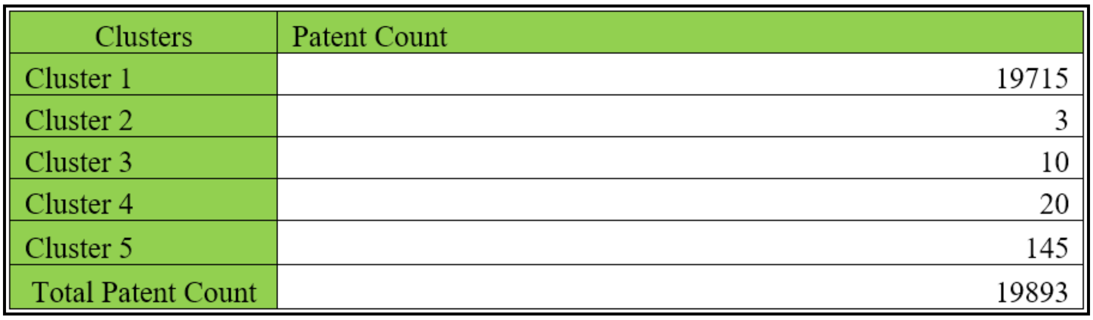
Figure 4.32: Visualization of relations among clusters


cloud - machine - location
signal - location - token
location - cloud - machine
location - cloud - machine
location - cloud - media
machine - cloud - l
electronic - location - cloud
cloud - machine - location
location - electronic - cloud
machine - location - media

4.9 POWER ITERATION 1000 WORDS 5 CLUSTERS (95 PERCENT VARIANCE)

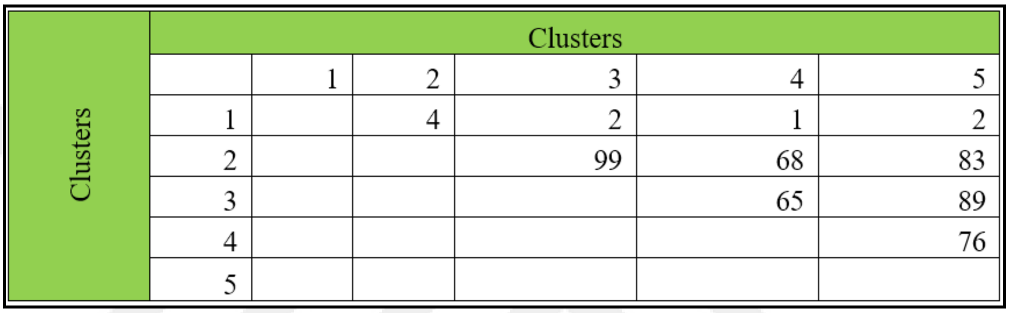
In this experiment, 1000 words used and 5 clusters were generated. The distribution of patent documents to the clusters obtained with power iteration clustering is shown in figure 4.33. As can be seen in the figure 4.33, almost all of the clusters have fall into the first cluster which shows that the algorithm could not divide the documents into distinct groups using the bag-of-words representation. Figure 4.34 shows the relations among each cluster.

Figure 4.33: Distribution of patents according to defined clustering algorithm



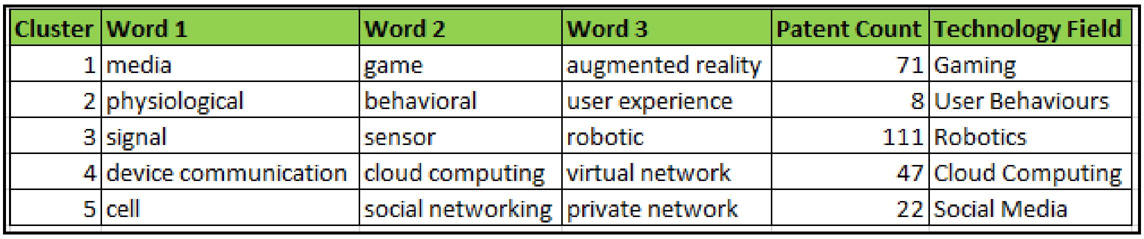
Clusters	Patent Count
Cluster 1	19715
Cluster 2	3
Cluster 3	10
Cluster 4	20
Cluster 5	145
Total Patent Count	19893

