Phesto Enock Mwakyusa

Semantic Annotation and Big Data Techniques for Patent Information Processing

Master's Thesis in Information Technology

October 10, 2017

University of Jyväskylä

Department of Mathematical Information Technology

Author: Phesto Enock Mwakyusa

Contact information: phesto@qusaz.com

Supervisors: Michael Cochez, and Vagan Terziyan

Title: Semantic Annotation and Big Data Techniques for Patent Information Processing

Työn nimi: Semanttinen annotaatio ja Big Data menetelmiä patentti-informaation prosessointiin

Project: Master's Thesis

Study line: Mobile Technology and Business (MOTEBU)

Page count: 73+0

Abstract: This thesis analyzes approaches to generate semantic annotations on patent records, as well as on other structured data, by relying on the structure and semantic representation of documents. Information in patent records reflects how real-world technologies evolve, and the approximately 3 million annual new patent applications capture the global inventive frontier. The volume of this information is too big to be effectively analyzed purely with human effort, necessitating Big data approaches to analyze it with computer aided tools and techniques. Big data is a term that describes a massive volume of structured, semi structured and unstructured data that is so large to the point that it is difficult to process using traditional database and software tools and techniques. Currently, technical information, such as patents, is typically stored in data repositories that do not support advanced Big data methods to structure and interpret documents. In the emerging Semantic technology, annotation, Web search, as well as interpretation and aggregation can be addressed by ontology-based semantic annotation. This thesis examines semantic annotation and other Big data methodologies, and their basic requirements, and reviews the current generation of semantic annotation and other Big data systems. As a use case, this thesis demonstrates how semantic annotation and other Big data techniques are employed to enhance the human processes whereby people retrieve information, carry out analysis or discovery within a large collection of patent information.

Keywords: Big Data, Semantic Annotation, Patent information, Data Mining

Suomenkielinen tiivistelmä: Tämä tutkielma analysoi miten luoda semanttisia annotaatioita patenttietueisiin, tai muuhun ei-strukturoituun dataa, hyödyntämällä tietueiden rakennetta tai semanttista representaatiota. Patenttitietueet sisältävät kokonaisuutena informaation siitä, miten reaalimaailman teknologiat kehittyvät ja muuttuvat, ja vuosittain globaalisti julkaistavat noin 3 miljoonaa uutta patenttihakemusta kuvaavat hyvin globaalin keksintörintaman kehitystä. Tämä informaatio on volyymiltaan liian laaja, jotta sitä voisi tehokkasti analysoida ja käsitellä puhtaasti ihmisvoimin. Tästä syystä sen analysointiin tarvitaan erityisiä Big data lähestymistapoja, jotka hyödyntävät tietokoneavusteisia työkaluja ja -prosesseja. Big data on termi joka kuvaa erittäin suurta volyymia strukturoitua, osittain strukturoitua tai strukturoimatonta dataa, joka on niin suuri että sen prosessointi perinteisin tietokanta- tai ohjelmistoteknisin työkaluin tai tekniikoin on vaivalloista. Nykyisin tekninen informaatio, kuten patentit, säilytetään datakokoelmissa, jotka eivät tue edistyneitä Big data menetelmiä strukturoida ja tulkita dokumentteja. Nousevassa Semanttisessa teknologiassa annotaatio, web-haku, sekä tulkinta ja koostaminen käsitellään ontologia-pohjaisella semanttisella annotaatiolla. Tämä tutkielma käsittelee semanttista annotaatiota ja muita Big data menetelmiä ja niiden perusedellytyksiä, sekä tarkastelee nykyaikaisia semanttisen annotaation ja muiden Big data menetelmien järjestelmiä. Tapaustutkimuksena tämä tutkielma osoittaa, miten semanttista annotaatiota ja muita Big data tekniikoita voidaan hyödyntää parantamaan prosesseja, joiden avulla ihmiset hakevat tietoa, tekevät analyysiä tai hakuja erittäin suuresta patentti-informaation kokoelmasta.

Avainsanat: Big Data, Semanttinen annotaatio, Patentti-informaatio, Tiedonlouhinta

Preface

Writing this thesis has been a wonderful adventure, filled with joy and tears, ups and downs, but here we are, at last. Wonderful people have played a great role into supporting my efforts to finish this work, it was not an easy task but, they made it possible. The first line of text in this thesis was written in 2014, being an entrepreneur, my work took precedence and the writing halted for two years before I resumed my writing in the beginning of 2016. In February, I encountered a very traumatizing event that made me pause my research again until July 2017, this has been such an experience.

I would like to thank God for His grace on my health mentally and physically, His protection to me and my family during the time of studies as well as writing this thesis. I was saved from a deadly tragedy in 2016 February in the ways that I cannot comprehend.

My high regarded appreciation goes to my supervisors Michael Cochez and Vagan Terziyan, they have been such instrumental mentors in guiding me thorough out my 4 years of writing this thesis. Despite the time it took for me to finish, they never left me alone. I would like to personally convey my special thanks to Michael Cochez, he has been working tirelessly with me even in odd hours and weekends, providing me with all the support and guidance I could ever need during this process, it has indeed been a pleasure being under their supervision.

I would love to express my appreciations to my lovely family, which has been a great support and encouragement to me during the difficult moments, keeping my spirit high and supporting me the way they could, my lovely wife Zelda, you are one of the kind. I love you.

My sisters, Jenny, Leah, Neema and Zena and my mother "The iron lady FROIDAH" their prayers and encouragement during all the challenging moments in life parallel to doing this work.

Special thanks to my sons Patrick (Lutengano), Presley(Lughano) and Powell(Lusekelo) for aew.sfkh.earwi /vyiewl.g54b78236576alksdfasiwjea akjshfdoaul akjs hfdlsakh, giving extra work to proof read and find their inputs on my work;

My appreciation to the management of TEQMINE Analytics for their support in providing

me with resources to process data and experiments for this thesis. They have done a tremendous job keeping me busy being an entrepreneur but at the same time, they managed to show me that it takes what it takes to be successful. My CEO Hannes Toivannen has been an instrumental part of supporting me in many ways including advising and helping me get time to finish this work.

Jyväskylä, October 10, 2017

Phesto Enock Mwakyusa

Glossary

Annotation
Meta-data added to a specific span of text

Big Data
Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them

Computer Science
Computer science is the study of how to manipulate, manage, transform and encode information.

Corpus
A collection of written texts, especially the entire works of a particular author or a body of writing on a particular subject.

Domain
Domain is the subject matter of the document in question. The concept of domain helps us identify context for the content. Example, if the document is in the "domain" of computer science, the word "root" is more likely to be relevant to a file system than a tree. Semantic annotation takes these relationships into account when applying annotations. We often use the terms, "domain knowledge" or "domain expert" to describe annotators or curators who really "understand" the content of a given document

Entity
Something that has a distinct, separate existence independent of the text. It can be either implicit (not mentioned, but its existence may be inferred from the text), or explicit (mentioned directly in the text). See also *Semantic Annotation*

EPO
"The European Patent Office (EPO) offers inventors a uniform application procedure which enables them to seek patent protection in up to 40 European countries. Supervised by the Administrative Council, the Office is the executive arm of the European Patent Organization." (EPO 2016)

Gensim
is a free Python library designed to automatically extract semantic topics from documents, as efficiently (computer-wise) and painlessly (human-wise) as possible. Gensim is designed

	to process raw, unstructured digital texts (“plain text”).
Innovation	Innovation, for its part, can refer to something new or to a change made to an existing product, idea, or field
IOT	Internet of things
KR	Knowledge Representation
meta-data	Meta-data is "data [information] that provides information about other data." Two types of meta-data exist: structural meta-data and descriptive meta-data.
NLP	Natural Language Processing
OCR	Optical Character Recognition
Ontology	Ontology is the science of things existing, or things existing permanently
OWL	(Web Ontology Language) The schema language, or knowledge representation (KR) language, of the Semantic Web.
patent	A patent is an exclusive right given by law to inventors to make use of, and exploit, their inventions for a limited period of time. By granting the inventor a temporary monopoly in exchange for a full description of how to perform the invention, patents play a key role in developing industry around the world.
PCT	"The Patent Cooperation Treaty (PCT) assists applicants in seeking patent protection internationally for their inventions, helps patent Offices with their patent granting decisions, and facilitates public access to a wealth of technical information relating to those inventions. By filing one international patent application under the PCT, applicants can simultaneously seek protection for an invention in a very large number of countries." (PCT 2016)
RDF	(Resource Description Framework) The data modeling language for the Semantic Web. All Semantic Web information is stored and represented in the RDF.
Semantic Annotation	Semantic annotation connects a word or span of text to a se-

mantic database or ontology where additional information is stored. Semantic annotations transforms the target text into an entity, which is a specific data element in a universe of data elements. Semantic annotation also provides an anchor point in a text document for examples or such entities. Semantically annotated documents, therefore, can be connected to a wealth of searchable information useful i many information management contexts. in semantic annotation, the traditional annotation type and features may be replaced by an address in form of URI(Universal Resource Indicator).

SPARQL (SPARQL Protocol and RDF Query Language): The query language of the Semantic Web. It is specifically designed to query data across various systems.

USPTO "The United States Patent and Trademark Office is an agency in the U.S. Department of Commerce that issues patents to inventors and businesses for their inventions, and trademark registration for product and intellectual property identification." (USPTO 2016b)

VSM Vector Sparse Model

XML XML Extensible Markup Language is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable.

List of Figures

Contents

1 Introduction

Patent information processing is of high importance for a variety of reasons within academia, business, law, and government, as well as beyond these cases. Inventors must search prior-art in order to secure they are not reinventing the wheel. Companies must explore that they are not commercializing products that infringe on the rights of other patent holders. Information retrieval from patent information has also an important role for product innovation design and development. And so forth. Effective and accurate methods to process patent information is fundamental for all of these user scenarios, for which the large number of existing patent records and the rapid growth of patent information poses a serious challenge.

The current approaches to patent information processing lack the semantic association and comprehension to a large degree, making it difficult to capture the implicitly useful knowledge at a semantic level. In order to improve traditional patent search methods, this thesis analyzes approaches to generate semantic annotations on patent records, as well as on other structured data, by relying on the structure and semantic representation of documents. To this end, this thesis demonstrates the use of Latent Dirichlet Allocation (LDA) to semantically analyze a very large collection of patent records, and how to use it to construct an improved patent information processing service with the objective to find similar, relevant patents.

The selected approach utilizes template schemes to extract the structure information from patent documents. It then identifies semantics of entities and relations between entities from the content based on natural language processing techniques and domain knowledge. Finally, it employs a heuristic pattern learning method to abstract patent technical features.

As a use case, this thesis demonstrates how semantic annotation and other Big data techniques are employed to enhance the human processes whereby people retrieve information, carry out analysis or discovery within a large collection of patent information.

The results are discussed in the context of Semantic annotation and other Big data methods to process patent information. Big data is a term that describes a massive volume of structured, semi structured and unstructured data that is so large to the point that it is difficult to process using traditional database and software tools and techniques.

2 Patents

According to Intellectual Property Law (2010) a patent is a legal document that is granted by an authorized government entity giving the recipient known as patentee, a set of specific exclusive rights, patent rights. These are rights to exclude others from making, using, offering, producing or selling the invention throughout the region where the granted protection right of the patent is valid.

Clarivate Analytics (2017)A patent is granted to inventions for a limited period of time. By granting the inventor a temporary monopoly in exchange for a full description of how to perform the invention, patents play a key role in developing industry around the world. Once the owner of an invention has been granted a patent in any particular country, they then have the legal authority to exclude others from making, using, or selling the claimed invention in that country without their consent, for a fixed period of time. In this way, inventors can prevent others from benefiting from their ingenuity and, ultimately, sharing in profits from the invention, without their permission. In return for these ownership rights, the applicant must make public the complete details of the patented invention. These include:

- Background information (the 'state of the art')
- The nature of any technical problems solved by the invention
- Description of the invention and how it works
- Illustrations of the invention where appropriate

Patent protection in a given country does not extend to other countries -inventors must file an application in each territory where they want their patent to be effective. To maintain the validity of a patent, the owner needs to pay fees to each appropriate patent authority; failure to do so causes the patent rights to lapse. Most countries also require that the patent is "worked." This means that the protected invention is put to commercial use, within a specified period of time.

A patent does not give a right to make or use or sell an invention, rather it provides a legal stand point, the right to exclude others from using, selling, making, offering for sale, or importing the patented invention for the validity period of the patent, usually 20 years from the filling date. A patent is a limited property right the government gives inventors in exchange for their agreement to share details of their invention with the public. So like any other property right, it may be licensed, sold, assigned or transferred, given away or just abandoned. The patent owner may give permission to, or license, other parties to use the invention on mutually agreed terms. The owner may also sell the right to the invention to someone else, who will then become the new owner of the patent. Once a patent expires, the protection ends, and an invention enters the public domain; that is, anyone can commercially exploit the invention without infringing the patent.

2.1 Types of patents

Generally, there are three main types of Patents issued by patent offices or patent region offices worldwide. USPTO (2017b)

- Utility Patent (See subsection 2.1.1)
- Design Patent (See subsection 2.1.2)
- Plant Patent (See subsection 2.1.3)

2.1.1 Utility patents

Utility patents may be granted to anyone who invents or discovers any new and useful process, machine, article of manufacture, or composition of matter, or any new and useful improvement thereof; Utility patents are grouped in five categories: a process, a machine, a manufacture, a composition of matter, or an improvement of an existing idea. Often, an invention will fall into more than one of the categories. For instance, computer software can usually be described both as a process (the steps that it takes to make the computer do something) and as a machine (a device that takes information from an input device and moves it to an output device). Regardless of the number of categories in which an invention falls, only one utility patent may be issued on it. Among the many types of creative works that might qualify for a utility patent are biological inventions, new chemical formulas, processes, or procedures; computer hardware and peripherals; computer software; cosmetics; electrical inventions; electronic circuits; food inventions; housewares; machines; and magic tricks. If you acquire a utility patent, you can stop others from making, using, selling and importing the invention. A utility patent last for 20 years from the date that the patent application is filed.

2.1.2 Design patents

Design patents may be granted to anyone who invents a new, original, and ornamental design for an article of manufacture; A design patent is granted for product designs—for example, an IKEA chair, Keith Haring wallpaper, or a Manolo Blahnik shoe. You can even get a design patent for a computer screen icon. There are strings attached to a design patent, too. As noted, the design must be ornamental or aesthetic; it can't be functional. Once you acquire a design patent, you can stop others from making, using, selling and importing the design. You can enforce your design patent for only 14 years after it's issued.

2.1.3 Plant patents

Plant patents may be granted to anyone who invents or discovers and asexually reproduces any distinct and new variety of plant. Asexual reproduction is the propagation of a plant to multiply the plant without the use of genetic seeds to assure an exact genetic copy of the plant being reproduced. Any known method of asexual reproduction which renders a true genetic copy of the plant may be employed. This may include cultivating different types of plants to create mutants or hybrids and also newly found seedlings. This patent protects the owner by keeping other individuals or businesses from creating the type of plant or profiting from the plant for at least 20 years from the date of the application.

2.2 World wide Patenting statistics

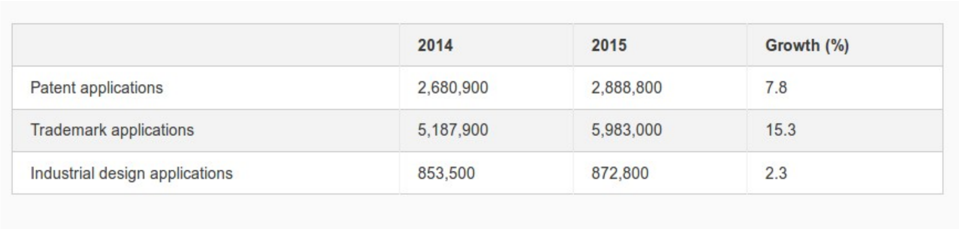
Worldwide filings of patent applications have grown at a substantial rate i.e. from 12,601,187 applications during the period 1995-2005 to 15,206,132 applications in the subsequent period between 2005-2015, which is nearly an increase of 2.6 million applications.

According to WIPO (2015), Around 2.68 million patent applications were filed worldwide in

2014, up 4.5% from 2013 1. Driving that strong growth were filings in China, which received 103,000 of the 116,100 additional filings and accounted for 89% of total growth, whereas the United States of America (US) contributed 6% of total growth. The 4.5% growth in filings in 2014 is lower than the growth rate in each of the previous four years, which varied between 7% and 10%. period.

The figure 1 shows the total number of new patent applications filed annually across 102+ patent offices in the last 20 years.

There is substantial rise in the number of new applications filed in the last three years. The number of patent applications filed in 2013-2014 totaled 4.4 million. This represents roughly 6.2% rise in the applications filed in the previous period between 2011-2012. The long term trends shows continuous growth in the number of applications filed, with the exception of slight decrease during 2007-2008. During the last 20 years total number of applications has tripled from where they were in 1996. (Technologies 2016)



	2014	2015	Growth (%)
Patent applications	2,680,900	2,888,800	7.8
Trademark applications	5,187,900	5,983,000	15.3
Industrial design applications	853,500	872,800	2.3

Figure 1. Total number of new patent applications filed in the period of year 2014 and 2015

As shown in the Figure 1 it is evident that there is a huge number of patents world wide, and this means a rich information pool of technology and discovery development around the world. The rate of patent number growth is not proportional with the technology and infrastructure that should allow the data to be accessible and useful to both researchers in information retrieval and other areas of computer science as well as professionals seeking to broaden their knowledge of patent search. According to WIPO (2016b) the report indicates that innovators filed some 2.9 million patent applications worldwide in 2015, up 7.8% from 2014, higher than the 4.5% growth rate in 2014. Also resident filings, where innovators filed for protection in their home economy, accounted for around two-thirds of the 2015 total.

The report continues to indicate that China's patent office received 1,101,864 filings in 2015, making it the first office to receive the filing of more than a million applications in a single year – including both filings from residents in China as well as from other countries innovators seeking patent protection inside China. This totaled almost as many applications as the next three offices combined: the U.S. (589,410), Japan (318,721) and the Republic of Korea (213,694).

Intellectual property rights are designed to encourage economic growth. Economists do not agree if they actually do so or if they are in fact harmful for economic growth. However, the fundamental justification for the regulation of ownership of technical ideas is to benefit the public. This reasoning proposes that the government must provide incentives for inventors to invest in technical ideas and their development, and that such incentives must combat the problem of copying: If free copying of technical ideas would be allowed, people would hesitate to invest in developing ideas, and the society at large would miss the benefits of technological progress. Landes and Posner (2003)

Patents and the ownership of technological ideas poses several ethical, economic, social and legal problems. One fundamental, global, one is that economies at different stage of development benefit asymmetrically from intellectual property rights. At very simple level, the advanced economies, such as Finland, the United States, Japan, and others, who have enjoyed long history of technology driven economic development, stand to benefit from global imposition of strong intellectual property rights, because they could extract rents from less advanced countries using or buying products developed in advanced economies. On the other hand, less developed countries, like many African or South American countries, could accelerate their economic development if they could use advanced technologies without paying patent fees. Mario Cimoli and Primi (2009)

The most known example of this problem is the decision of large developing countries, such as Brazil, to break the patent protection of HIV/AIDS drugs in order to provide affordable care for inflicted people. The US and European patent holders to these life saving drugs demanded unaffordable prices, causing several thousand people to die in lack of medication. Only the government decision to break patent monopolies provided broad access to critical medicines. Mario Cimoli and Primi (2009)

2.3 Patent information

Patent information refers to the information found in patent applications and granted patents. This information may include bibliographic data about the inventor and patent applicant of patent holder, a description of the claimed invention and related developments in the field of technology, and a list of claims indicating the scope of patent protection sought by the applicant. The requirement that a patent applicant disclose information about their inventions is very important for te continuous development of the technology. This information provides a basis on which new technical solutions can be developed by other inventors.

Patent documents contain technological information that is often not divulged in any other form of publication, covering practically every field of technology. They have a relative standardized format and are classified according to technical fields to make identifying relevant documents easier. A large percentage of information that is found in patents is not published anywhere else, this makes patents to be one of the unique sources for discovering new technology information. Currently there are more than 35 million patents worldwide, and every year there is an average of one million new patent applications filed. "Moreover, the patent document provides much more detailed information about a technology than any other type of scientific or technical publication. And it is estimated that more than 70 percent of the information disclosed in patents is never published anywhere else." (WIPO 2017)

Unique insight into industry developments

In order to secure rights to an invention, the inventor must keep the details secret prior to filing the patent application. So publication of a patent is often the first time that an invention has ever been disclosed. Monitoring the vital information contained within published patent documents is a great way to stay on top of key industry developments. (Clarivate Analytics 2017)

Extensive References to Similar Inventions

Many patent documents include search reports prepared by patent examiners. These reports may cite or reference patents and other literature related to the subject matter of the invention. This supplementary information can provide valuable background information on the development of that particular technology, saving you time in researching that topic. (Clarivate Analytics 2017)

Detailed Descriptions of the Invention

In order to obtain a granted patent, the technical details of the invention must be fully disclosed in the text and drawings of the patent application. The detail must be sufficient to enable an expert specializing in the same field to re-create the invention. By browsing through these full and practical descriptions, you may discover details that prompt new groundbreaking ideas of your own. (Clarivate Analytics 2017)

The information contained in patent document can be very useful to researchers, entrepreneurs, and many others, helping to:

1. Avoid duplication of research and development work
2. Build on and improve existing products or processes
3. Assess the state-of-the-art in a specific technological field, e.g. to get an idea of the latest developments in this field.
4. Evaluate the patentability of inventions, in particular the novelty and inventiveness of inventions (important criteria for determining their patentability), with a vew to applying for patent protection domestically or abroad
5. Identify inventions protected by patents, in particular to avoid infringement and seek opportunities for licensing.
6. Monitor activities of potential partners and competitors both within the country and abroad.
7. Identify market niches or discover new trends in technology or product development at an early stage.

Patent documents are published by national and regional patent offices, usually 18 months after the date on which a patent application was first filed or once a patent has been granted for the invention claimed by the patent applicant. Some patents offices publish patent documents through free-of-charge online databases, making the information easily accessible by public.

WIPO’s PATENTSCOPE database provides free of charge online access to millions of international patent applications filed under Patent Cooperation Treaty (PCT) System as well as patent document filed at national and regional patent offices such as the European Patent Office and the United States Patent and Trademark Office.

Though accessibility of patent information has grown as more and more patent offices make their patent document available through online databases, certain skills are still required in order to make effective use of this information, including carrying out targeted patent searches and providing meaningful analysis of patent search results. As a result, it may be advisable to contact a patent information professional for assistance where business-critical decisions are at stake. WIPO Patent Information Services (WIPIS) provide free-of-charge patent search services for individuals and institutions in developing countries.

2.4 Benefits of patents

Patented inventions have, in fact, pervaded every aspect of human life, from electric lighting (patents held by Edison and Swan) and plastic (patents held by Baekeland), to ballpoint pens (patents held by Biro), and microprocessors (patents held by Intel, for example). Patents provide incentives to and protection for individuals by offering them recognition for their creativity and the possibility of material reward for their inventions. At the same time, the obligatory publication of patents and patent applications facilitates the mutually-beneficial spread of new knowledge and accelerates innovation activities by, for example, avoiding the necessity to “re-invent the wheel”.

Once knowledge is publicly available, by its nature, it can be used simultaneously by an unlimited number of persons. While this is, without doubt, perfectly acceptable for public information, it causes a dilemma for the commercialization of technical knowledge. In the absence of protection of such knowledge, “free-riders” could easily use technical knowledge embedded in inventions without any recognition of the creativity of the inventor or contribution to the investments made by the inventor. As a consequence, inventors would naturally be discouraged to bring new inventions to the market, and tend to keep their commercially valuable inventions secret. A patent system intends to correct such under-provision of innovative activities by providing innovators with limited exclusive rights, thereby giving the innovators the possibility to receive appropriate returns on their innovative activities. In a wider sense, the public disclosure of the technical knowledge in the patent, and the exclusive right granted by the patent, provide incentives for competitors to search for alternative solutions and to "invent around" the first invention. These incentives and the dissemination of knowledge about new inventions encourage further innovation, which assures that the quality of human life and the well-being of society is continuously enhanced. (WIPO 2016a)

There are many ways in which an inventor might be compensated for a patent. An inventor might bring the patented product to market under the protection of the monopoly created by the patent. The inventor may license a patent to another entity for an up front fee, an ongoing royalty or other consideration. The inventor may also sell the patent outright. This core incentive to inventors, is a main factor fueling the efforts of them to continue bringing more revolutionary inventions into the technology pool, because there is a reward into their hard and valuable work. But also the patent system enables the world to have access to inventions from all over the world and have a opportunity to invent around or over the current inventions, and minimize the redundancy of same thing that might have been done in the other part of the world.