Figure 4.34: Cluster-relation table according to defined algorithm



		Clusters				
Clusters		1	2	3	4	5
	1		4	2	1	2
	2			99	68	83
	3				65	89
	4					76
	5					

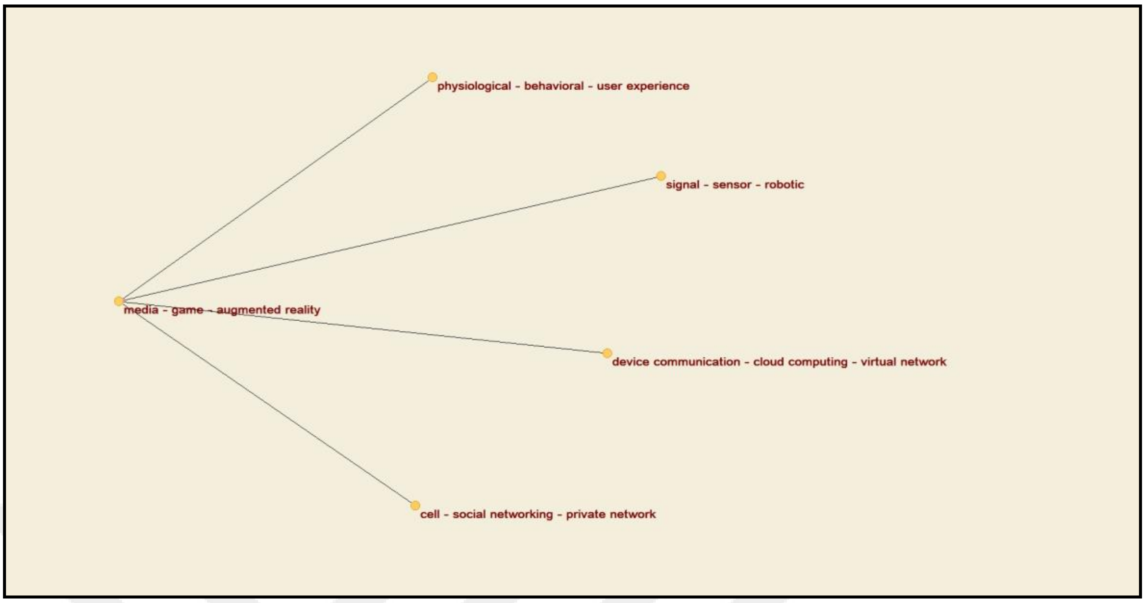
Figure 4.35: Top 3 words of the result of defined algorithm



Cluster	Word 1	Word 2	Word 3	Patent Count	Technology Field
1	media	game	augmented reality	71	Gaming
2	physiological	behavioral	user experience	8	User Behaviours
3	signal	sensor	robotic	111	Robotics
4	device communication	cloud computing	virtual network	47	Cloud Computing
5	cell	social networking	private network	22	Social Media

In the figure 4.35, there are 5 technology fields such as Gaming, User Behaviours, Robotics, Cloud Computing, and Social Media. The most number of patents belongs to Robotics field. The least number of patents belongs to User Behaviours. However, these results are not reliable since the distribution of patent documents in the clusters is not balanced and four of the five clusters contain a few number of patent documents.

Figure 4.36: Visualization of relations among clusters


physiological - behavioral - user experience
signal - sensor - robotic
media - game - augmented reality
device communication - cloud computing - virtual network
cell - social networking - private network

The image of this method is not successful because of the weak cluster distribution but there are some relations such as Robotics and Gaming field. To verify this relation, the patents are shown below.

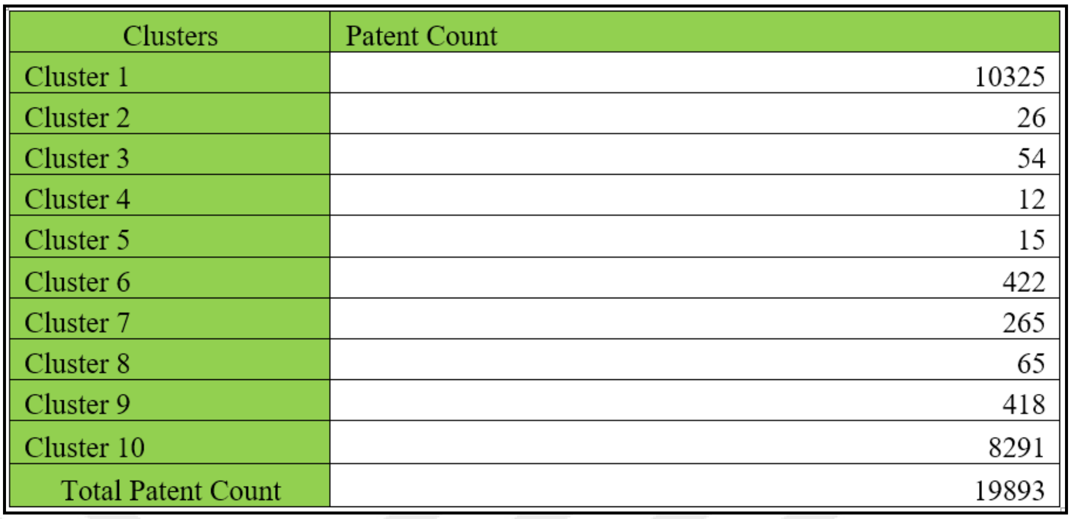
a) Virtual ankle and balance trainer system
b) Mobile Camera Localization Using Depth Maps
c) Method and system for obtaining positioning data

Although the patent distributions are not good, the technology extraction could perform well.

4.10 POWER ITERATION 1000 WORDS 10 CLUSTERS (95 PERCENT VARIANCE)

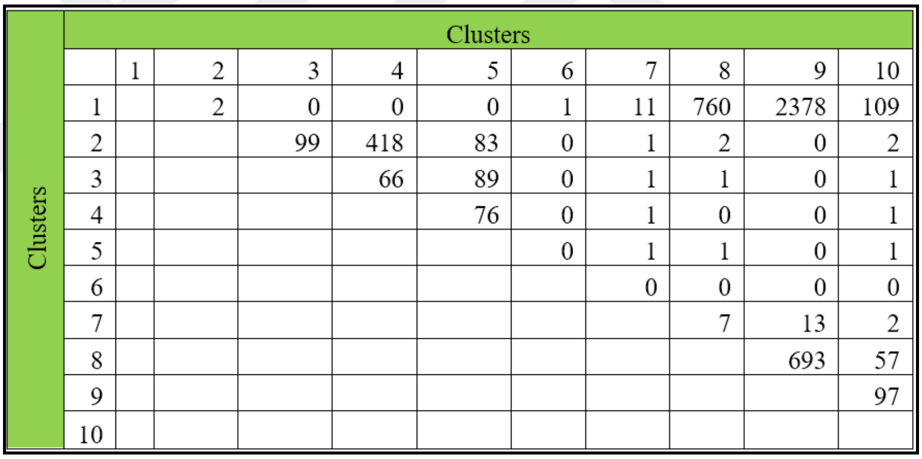
In this experiment, 1000 words used and 10 clusters were generated. After applying the method, the patent cluster distribution can be seen in figure 4.37. It is seen that most of the patent documents fall into two clusters (cluster 1 and cluster 10).