For a technology based enterprise, they are more likely to be developing new products, services or processes, they invent and innovate. But before committing resources to expensive development work, it is important to check whether anyone else has invented or worked on the same idea. If it happens to the fact, it is not necessarily the end of the road, but it may prevent the company from filling a patent, and they cannot copy someone else's patented invention without consent of the patent owner. By having this information beforehand, it will save the company the cost and time to invent and file a patent only to find out that similar invention has been filed already. (WIPO 2016c)

Reasons for patenting the inventions

- Exclusive rights - Patents provide the exclusive rights which usually allow a inventor to use and exploit the invention for twenty years from the date of filing of the patent application.

- Strong market position - Through these exclusive rights, the patent owner is able to prevent others from commercially using your patented invention, thereby reducing competition and establishing the companies position in the market as the pre-eminent player.
- Higher returns on investments - Having invested a considerable amount of money and time in developing innovative products, a company could, under the umbrella of these exclusive rights, commercialize the invention enabling itself to obtain higher returns on investments.
- Opportunity to license or sell the invention - If te patent owner chose not to exploit the patent, it may sell it or license the rights to commercialize it to another enterprise which will be a source of income for the company.
- Increase in negotiating power - If the company is in the process of acquiring the rights to use the patents of another enterprise, through a licensing contract, its patent portfolio will enhance the bargaining power. That is to say, its patents may prove to be of considerable interest to the enterprise with whom the company is negotiating and it could enter into a cross licensing arrangement where, simply put, the patent rights could be exchanged between your enterprise and the other.
- Positive image for the enterprise - Business partners, investors and shareholders may perceive patent portfolios as a demonstration of the high level of expertise, specialization and technological capacity within your company. This may prove useful for raising funds, finding business partners and raising company's market value.

2.5 Disadvantages of Patents

The idea of protection of invention to inventors and limiting others from freely using the design or processes invented by others, is good and beneficial to the inventors, but it comes with some disadvantages. There are cases where some great inventions that could be very useful for human kind, either by improving their life or providing a much needed service to the society can be prevented from being implemented just because the patent owner decides to shelve it. Many big companies shelves patents just because they do not have a good business opportunity in terms of profits, regardless of their impact to the society.

This same idea is debated to hinder the development of innovation, if inventions ware free like Open source software, anybody would innovate further the inventions without limitations, and by being able to do that, inventions would be fast innovated, tested and made better. With patents, others are not allowed to use the invention, test and build on top of that more advanced ideas.

The fact that patents are valid for 20 years and 14 years for design patents, that is the exact time that might keep the invention stall with no innovation from others without the permission of the patent holder. The licensing of patents is not cheap, this means innovators with no capital or capability to pay for licensing a patent cannot innovate on the already patented invention.

It costs time and money to apply and maintain a patent. Before applying for a patent it has to be researched to ensure there are no existing patents of a similar nature – this discovery process involves legal fees Not possible to guarantee that once a patent is valid and granted, it is the end of it. The patent can still be legally challenged and revoked with no refunds It is still up to the inventor to protect a patent if an infringement has been discovered – the patent office does not take sides. Also a granted patent does not mean that the invention has merits of commercial value. Some product processes can be slightly changed or modified around a patented invention to get around the wording of patents.

2.6 Applying for Patent

A patent is requested by filing a written application at the relevant patent office. The person or company filing the application is referred to as "the applicant." The applicant may be the inventor or its assignee. The application contains a description of how to make and use the invention that must provide sufficient detail for a person skilled in the art (i.e., the relevant area of technology) to make and use the invention. In some countries there are requirements for providing specific information such as the usefulness of the invention, the best mode of performing the invention known to the inventor, or the technical problem or problems solved by the invention. Drawings illustrating the invention may also be provided. The application also must include one or more claims that define what a patent covers or the "scope of protection." After filing, an application is often referred to as "patent pending." While this term does not confer legal protection, and a patent cannot be enforced until granted, it serves to provide warning to potential infringer's that if the patent is issued, they may be liable for damages.

Once filed, a patent application is examined. A patent examiner reviews the patent application to determine if it meets the patentability requirements of that country. If the application does not comply, objections are communicated to the applicant or their patent agent or attorney, to which the applicant may respond. The number of Office actions and responses that may occur vary from country to country, but eventually a final rejection is sent by the patent office, or the patent application is granted, which after the payment of additional fees, leads to an issued, enforceable patent. In some jurisdictions, there are opportunities for third parties to bring an opposition proceeding between grant and issuance, or post-issuance. Once granted the patent is subject in most countries to renewal fees to keep the patent in force. These fees are generally payable on a yearly basis. Some countries or regional patent offices (e.g. the European Patent Office) also require annual renewal fees to be paid for a patent application before it is granted.

A patent is granted by a national patent office or by a regional office that carries out the task for a number of countries. Currently, the following regional patent offices are in operation, according to (WIPO 2016)

- African Intellectual Property Organization (OAPI)
- African Regional Intellectual Property Organization (ARIPO)
- Eurasian Patent Organization (EAPO)
- European Patent Office (EPO)
- Patent office of the Cooperation Council for the Arab States of the Gulf (GCC Patent Office)
- Nordic Patent Institute (NPI)

Under such regional systems, and applicant requests protection for an invention in one or more member states of the regional organization in question. The regional office accepts these patents applications, which have the same effect as national applications, or grants patents, if all the criteria for the grants of such a regional patent are met.

There are a number of conditions that must be met in order to obtain a patent and it is not possible to compile an exhaustive, universally applicable list. However, some of the key conditions include the following:

- The invention must show an element of novelty; that is, some new characteristic which is not known in the body of existing knowledge in its technical field. This body of existing knowledge is called “prior art”.
- The invention must involve an “inventive step” or “non-obvious”, which means that it could not be obviously deduced by a person having ordinary skill in the relevant technical field.
- The invention must be capable of industrial application, meaning that it must be capable of being used for an industrial or business purpose beyond a mere theoretical phenomenon, or be useful.
- Its subject matter must be accepted as “patentable” under law. In many countries, scientific theories, aesthetic creations, mathematical methods, plant or animal varieties, discoveries of natural substances, commercial methods, methods for medical treatment (as opposed to medical products) or computer programs are generally not patentable.
- The invention must be disclosed in an application in a manner sufficiently clear and complete to enable it to be replicated by a person with an ordinary level of skill in the relevant technical field. (WIPO 2016a)

2.7 Patent application costs

The costs of preparing and filing a patent application, examining it until grant and maintaining the patent vary from one jurisdiction to another, and may also be dependent upon the type and complexity of the invention, and on the type of patent. The details in 2.7.1 are based on the US patent office only for the year 2016, other regions and countries might have their own pricing tariffs.

2.7.1 Application fees

For utility patents (see 2.1.1), the small entity fees include a $165 filing fee ($82 if filing electronically), as well as a search fee of $270 and an examination fee of $110. For large entities, the filing fees are a $330 filing fee, a search fee of $540, and an examination fee of $220. In addition, both small and large entities must pay more fees for claims in excess of 20 and for multiple dependent claims. (NOLO 2016)

For design patents 2.1.2, the small entity fees include a $110 filing fee, as well as a search fee of $50 and an examination fee of $70. For large entities, the filing fees are a $220 filing fee, a search fee of $100, and an examination fee of $140. In addition, both small and large entities must pay more fees for a design patent application that exceeds 100 pages. (NOLO 2016)

For plant patents 2.1.3, the small entity fees include a $110 filing fee, as well as a search fee of $165 and an examination fee of $85. For large entities, the filing fees are a $220 filing fee, a search fee of $330, and an examination fee of $170. (NOLO 2016)

2.7.2 Maintenance fees

Using USPTO as an example, fees must be paid to the U.S. Patent and Trademark Office (USPTO) (or the patent office of another country where a patent has been obtained) to keep an issued patent in effect. As of September 2008, the maintenance fees for U.S. utility patents (there are no maintenance fees for design or plant patents) are as follows:

- due at 3.5 years, $980 for large entities and $490 for small entities
- due at 7.5 years, $2,480 for large entities and $1,240 for small entities, and
- due at 11.5 years, $4,110 for large entities and $2,055 for small entities.

Effective with applications filed after June 7, 1995, the patent term changed from 17 years from the date of issue to 20 years from the date of filing. This means that the final maintenance fee may extend beyond the 17th year until the patent term actually expires. (NOLO 2016)

2.8 Big Data

Big data is a broad term, or catch-phrase, used to describe a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques

According to SAS Institute Inc (2015) while the term “big data” is relatively new, the act of gathering and storing large amounts of information for eventual analysis is ages old. The concept gained momentum in the early 2000s when industry analyst Doug Laney articulated the now-mainstream definition of big data as the three Vs:

Volume Organizations collect data from a variety of sources, including business transactions, social media and information from sensor or machine-to-machine data. In the past, storing it would’ve been a problem – but new technologies (such as Hadoop) have eased the burden.

Velocity Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time.

Variety Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.

2.9 Patents as an example of Big Data

In essence a patent is a combination of data and one or several processes, each patent illustrates a process that uses certain data, resulting in a qualification of being useful and new invention, according to patent examining office. In addition, a patent is a process, however a patent is a particular process, in that process itself it is kept locked in a frozen state, it is a snapshot of technological invention (process and data) and freezes in time. So only that particular snapshot of the process is protected by the patent’s exclusive rights. In the true meaning of big data, these frozen processes are least innovative, meaning they do not evolve to make changes every time, they are kept constant in time.

Patents documents contains abstract of invention, invention claims and full text invention description, the length of each patent full text description varies from one patent to another, depending on the nature of the invention as well as its complexity. Some patents have more graphs, images and illustrations than text, some have chemical formulas, while other have long mathematical calculations. Some patent documents have long text than average books, for example a US patent No US2003173072A1 Espacenet (2017) has 1025 pages of full text description, having a few millions of these documents qualifies patents as big data. The enormous data that is contained in the patent documents and their uniqueness, requires a sophistication methods and techniques of handling the data, structuring and information processing and extraction from them.

For human to process this kind of volume of data from patents is tedious, less efficiency and the results are average in quality. Often high percentage of the quality of information relevance obtained from processing patent document in traditional methods is low. Patents are filled with rich technological information from complex mathematical and chemical formulas to cutting edge product design specification that needs high accuracy means of swifting through the jungle of content within patents.

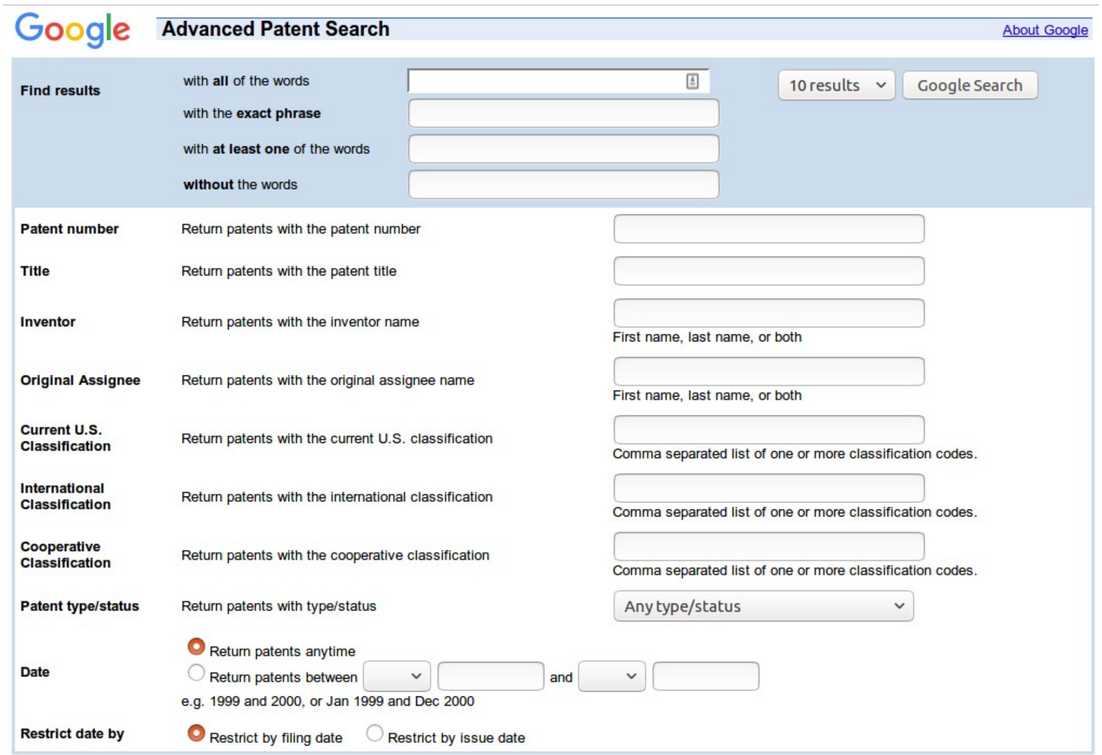
3 Current state of patent search

The current search methodology that is vastly used for patent search purposes by many patent information seekers is by the use of traditional keyword search. While the advanced part of the searching is just the additional of extra attributes and meta information of patents. Patent classification, phrases, inventor names and keyword exclusion are some of the fields that maybe applied to advance the complexity of searching for information from patents, or patents themselves.

The figure 2 shows a screen shot of Google's advanced patent search interface, from the illustration it is show that key fields used in the advanced search of patents are Patent number, Title, Inventor, Original assignee, classification, patent status or type, date of publication, just to mention a few. All these fields require that a searcher does have prior knowledge of what to search, the searching person should have a knowledge of classification, inventor name and if possible publication dates or a patent number. This is a reason why average people find it impossible to conduct a productive and informative search using these methods without having an intimate knowledge of patents. For this reason, these tools benefits a limited number of patent search experts who have a good experience and knowledge of patents, and can figure out ways to go about searching patents form the patent databases.

The major limitation of these methods and techniques is that, they inhibits the ability to do knowledge and information discovery without known parameters. If the patents would have a mens to search within its content contextually, users would not have to have a searching expertise to get what they want from the patent documents. One could just provide with a plain text description of knowledge or product description, and get back relevant similarity with patents with potentially high relevance including many that was unknown by classification, inventor names and all fields that are otherwise needed to do the advanced searching.

Many of business, legal, research and management decision need to be made throughout the life cycle of a patent. Even before having an invention, the company or individual inventors need to evaluate what has already been patented in the related industry in order to know what areas of their industry to focus the innovation efforts and resources.



Google Advanced Patent Search

About Google

Find results	with **all** of the words	
	with the **exact phrase**	
	with **at least one** of the words	
	without the words	
Patent number	Return patents with the patent number	
Title	Return patents with the patent title	
Inventor	Return patents with the inventor name	First name, last name, or both
Original Assignee	Return patents with the original assignee name	First name, last name, or both
Current U.S. Classification	Return patents with the current U.S. classification	Comma separated list of one or more classification codes.
International Classification	Return patents with the international classification	Comma separated list of one or more classification codes.
Cooperative Classification	Return patents with the cooperative classification	Comma separated list of one or more classification codes.
Patent type/status	Return patents with type/status	Any type/status
Date	● Return patents anytime ○ Return patents between and e.g. 1999 and 2000, or Jan 1999 and Dec 2000	
Restrict date by	● Restrict by filing date ○ Restrict by issue date	

Figure 2. Google Advanced patent search Google Inc (2017)

A company may already be involved in research and development for a technology or product and may need to know how they should design around the boundaries already protected by other in-force patents. When approaching a large product roll-out, a company may need to conduct one last check to be sure that the features of the product can be made, used, sold, or distributed without infringing upon other in-force patents. The business decisions relating to product roll-outs or product designs can have major financial implications. Prior to filing or even drafting a patent application, an inventor and their patent practitioner may want to gauge the success which the hypothetical patent application may have when it is sent to be examined by a patenting authority. In the preceding stages of either protecting a company's patent portfolio or in seeking licensing agreements, the company may seek evidence of the company's already patented technology being made, used, sold, or distributed by others. In the event that a company is sued by another for patent infringement the defendant may attempt to find prior art that precedes the plaintiff's patents to demonstrate that the patents are invalid and unenforceable. (Lupu et al. 2011, p. 18)

While the term “patent searching” can mean “the act of searching patent information” or “searching for patents”, the phrase is more commonly used to describe searching and filtering a body of information in light of and guided by an intellectual-property related determination. (Lupu et al. 2011, p. 18)

According to Lupu et al. (2011, p. 18) there are more than one million patents applied for worldwide each year, the amount of information available to researchers and the opportunity to derive business value and market innovative new products from detailed inventions is huge. However, patent documents present several peculiarities and challenges to effective searching, analysis and management:

- They are written by patentees, who typically use their own lexicon in describing their inventive details. D. Alberts et al. (Lupu et al. 2011) pg 6
- They often include different data types, typically drawings, mathematical formulas, biosequence listings, or chemical structures which require specific techniques for effective search and analysis. (Lupu et al. 2011)
- In addition to the standard metadata (e.g., title, abstract, publication date, applicants, inventors), patent offices typically assign some classification coding to assist in managing their examination workload and in searching patents, but these classification codes are not consistently applied or harmonized across different patenting offices. (Lupu et al. 2011)

3.1 Why do patent search

Any technological based enterprise, is likely to be developing new products, services or processes, inventing and innovation. Before committing resources (Money and time) to costly development task, there is an importance to check whether anyone else has come up with the same ideas. If there is, it is not necessarily a gate block, but it may prevent a company from getting a patent of their own, and they certainly cannot copy some one else’s invention without invention owners permission. This means, it is always important and necessary to know if other similar inventions exist. (see Jolly and Pholpott 2009)

- Learning more about a new field of technology.

Before improving or innovating more on an existing technology or technical aspect, you must first understand exactly how the existing technology, implementation and design works. Companies need to search patents related to their areas of interest in order to better understand their invention, so that a better solution, innovation or advancement can be achieved.

- For market information.

  Searching patents can also help to find out which other companies are working in similar fields competing directly or indirectly to you. When you are equipped with this knowledge of the real fact about the state of innovation from patents, then you can be well informed on how to approach your solutions, face up the competition or just try to collaborate and join.

- In order to track the intellectual property of competitors.

  There is a tremendous amount of information that can be obtained from patent documents, technological trends, technology maps and growth can be tracked from patents. For a company that needs to understand the competition technologically, know who is investing heavily in what area of innovation, monitor and track the innovation trend. All these statistics can be obtained from patent documents.

- Legal purposes

  Patents have legal consequences, an inventor must first be aware whether a given patent is in force, and where. This legal status information can be found when searching patent and may have an influence on your business opportunities. Being aware of the legal status the invention, puts you in a better position to avoid law suits which might cause great financial damages to the company. On the hand, knowing the boundaries on which the existing patent is in force, a company might understand the freedom to operate regions, and use its inventions on the areas where the existing patents are not in force.

3.2 Patent searching

The challenge in searching in patents is that, patents cover a wide range of technological inventions and methods of their implementations and usability, and each area of technology has its own range of terminologies in every language, often giving words a different meaning from their ordinary dictionary definition. The English word "furnish", for example, is used in the paper making industry to indicate the materials of which paper is made. Unless a search is limited to the technological context of the subject matter being searched, the results will not be sufficiently precise. Better precision can be achieved by searching text terms in combination with patent classification codes or other indications of context.

Searching from full-text patent data requires a carefully planned strategy and being constantly aware of how a technology can be described from a scientist's or engineer's perspective versus how a technology can be described in the language of patent writers. (Lupu et al. 2011, p. 34)

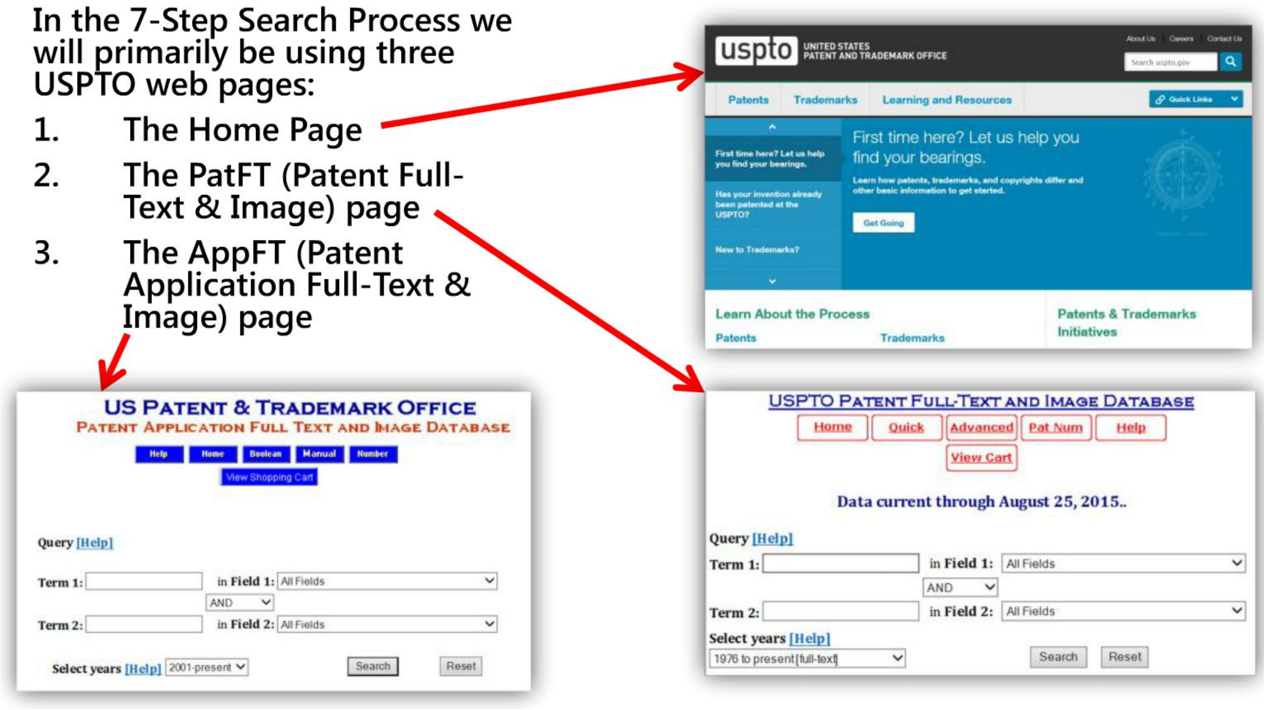
This is the main reason of why patent information search is fundamentally a complicated process, and the traditional method of keyword search is not sufficient for optimal results in the patent information context. Web based searches using search engines are wild card search, always the user will decide which returned answer is relevant from a pool or results. Mainly because this kind of search is more generic and it depends on the search keywords used.

3.2.1 Traditional search process

The following is the explanation of the current suggested and recommended search procedure by USPTO and other Patent organizations. USPTO (2017a) Titled "How to conduct a preliminary U.S. Patent search A step by step strategy" To avoid pitfalls of keyword searching, and to conduct a more thorough preliminary patent search, a classification search should be done. The following is the recommended 7-step search strategy using free web-based resources.