Figure 4.37: Distribution of patents according to defined clustering algorithm



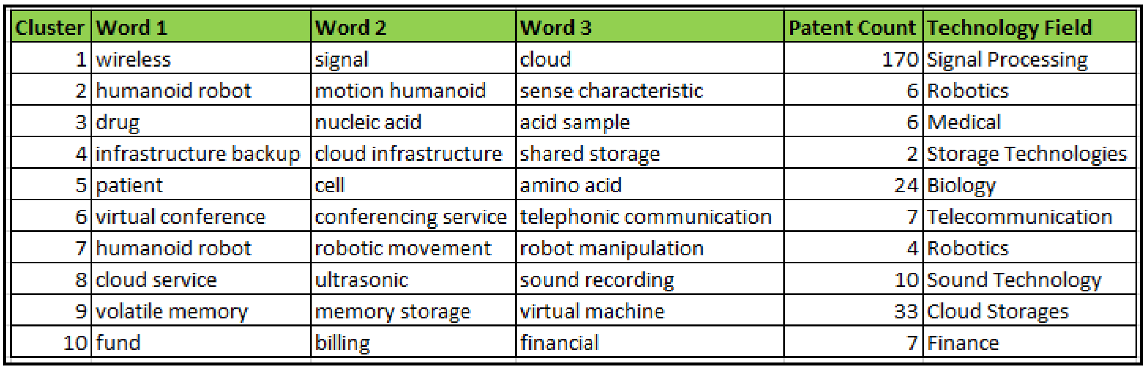
Clusters	Patent Count
Cluster 1	10325
Cluster 2	26
Cluster 3	54
Cluster 4	12
Cluster 5	15
Cluster 6	422
Cluster 7	265
Cluster 8	65
Cluster 9	418
Cluster 10	8291
Total Patent Count	19893

Figure 4.38: Cluster-relation table according to defined algorithm



		Clusters									
		1	2	3	4	5	6	7	8	9	10
Clusters	1		2	0	0	0	1	11	760	2378	109
	2			99	418	83	0	1	2	0	2
	3				66	89	0	1	1	0	1
	4					76	0	1	0	0	1
	5						0	1	1	0	1
	6							0	0	0	0
	7								7	13	2
	8									693	57
	9										97
	10										

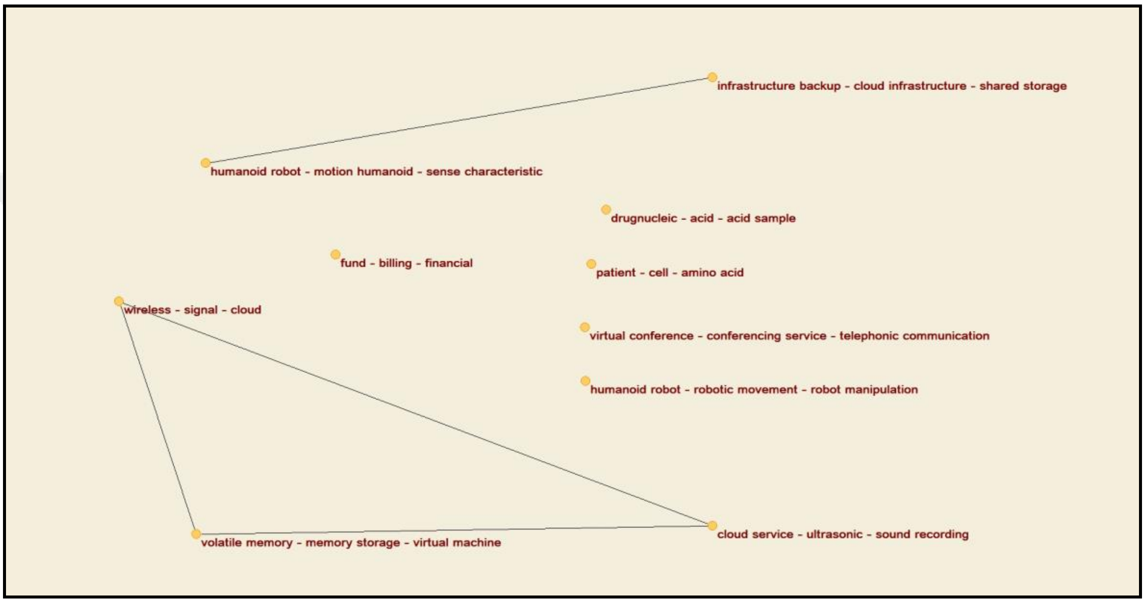
Figure 4.39: Top 3 words of the result of defined algorithm