In the case of this illustration, three web pages will be used

- The uspto homepage (USPTO 2016a)
- The PatFT (Patent Full-text and image) page (USPTO 2017c)
- The AppFT (Patent Application Full-text and image) page



In the 7-Step Search Process we will primarily be using three USPTO web pages:

1. The Home Page
2. The PatFT (Patent Full-Text & Image) page
3. The AppFT (Patent Application Full-Text & Image) page

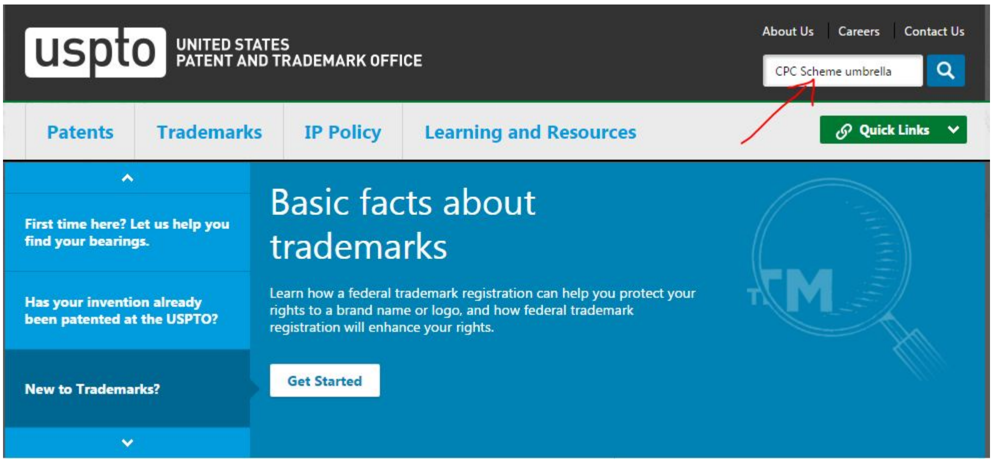
uspto UNITED STATES PATENT AND TRADEMARK OFFICE
About Us
Careers
Contact Us
Search uspto.gov
Patents
Trademarks
Learning and Resources
Quick Links
First time here? Let us help you find your bearings.
Has your invention already been patented at the USPTO?
New to Trademarks?
First time here? Let us help you find your bearings.
Learn how patents, trademarks, and copyrights differ and other basic information to get started.
Get Going
Learn About the Process
Patents
Trademarks
Patents & Trademarks Initiatives
US PATENT & TRADEMARK OFFICE
PATENT APPLICATION FULL TEXT AND IMAGE DATABASE
Help
Home
Boolean
Manual
Number
View Shopping Cart
Query [Help]
Term 1:
in Field 1: All Fields
AND
Term 2:
in Field 2: All Fields
Select years [Help] 2001-present
Search
Reset
USPTO PATENT FULL-TEXT AND IMAGE DATABASE
Home
Quick
Advanced
Pat Num
Help
View Cart
Data current through August 25, 2015..
Query [Help]
Term 1:
in Field 1: All Fields
AND
Term 2:
in Field 2: All Fields
Select years [Help]
1976 to present [full-text]
Search
Reset

Figure 3. Screen-shot of three USPTO web pages, USPTO home page, PatFT home page and APPFT home page

If we could search for “An improvement in umbrella design” Before searching if similar inventions or claims exists, the searcher should have the following questions. What is the purpose of the invention? is it a utilitarian device or an ornamental design? Is the invention a process - a way of making something or performing a function - or is it just a product? What is the invention made of? what is the physical composition of the invention? How is the invention used? What keywords and technical terms that describe the nature of the invention?

Applying the questions

An umbrella that has a new rib design to eliminate the umbrella collapsing or inverting due to high winds. A product Framework with ribs, stretchers and a main frame, securing rings, mounting brackets, joint connectors, fabric connectors, fabric, linkage bar In addition to “umbrella”: Parasol, sunshade, support assembly or apparatus, windproof, wind resistant.



Basic facts about trademarks

Learn how a federal trademark registration can help you protect your rights to a brand name or logo, and how federal trademark registration will enhance your rights.

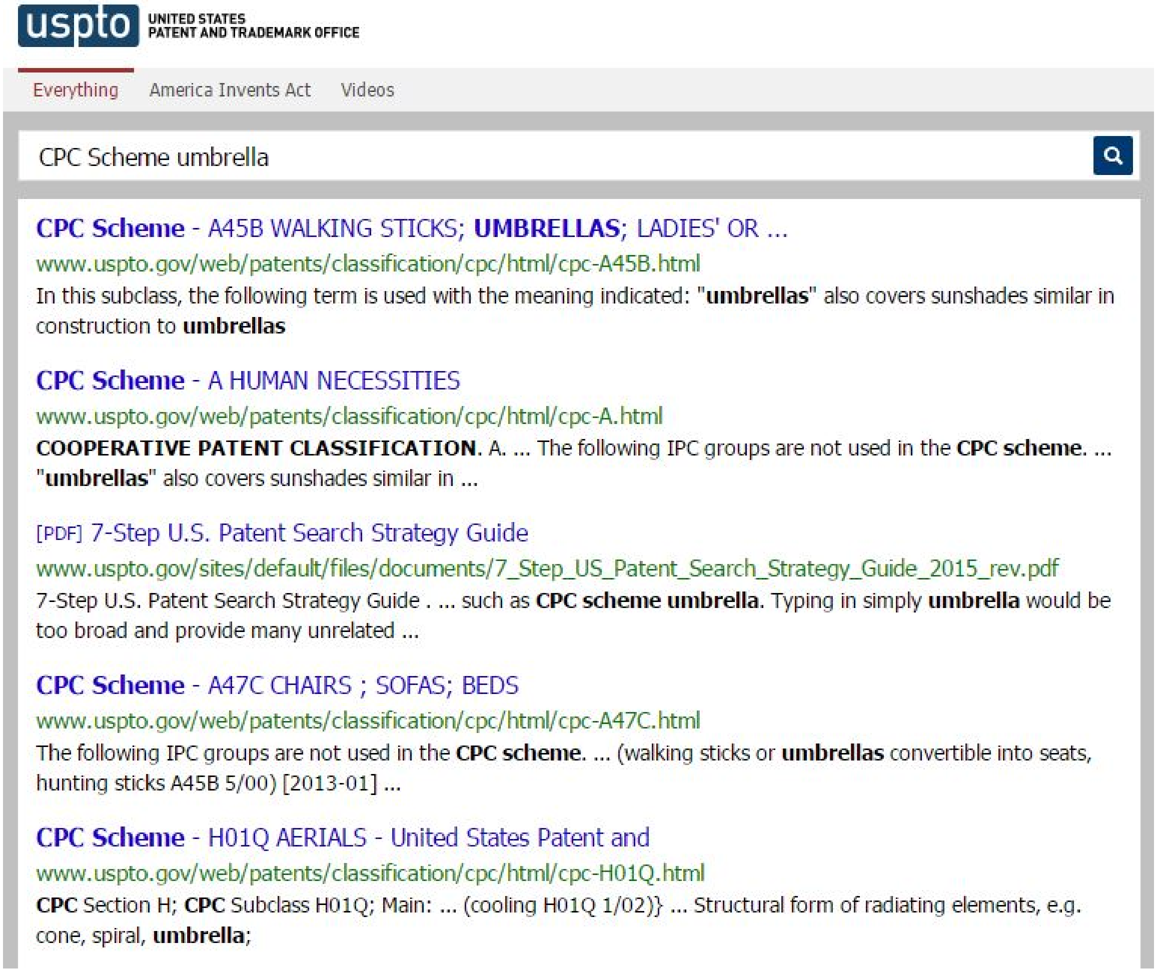
Figure 4. Screenshot of USPTO web page with a word “CPC scheme umbrella” on a search box (USPTO 2016c)

The USPTO website home page has generic search text box in the top right corner, CPC classification schema (schedules) can be searched using this text box. To achieve more desired results, we use specific language for the search terms, such as “CPC scheme umbrella” this search term will allow for results to be focused on the classification provided rather than just a plain keyword "Umbrella", typing in only “Umbrella” would be too broad as a result it will provide many unrelated search results.

From this results page, you can select an entry to access a CPC class subclass scheme page. Selecting a result from USPTO website search, click on the link for A45B which includes thw world “umbrellas” Review the entire Class-subclass A45B Scheme page. The class titles may provide additional information of cross references to other related CPC classifications. Then you need to review the Main group classifications fro umbrellas in the A45B scheme. For the full step by step searching for patents, please refer to (see USPTO 2016c)

Review classification definitions

Along the searching steps as described by USPTO, if you identified a class as relevant classification for your umbrella invention. It is time to access U.S patents that have been issued with that classification to see if someone else previously patented an invention similar to



CPC Scheme umbrella

CPC Scheme - A45B WALKING STICKS; **UMBRELLAS**; LADIES' OR ...
www.uspto.gov/web/patents/classification/cpc/html/cpc-A45B.html
In this subclass, the following term is used with the meaning indicated: "**umbrellas**" also covers sunshades similar in construction to **umbrellas**

CPC Scheme - A HUMAN NECESSITIES
www.uspto.gov/web/patents/classification/cpc/html/cpc-A.html
COOPERATIVE PATENT CLASSIFICATION. A. ... The following IPC groups are not used in the **CPC scheme**. ... "**umbrellas**" also covers sunshades similar in ...

[PDF] 7-Step U.S. Patent Search Strategy Guide
www.uspto.gov/sites/default/files/documents/7_Step_US_Patent_Search_Strategy_Guide_2015_rev.pdf
7-Step U.S. Patent Search Strategy Guide . ... such as **CPC scheme umbrella**. Typing in simply **umbrella** would be too broad and provide many unrelated ...

CPC Scheme - A47C CHAIRS ; SOFAS; BEDS
www.uspto.gov/web/patents/classification/cpc/html/cpc-A47C.html
The following IPC groups are not used in the **CPC scheme**. ... (walking sticks or **umbrellas** convertible into seats, hunting sticks A45B 5/00) [2013-01] ...

CPC Scheme - H01Q AERIALS - United States Patent and
www.uspto.gov/web/patents/classification/cpc/html/cpc-H01Q.html
CPC Section H; **CPC** Subclass H01Q; Main: ... (cooling H01Q 1/02)} ... Structural form of radiating elements, e.g. cone, spiral, **umbrella**;

Figure 5. Screenshot of USPTO search results for the word “CPC scheme umbrella”

yours. Remember, if a claimed invention has previously been publicly disclosed in “Prior Art” such as U.S patent, you cannot now get a patent on it yourself, because your invention will lack novelty.

Patent offices and other public and private sector providers of patent documents have placed tremendous emphasis on recent years on making access to and retrieval of these documents “as easy as possible.” There are two approaches to patent searching from patent databases for individuals and companies. At the entry level there are a number of free services on the Internet intended for the non-patent experts. They all have their advantages and disadvantages, but the essential common characteristic is that they are free to use, but with varying degrees of user friendliness. Not all databases contains all the patents available, some do focus on a specific country, some have patents from a certain region, i.e. European patents databases. But also some focuses only on a certain topic, i.e. Medical related patents.

Further up the patent searching hierarchy there are professional patent search services and service providers who will search patents for other clients. Users can buy into subscription-based databases, at a price of course, but there are not likely to be cost effective unless you are searching patents continuously or you are large enough to employ your own patent information specialist. (Jolly and Pholpott 2009)

3.2.2 How big is the patent search market

According to Researh and Development Magazine (2017) on its 58th annual global funding forecast estimates that global R&D investments will increase by 3.4% in 2017 to 2.066 trillion US dollars. The estimation is that about 3 to 4% of the estimated fund in R&D involves patent search, this is a significant amount of market share value that goes only to facilitate patent search. It is the very importance of understanding the value of an invention or a research, to have a clear picture of the technology in question, if it is worth the time and money of research and development.

There are specialized patent search companies build around providing search and patent discovery services to interested parties in need ot the service. Patent attorney offices also provide patent search and discovery services as an extension of service to patent filling. Companies such as CPA Global (https://www.cpaglobal.com/), Clarivate analytics (https://clarivate.com/), Thomson Reuters (https://www.thomsonreuters.com/en.html) and LexisNexis LexixNexis (2017) are just among many companies with millions of dollars turn over by providing Patent search related services.

Patent search and information processing service providers rely on patent database services to provide and conduct their search and discovery tasks. This is another area that generate a substantial revenue by providing patent databases for patent searchers. Services like espacenet and (LexixNexis 2017)

4 Semantic Annotation

Annotation is about attaching names, attributes, comments, descriptions to a document or to a selected part in a text. It provides additional information (metadata) about an existing piece of data.

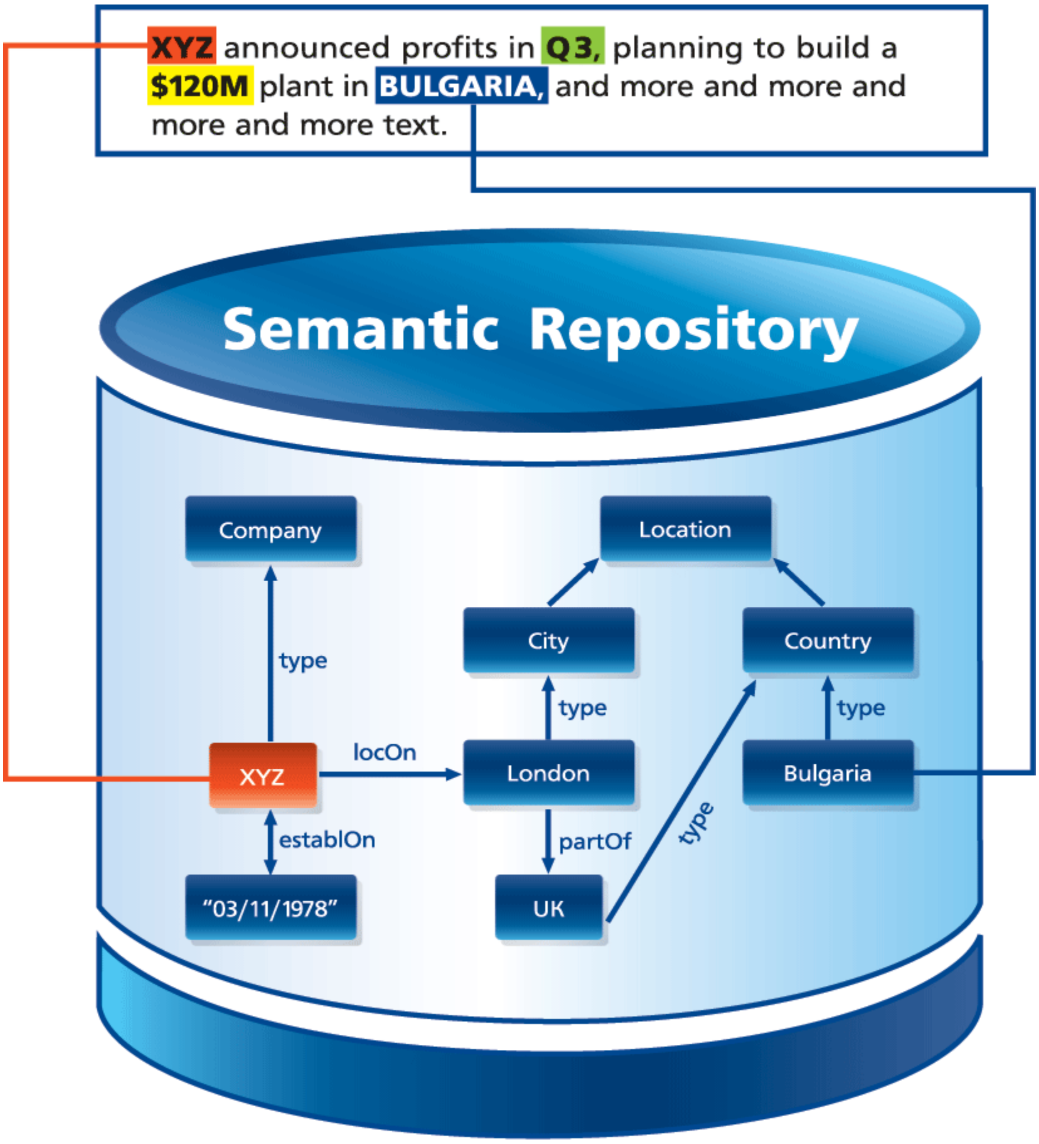


Figure 6. Annotation Diagram (Ontotex 2016)

According to Ontotex (2016) "Semantic annotation is the process of attaching additional information to various concepts (e.g. people, things, places, organizations etc) in a given text or any other content. Unlike classic text annotations for reader's reference, semantic annotations are used by machines to refer to. Semantic annotation enables several applications including semantic based information search, categorization, and composition of documents. When a document (or another piece of content, e.g. video) is semantically annotated it becomes a source of information that is easy to interpret, combine and reuse by our computers."

In a nutshell, semantic annotation is about assigning to the entities in the text links to their semantic descriptions (as presented in Ontotex (2016)Annotation Diagram). This kind of metadata provides both class and instance information about the entities. Whether these annotations should be called "semantic", "entity" or some other way, it is all a matter of terminology.

Up to now, there neither exists a well-established term for this task, nor there is a well-established meaning for the term "semantic annotation". What is more important is that the automatic semantic annotations enable many new types of applications: highlighting, indexing and retrieval, categorization, generation of more advanced metadata, smooth traversal between unstructured text and available relevant knowledge. Semantic annotation is applicable for any sort of text — web pages, regular (non-web) documents, text fields in databases, etc. Further, knowledge acquisition can be performed on the basis of the extraction of more complex dependencies — analysis of relationships between entities, event and situation descriptions, etc (Information resources management association 2016)

For instance, to semantically annotate chosen concepts in the sentence "Aristotle, the author of Politics, established the lyceum" means to identify Aristotle as person and Politics as a written work of political philosophy and to further index, classify and interlink the identified concepts in a semantic graph database. In this case Aristotle can be linked to his date of birth, his teachers, his works and Politics can be linked to its subject, to its date of creation etc. Given the semantic metadata about the above sentence and its links to other (external or internal) formal knowledge, algorithms will be able to automatically:

- Find out who tutored Alexander the Great.
- Answer which of Plato's pupils established the Lyceum.
- Retrieve a list of political thinkers who lived between 380 and 310 BC.

- Render a page about Greek philosophers and include Aristotle.

In the current state of data concentration, there is an amazing resource for all sorts of information that can be used for about anything, programming, learning to play music instrument, medical and many other useful applications. However there is another layer of information that is available and being communicated by means of blogs, tweets, journals, articles. Take the web for example, it contains the information in all kind of form, including texts, images, videos and audio, and from all these Language is the communication medium that enables human beings to understand the content and context as well as relate/link them from one media to another. Despite the fact that computers excellent at delivering this information to the interested users, the systems are inadequate in understanding the language itself.

According to Amber Stubbs (2012) "Theoretical and computational linguistics are focused on unraveling the deeper nature of language and capturing the computational properties of linguistic structures. Human language technologies (HLTs) attempt to adopt these insights and algorithms and turn them into functioning, high-performance programs that can impact the ways we interact with computers using language. With more and more people using the Internet every day, the amount of linguistic data available to researchers has increased significantly, allowing linguistic modeling problems to be viewed as ML tasks, rather than limited to the relatively small amounts of data that humans are able to process on their own."

However, it is not enough to simply provide a computer with a large amount of data and expect it to learn to speak—the data has to be prepared in such a way that the computer can more easily find patterns and inferences. This is usually done by adding relevant metadata to a dataset. Any metadata tag used to mark up elements of the dataset is called an annotation over the input. However, in order for the algorithms to learn efficiently and effectively, the annotation done on the data must be accurate, and relevant to the task the machine is being asked to perform. For this reason, the discipline of language annotation is a critical link in developing intelligent human language technologies. (Amber Stubbs 2012)

Datasets of natural language are referred to as corpora, and a single set of data annotated with the same specification is called an annotated corpus. Annotated corpora can be used to train ML algorithms. In this chapter we will define what a corpus is, explain what is meant by an annotation, and describe the methodology used for enriching a linguistic data collection with annotations for machine learning. Compared to tagging, which speeds up searching and helps you find relevant and precise information, semantic annotation goes one level deeper, It enriches the unstructured of semi-structured data with context that is further linked to the structured knowledge of an a domain. It allows results that are not explicitly related to the original search. (Pustejovsky and Stubbs 2012)

Semantic annotation on the other hand, it helps to bridge the ambiguity of the natural language when expressing notions and their computational representation in a formal language. By telling a computer how data items are related and how these relations can be evaluated automatically, it becomes possible to process complex filter and search operations. The difference between annotation and other forms of meta-data is that an annotation is grounded to a specific point in a document. For example, it might be considered that a folder name is a form of meta-data. But when a file is removed from a folder, it becomes separated from that meta-data and as a results, it loses some valuable context.

4.0.1 Ontologies

Referring to the book of Staab and Studer (2013) Ontology focuses on the nature and structure of things per se, independently of any further considerations, and even independently of their actual existence. The author continues to define Ontology as used in the computer science "We refer to an ontology as a special kind of information object or computational artifact. the account of existence in this case is a pragmatic one: For AI systems, what exists is that which can be represented." Computational ontologies are means to formally model the structure of a system, i.e., the relevant entities and relations that emerge from its observation, and which are useful to our purposes.

Ontology is the science of things existing, or things existing permanently; and its object is to determine what these things are, how we come to know them, and into how few classes they may be divided without doing violence to their essential differences. Ramsay (1870) The author continue to explain that, Now, as things cannot be known of themselves to be permanent, for everything within and without us in changeable, transitory, fleeting everything which we directly experience is but the moment, it follows that things permanent can be know only by inference. Although it is required from an ontology to be formally defined, there is no common definition of the term "ontology" itself. The definitions can be categorized into roughly three groups:

- Ontology is a term in philosophy and its meaning is “theory of existence”.
- Ontology is an explicit specification of conceptualization.
- Ontology is a body of knowledge describing some domain, typically common sense knowledge domain.

The first definition is the meaning in philosophy, however it has many implications for the Artificial Intelligence purposes. The second definition is generally accepted as a definition of what an ontology is for the AI community. The third definition views an ontology as an inner body of knowledge, not as the way to describe the knowledge.

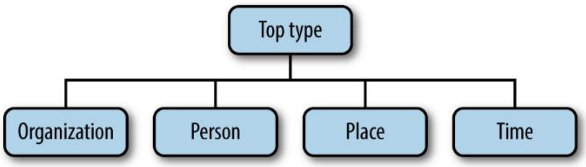


Figure 7. A simple Ontology diagram (Ontotex 2016)

4.0.2 Semantic Web

The Semantic Web is the name of a long-term project started by W3C with the stated purpose of realizing the idea of having data on the Web defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration, and reuse of data across various applications (W3C 2017). The Semantic Web is a Web-technology that lives on top of the existing Web by including machine-readable information in files without modifying the existing Web structure.

According to Semantics (2017) The Semantic Web, Web 3.0, the Linked Data Web, the Web of Data, the Semantic Web represents the next major evolution in connecting information. It enables data to be linked from a source to any other source and to be understood by computers so that they can perform increasingly sophisticated tasks on our behalf. The Semantic Web abstracts away the document and application layers involved in the exchange of information. The Semantic Web connects facts, so that rather than linking to a specific document or application, you can instead refer to a specific piece of information contained in that document or application. If that information is ever updated, you can automatically take advantage of the update. From a technical point of view, the Semantic Web consists primarily of three technical standards:

- RDF (Resource Description Framework): The data modeling language for the Semantic Web. All Semantic Web information is stored and represented in the RDF.
- SPARQL (SPARQL Protocol and RDF Query Language): The query language of the Semantic Web. It is specifically designed to query data across various systems.
- OWL (Web Ontology Language) The schema language, or knowledge representation (KR) language, of the Semantic Web. OWL enables you to define concepts composably so that these concepts can be reused as much and as often as possible. Composability means that each concept is carefully defined so that it can be selected and assembled in various combinations with other concepts as needed for many different applications and purposes.

4.1 Meta-data

Metadata is data that describes other data, or otherwise metadata is data about data. The main purpose of metadata is to summarize the basic information about data, or provide key details about a set of information or data, an image itself is a digital data, but it has metadata that describes the camera used to take it, time and date the picture was taken, exposure and such. this helps to make working with particular instances of data more manageable. The term metadata is used differently in different communities. It might be used to refer to machine understandable information, while on the other hand is is only used for records that describe electronic resources.

There are three main types of metadata:

- Descriptive metadata describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords.

- Structural metadata indicates how compound objects are put together, for example how pages are ordered to form chapters.
- Administrative metadata provides information to help manage a resource, such as when and how it was creates, file type and other technical information, and who can access it.

4.1.1 Patent Meta data

In Patents, Meta data is the data that describes or are related to the patents documents, they can be categorized into two types, explicit and implicit metadata. Explicit metadata is usually given in the front page of a patent document, this is most of the time referred to as bibliographic information. For example, Invention title, Inventors name, assignees name, classification, inventor country to where the invention is to be protected. While Implicit metadata according to Mark Giereth et al, has to be extracted from higher level associations between patent documents as well as from their textual content, for example patent or literature citations occurring in the patent content or the patent type extracted from the claims.

Patent metadata can further be classified into internal and external data, Internal metadata can be derived from a single patent document, whereas for external metadata other patent documents or data sources have to be taken into consideration. Examples for external metadata are events concerning the legal status of a patent or additional applicant or inventor information.