Cluster	Word 1	Word 2	Word 3	Patent Count	Technology Field
1	wireless	signal	cloud	170	Signal Processing
2	humanoid robot	motion humanoid	sense characteristic	6	Robotics
3	drug	nucleic acid	acid sample	6	Medical
4	infrastructure backup	cloud infrastructure	shared storage	2	Storage Technologies
5	patient	cell	amino acid	24	Biology
6	virtual conference	conferencing service	telephonic communication	7	Telecommunication
7	humanoid robot	robotic movement	robot manipulation	4	Robotics
8	cloud service	ultrasonic	sound recording	10	Sound Technology
9	volatile memory	memory storage	virtual machine	33	Cloud Storages
10	fund	billing	financial	7	Finance

The figure 4.39 shows the top 3 words of the result of the defined algorithm. Clustering results were also visualized in the following figure. In this diagram, it can be seen that there are 10 different technology fields. The most number of patents belongs to Signal Processing field. The least number of patents belong to Storage Technologies. The technology extraction is as good as K-Means method.

Figure 4.40: Visualization of relations among clusters


infrastructure backup - cloud infrastructure - shared storage
humanoid robot - motion humanoid - sense characteristic
drugnucleic - acid - acid sample
fund - billing - financial
patient - cell - amino acid
wireless - signal - cloud
virtual conference - conferencing service - telephonic communication
humanoid robot - robotic movement - robot manipulation
cloud service - ultrasonic - sound recording
volatile memory - memory storage - virtual machine

It can be seen in the figure 4.40, the diagram is not good because of the weak cluster distributions. There are only 4 relations that can be seen. To verify the some of the relations, patents can be found below. For example, there is a relation between Sound Technology and Signal Processing fields.

a) Heart sound detecting system based on cloud computing
b) Apparatus and method for providing a game service in cloud computing environment
c) Dynamic negotiation and authorization system to record rights-managed content

5. CONCLUSION AND DISCUSSION

In this study, different clustering algorithms were applied to the patent documents including "cloud computing" term. The methods were performed on two different datasets which are last two years' patents and last three years' patents. This study aims to analyze cloud computing technology patents in terms of the technology fields it has been applied to. With the clustering methods, related words and patents summed up together. The idea of summing up all words and related patents shows categorized documents, so it can be defined the related technology or technologies.

Three main steps were followed in this study. The first step is related to gathering data from the internet and pre-processing it. It is also a big deal to gather data from patent websites. This is the main step of the study because all results are dependent on data quality. The search terms and resources specify the fate of the results. In the total study, 65868 patents gathered for all years. All patents are related to cloud computing. It is important to gather patents from different and high-quality patent sources.

The second step is to process patents and texts. Each patent processed separately. Each of the steps applied for all patents. There are many steps for processing and pre-processing phase as described in the material method section. Selected text mining algorithms specify the quality of the texts, so texts give the final results after processing step.

The final step contains clustering and visualization steps. This step is as important as a pre-processing step. The texts make no sense without clustering step. Many clustering algorithms were examined in this step. So, each clustering algorithm was performed with different parameters. Visualization is as fundamental as clustering step. It helps the data to be interpreted.

The results of the study are very interesting and surprising. In this study, K-Means, Spectral, and Power iteration were applied. Each algorithm tested many times and with different parameters. The results show that K-Means does the best distribution of patents. The distribution of the patents is very important. It can be seen that the most of the patents get in the one cluster and the other clusters have a small number of patents. It is so important to distribute patents to each cluster as well.

The similarity metric is also important to find the optimum cluster distribution. As can be seen in K-means results, the Cosine similarity distance is better than the Euclidian distance. This study suggests using the cosine similarity metric with K-Means for patents. The cloud computing is just a case study to apply our patent mining engine. The subject can be anything instead of the Cloud computing. This study shows that the best result can be taken by 95 percent Variance data with cosine similarity metric. It can be seen in the figures that different categories are related each other. For example, the words are related to the game word are linked with the words are “mobile communication”, “wearable device” and “augmented reality”. Also the words like “cell”, “sequence”, “disease” have strong relation with “sensor”, “signal”, “biometric”. It can be easily seen the relations between different kind of technologies. Although all patents are related to cloud computing technology, the results also show the real categories of the patents. Therefore, these categories are related to cloud computing field directly or indirectly.

This study fills the gap that is between patent industry and data science. Many studies only use the international patent codes. This study uses the content of a given patent and tries to extract information. It is hard to process pure text documents in data science. This study uses some NLP and text mining techniques to evaluate documents and classify them. The results show that K-means clustering algorithm with the cosine similarity metric give the optimum results.

REFERENCES

Books

Li Y., He C., Fan X., Huang X., Cai Y., 2015. *HCloud, a healthcare-oriented cloud system with improved efficiency in biomedical data processing*. Cloud Computing with E-science Applications. Vol. 1, England: CRC Press/Taylor&Francis.

Porter A. L., Cunningham S. W., 2004. *Tech mining: multiple ways to exploit science, technology & information resources*. Vol. 1, Canada: Wiley

Periodicals

Chen C., Zhang J., & Vogeley M. S., 2009. *Visual analysis of scientific discoveries and knowledge diffusion*, In The 12th international conference on scientometrics and informetrics. 14–17

Fall C. J., Törcsvari A., Benzineb K., 2003. Karetka G., *Automated Categorization in the International Patent Classification*. ACM SIGIR Forum, (Vol. 37), 10-25

Gao L., Porter A. L., Wang J., Fang S., Zhang X. Ma T., 2013. *Technology life cycle analysis method based on patent documents*. Technological Forecasting & Social Change. (80), 398–407

Huang J., 2016. *Patent Portfolio analysis of the cloud computing industry*. Elsevier. (Vol. 39). 45–64

Jie G., Mengyang S., Zhaofeng Z., Xiaoping L., 2010. *IPC Co-occurrence based Technological Trends Discovery*. 3rd International Conference on Information Management, Innovation Management and Industrial Engineering

Jivani A. G., 2011. *A Comparative Study of Stemming Algorithms.* IJCTA, Vol2(6), 1930- 1938

Jun S., 2012, *A Clustering Method of Highly Dimensional Patent Data Using Bayesian Approach*. IJCSI International Journal of Computer Science Issues, (Vol. 9), 7-11

Kakimoto K., 2003. *Intellectual Property Cooperation Center*, personal communication

Kasravi K., Risov M., 2007. *Discovery of Business Value from Patent Repositories.* Proceedings of the 40th Hawaii International Conference on System Sciences.

Larkey L. S., 1998. *Some Issues in the Automatic Classification of U.S. patents*, Working Notes for the Workshop on Learning for Text Categorization, 15th Nat. Conf. on Artif. Intell. (98)

Lin F., Wei C., Lin Y., Shyu Y., 2008. *Deriving Technology Roadmaps with Tech Mining Techniques*. PACIS 2008 Proceedings. (Paper 255)

LIU L., KANG J., YU J., WANG Z., 2005. *A Comparative Study on Unsupervised Feature Selection Methods for Text Clustering*. IEEE, 597-601

Qu P., Zhang J., He J., Zeng W., Xu H., 2014. *Term Extraction Using Co-occurrence in Abstract and First Claim for Patent Analysis*. International Conference on Identification, Information and Knowledge in the Internet of Things

Rotoloa D., Hicksb D., Martina B.R., 2015. *What is an emerging technology*. Research Policy, (44), 1827–1843

Trumbach C. C., Payne P., Kongthon A., 2006. *Tech mining for small firms: Knowledge prospecting for competitive advantage*. Technological Forecasting & Social Change. (73), 937–949

Other Sources

Bloomberg Data, *Car Companies and Their Countries*, 2015, Strategy Business, http://www.strategy-business.com/feature/00370?gko=e606a

Chiou F.C., 2010. *Technology Life of Cloud Computing Using Literature Analysis*, Master Dissertation. National Dong-hwa University, Hualien, Taiwan.

Moore G., 1975, *Moore Law*, Wikipedia, https://en.wikipedia.org/wiki/Moore%27s_law (2016)

Newbies D., 2004. *Maps and Science*, http://www.davosnewbies.com/posts/davos-newbies-home-514/