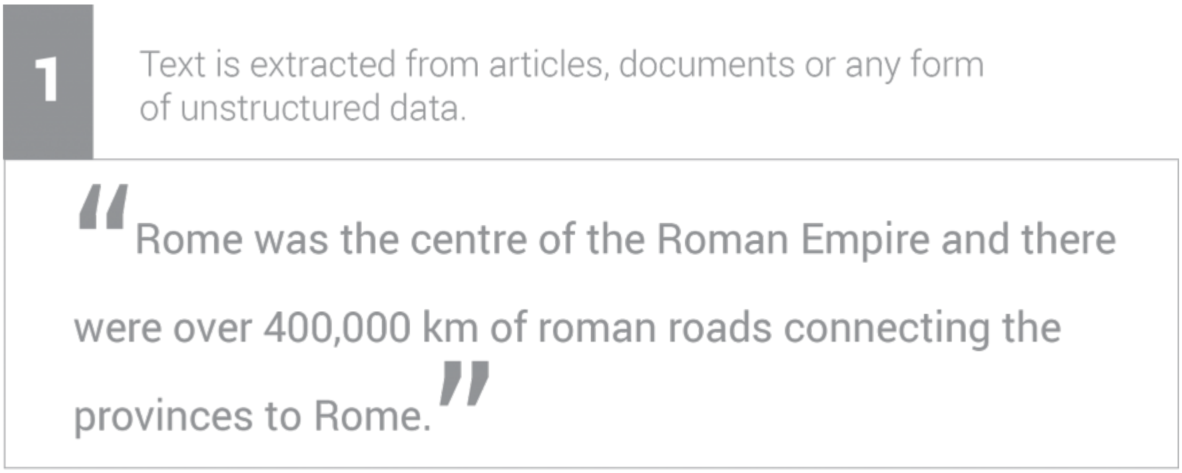
4.2 How to annotate

Annotation can be done by means of flat tagging or labeling of content with meta-tags, this method is currently used extensively on blogs and news content to support the contextual surfacing of content. Semantic annotation is a more sophisticated step than flat tagging. Semantic annotation is the application of ontologically modeled references to digital media content.

An example given by datalanguage, instead of simply applying a free text word of phrase as a tag, you apply associate to the relevant piece of content a URI reference to an instance of a “thing” that has been adequately domain modeled. As an example, if a free text term “Elvis Presley” is applied to music article, then it would be able to some extent to allow users to know that the article was about Elvis Presley, and maybe be able to explore for other articles that have been tagged with the exact phrase. Although if another article was tagged with “Elvis” there most likely be no correlation between the two terms, but if however, we associated a URI referencing “Elvis Presley” in some ontological domain model to the content, then we also gain all knowledge about Elvis from the underlying model.

Semantic annotation enriches content with information that can be machine processed by connecting background information to extracted concepts. The concepts that are found in a document or any other content are unambiguously defined and related to each other withing and outside the content.

Text identification



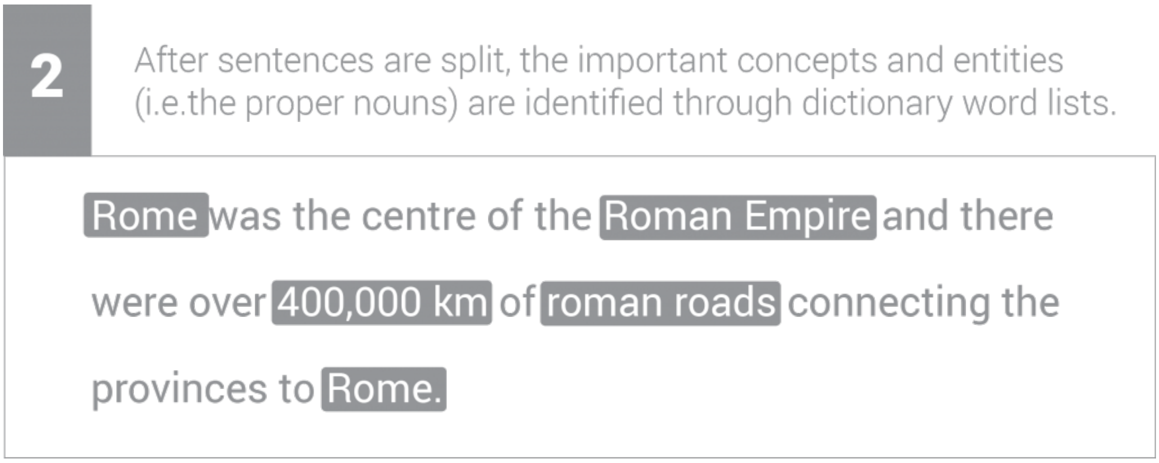
1 Text is extracted from articles, documents or any form of unstructured data.

“Rome was the centre of the Roman Empire and there were over 400,000 km of roman roads connecting the provinces to Rome.”

Figure 8. Text example to be annotated (Ontotex 2016)

Then text is extracted from various sources, including non-textual sources such as PDF files, videos, documents and voice files.

Sentences are split by algorithms and concepts, such as people, things, places, events, num-



2

After sentences are split, the important concepts and entities (i.e.the proper nouns) are identified through dictionary word lists.

Rome was the centre of the Roman Empire and there were over 400,000 km of roman roads connecting the provinces to Rome.

Figure 9. Text analysis (Ontotex 2016)

bers are identified.

Concept Extraction

All recognized concepts are classified, that is they are defined as people, organizations, numbers etc. Next, they are disambiguated, that is they are unambiguously defined according to a domain-specific knowledge base. For example, Rome is classified as a city and further disambiguated as Rome, Italy not Rome, Iowa. This is the most important stage of semantic annotation. It very much resembles Named Entity Recognition but is different for it not only recognizes text chunks but also makes them machine-processable and understandable data pieces by linking them to a broader sets of already existing data.

Relationship Extraction

The relationships between the extracted concepts are identified and interlinked with related external or internal domain knowledge.

Indexing and storing in a semantic graph database

All the recognized and enriched with machine-readable data mentions of people, things, numbers etc and the relationships between them are indexed and stored in a semantic graph database for further reference and use.

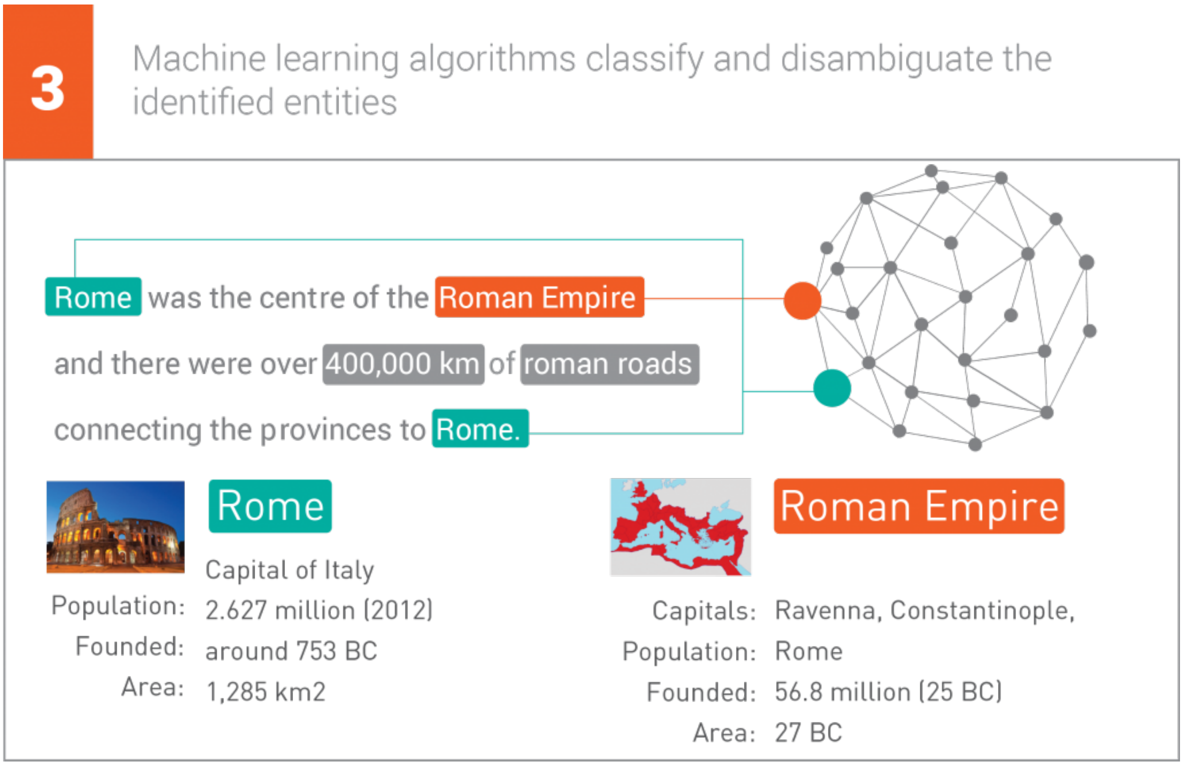

3
Machine learning algorithms classify and disambiguate the identified entities
Rome was the centre of the Roman Empire
and there were over 400,000 km of roman roads
connecting the provinces to Rome.
Rome
Capital of Italy
Population: 2.627 million (2012)
Founded: around 753 BC
Area: 1,285 km2
Roman Empire
Capitals: Ravenna, Constantinople,
Population: Rome
Founded: 56.8 million (25 BC)
Area: 27 BC

Figure 10. Annotation (Ontotex 2016)

4.2.1 Document annotation

The traditional domain of document annotation covers the annotation of arbitrary textual documents, or parts of them. Annotations can be manual (performed by one or more people), semi-automatic (based on automatic suggestions), or fully automatic. Manual annotation tools allow users to add annotations to web pages or other resources, and share these with others. An example annotation would relate the text “Paris” to an ontology, identifying it as a city and as capital of France. Automatic tools can perform similar annotations (such as named-entity recognition) without manual intervention. (Oren et al. 2016)

4.2.2 Manual annotation

Manual annotation (MA) is a methodology that transforms the existing syntactic resources into interlinked knowledge structures by adding information to some level of document (word, phrase or paragraph) which constitutes metadata. The process of manual annotation is expensive, and regularly does not consider that multiple standpoints of a data source,

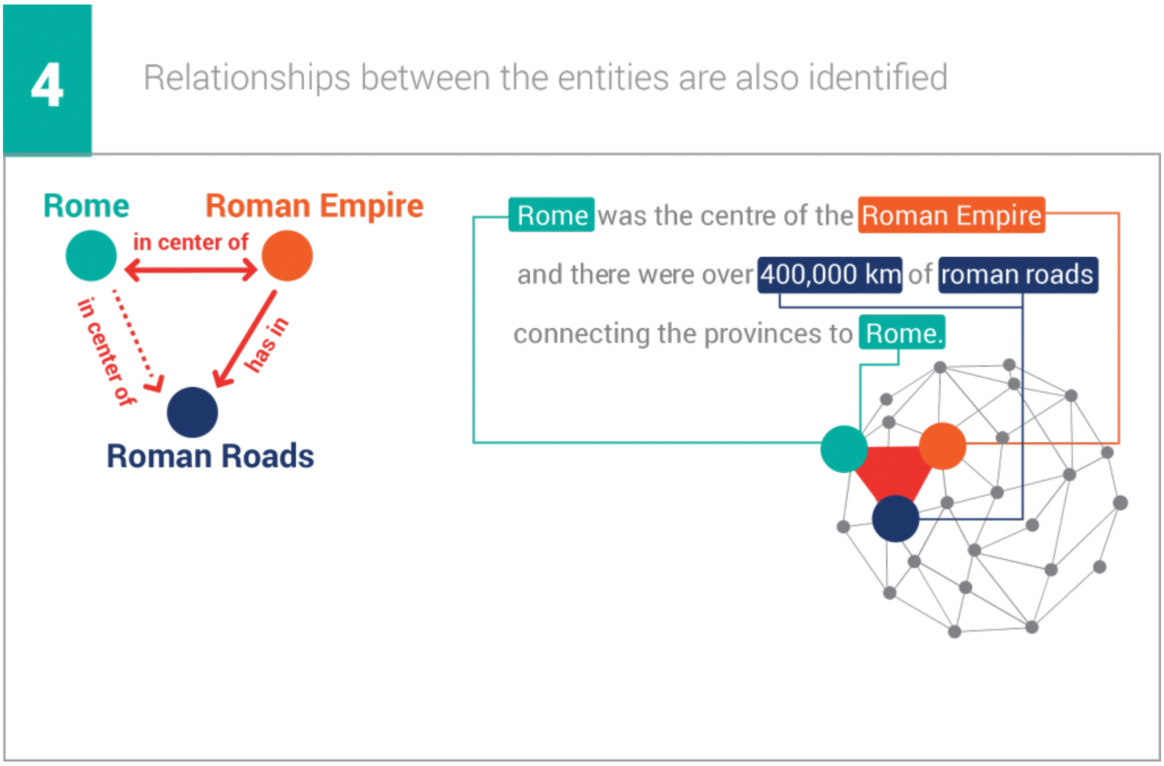

4
Relationships between the entities are also identified
Rome
Roman Empire
in center of
in center of
has in
Roman Roads
Rome was the centre of the Roman Empire
and there were over 400,000 km of roman roads
connecting the provinces to Rome.

Figure 11. Relationship Extraction, Figure from ontotext website(Ontotex 2016)

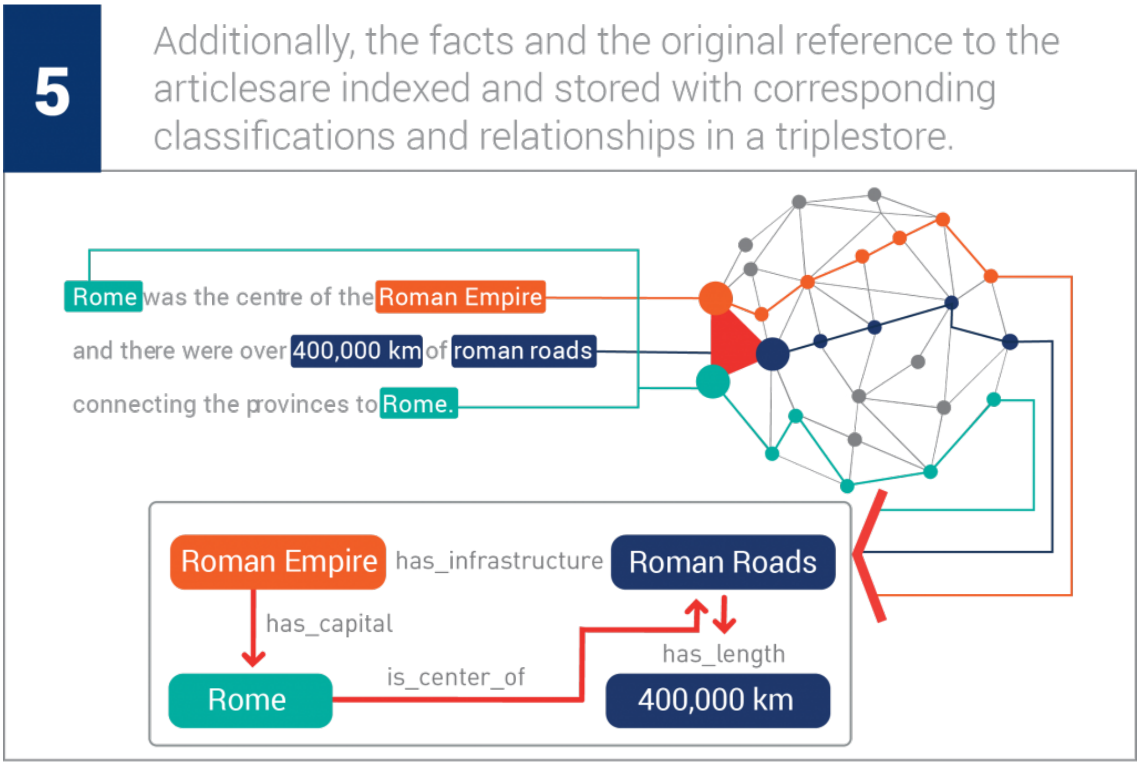
involving multiple ontologies, can be useful to support the requirements of different users. Furthermore, MA is more easily feasible today, by means of authoring tools such as Semantic Word. MA is more precise compared to automatic annotation, but is very labor-intensive.

4.2.3 Automatic Annotation

Automatic semantic annotation is an ideal solution to the big data and large datasets problem, because it is obvious nearly impossible to apply manual semantic annotation to a relative large dataset. However, the fully automatic creation of semantic annotations is also an unsolved problem. Hence, semi-automatic creation of annotations is the method mostly used in current systems.

There are many automatic annotation methods that have been proposed, including:

supervised machine learning based method The supervised machine learning based method



5

Additionally, the facts and the original reference to the articlesare indexed and stored with corresponding classifications and relationships in a triplestore.

Rome was the centre of the Roman Empire and there were over 400,000 km of roman roads connecting the provinces to Rome.

Figure 12. Indexing and storing in a semantic graph database (Ontotex 2016)

consists of two stages: annotation and training. In annotation, we are given a document in either plain text or semi-structured (e.g. emails, web pages, forums, etc.), and the objective is to identify the entities and the semantic relations between the entities. In training, the task is to learn the model(s) that are used in the annotation stage. For learning the models, the input data is often viewed as a sequence of units, for example, a document can be viewed as a sequence of either words or text lines (depending on the specific applications). In the supervised machine learning based method, labeled data for training the model is required.

Unsupervised machine learning based method The unsupervised machine learning based method tries to create the annotation without labeled data. The generalized patterns can then be used to extract the data from the Web.

ontology based method The ontology based method employs the other knowledge sources like thesaurus, ontology, etc. The basic idea is to first construct a pattern-based ontology, and then use the ontology to extract the needed information from the web page. (Tang et al. 2009)

4.2.4 Mixed Annotation

The mixed method automation sometimes referred to as semi-automatic annotation process requires human intervention at some annotation level. Generally the context that is easily automated would be automatically annotated, but there are some complicated context where a human intervention is needed to provide a more meaningful annotation, This category of annotation systems differs in their architecture, methods and tools of information extraction, the manual work amount required to achieve annotation, performance, storage management and other features. (Slimani 2013)

4.2.5 The challenge of novelty

Novelty is a requirement criteria for a patent claim to be patentable and get granted, an invention is not regarded as new it was known to the public before it was filed, or before its date of priority if the applicant claims priority of an earlier patent application. In order for an invention to be patentable it must be new as defined in the patent law, which provides that an invention cannot be patented if:

"(1) The claimed invention was patented, described in a printed publication, or in public use, on sale, or otherwise available to the public before the effective filing date of the claimed invention" or See [page 342] (Bouchoux 2016)

"(2) The claimed invention was described in a patent issued [by the U.S.] or in an application for patent published or deemed published [by the U.S.], in which the patent or application, as the case may be, names another inventor and was effectively filed before the effective filing date of the claimed invention." See [page 342] (Bouchoux 2016)

Given the fact that creating ontologies for data to be annotated is a rather complicated process, It needs resources and expertise to facilitate the creation of ontology. The ontology has to match with the context of the data that is being annotated, making sure that the meaning and classes are optimal to give the correct meaning of the date being annotated. From the novelty point of view, the patent information is newer to every new Patent that is been filed, inventions and new discoveries means there are new terms, topics, and words that are introduced into the technology sphere. This on itself makes ontology based solution to be a challenge, because in order to cope with the new inventions, new ontologies needs to be generated or created for every new invention or patent being filed. This work itself proves to be a challenge rather than a solution to the problem. Due to the infancy of ontology and semantic technology, tools and means of automatic ontology generation is still a work in progress, this is not a completely impossible task, rather a complicated process at this particular state of technology level on the subject. With more research and innovative solution on the tools, methods and ways to solve and improve the ontology generation from existing and new content, we will witness a tremendous development and implementation of various ontological and semantic solutions.

The semantic ontology for Patents is similar to a rear view mirror while driving a car, inventions are the way forward and the rear view mirror shows a history of what has been done. In this case Ontology can solve the problem of historical data discoveries and adding semantics to it, but not so much to the future of information that might not be present currently.

5 Improving patent search using annotation

5.1 Opportunities

Whenever there is a challenge or a problem, there is an opportunity. This is no different with patent search and other big data repositories, searching and information extraction from data collection. In the last 15 years there has been a tremendous explosion of amount of digital data generated, from a variety of source such as Internet, mobile devices, social media, different equipments and data sensors and this is the beginning of Big Data as known in computer and digital world.

Many large corporations as well as startups are investing into big data management in terms of searching and knowledge extraction. But there is even more need and usability of big data to the technology environment, the introduction of semantic web, and now Internet of Things (IoT) brings a high demand in digital data and conceptualization of data. Digital systems and machines needs to not only use data, but understand the concept and meaning of data. Any dataset, big or small it will useful as long as it is reliable, and there are a good number of data quality assurance opportunities.

Natural language processing (NLP) and Machine deep learning are among the areas that many scientific researchers are putting their efforts to give data the properties that will enable systems to understand the meaning of data.

There is a very wide possibilities of what has to be done to take advantage of the big data and its potential to the new technology stack of users and systems, researchers, innovators, inventors and technology and information stake holders to capitalize on the opportunity. Patents are one example of this, there are multi billion companies that are providing patent data and traditional search databases for patent information. And companies like TEQMINE ANALYTICS LTD are frontiers of introducing and implementing unsupervised machine learning aiding to patent similarity search. There is a high demand for the services that solves this patent information searching, as well as other types and categories of big data sets.

5.2 Topic Modeling

According to Balagopalan (2012) "Topic modeling is a machine learning technique that can be used to analyze large collections of unlabeled documents in an unsupervised setting. Topics consist of groups of words that co-occur frequently in documents. Latent Dirichlet Allocation (LDA) is perhaps the most widely used topic model."

Large amounts of patent data are created and collected everyday, reflecting the rate of global inventive effort, as well as the overall effort put into research and development continuously. As more information becomes available, the challenge to analyze and digest the new information increases, and is manifest in technical difficulties to carry out patent search and other information processing. Thus, we need tools and techniques to organize, search and understand large quantities of information and data that is put into our disposal for human and computer systems to process.

Topic modeling is a machine learning method that provides us with methods to organize, understand and summarize large collections of textual information. It helps in:

- Discovering hidden topical patterns that are present across the collection
- Annotating documents according to these topics
- Using these annotations to organize, search and summarize texts

There are many techniques that are used to obtain topic models.

In machine learning and natural language processing, a topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body. Intuitively, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently: "dog" and "bone" will appear more often in documents about dogs, "cat" and "meow" will appear in documents about cats, and "the" and "is" will appear equally in both.

A document typically concerns multiple topics in different proportions; thus, in a document that is 10% about cats and 90% about dogs, there would probably be about 9 times more dog words than cat words. The "topics" produced by topic modeling techniques are clusters of similar words. A topic model captures this intuition in a mathematical framework, which allows examining a set of documents and discovering, based on the statistics of the words in each, what the topics might be and what each document's balance of topics is.

Topic models can also referred to as probabilistic topic models, which refers to statistic algorithms for discovering the latent semantic structures of an extensive text body. Given the vast amount of information generated everyday from unprecedented number of data sources and sensors, the amount of the written material we encounter each day is simply beyond our processing capacity. Topic models can help to organize and offer insights for human to understand large collections of unstructured text bodies. Originally developed as a text-mining tool, topic models have been used to detect instructive structures in data such as genetic information, images, and networks.

5.2.1 Latent Dirichlet Allocation (LDA)

Blei, Ng, and Jordan (2003) describes latent Dirichlet allocation (LDA), as a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. We present efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation. We report results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model.

Suppose you have the following set of sentences: (Chen 2017)

- I like to eat broccoli and bananas.
- I ate a banana and spinach smoothie for breakfast.
- Chinchillas and kittens are cute.
- My sister adopted a kitten yesterday.
- Look at this cute hamster munching on a piece of broccoli.

Latent Dirichlet allocation is a way of automatically discovering topics that these sentences contain. For example, given these sentences and asked for 2 topics, LDA might produce something like the following

Sentences 1 and 2: 100% Topic A
Sentences 3 and 4: 100% Topic B
Sentence 5: 60% Topic A, 40% Topic B
Topic A: 30% broccoli, 15% bananas, 10% breakfast, 10% munching, ... (at which point, you could interpret topic A to be about food)
Topic B: 20% chinchillas, 20% kittens, 20% cute, 15% hamster, ... (at which point, you could interpret topic B to be about cute animals)

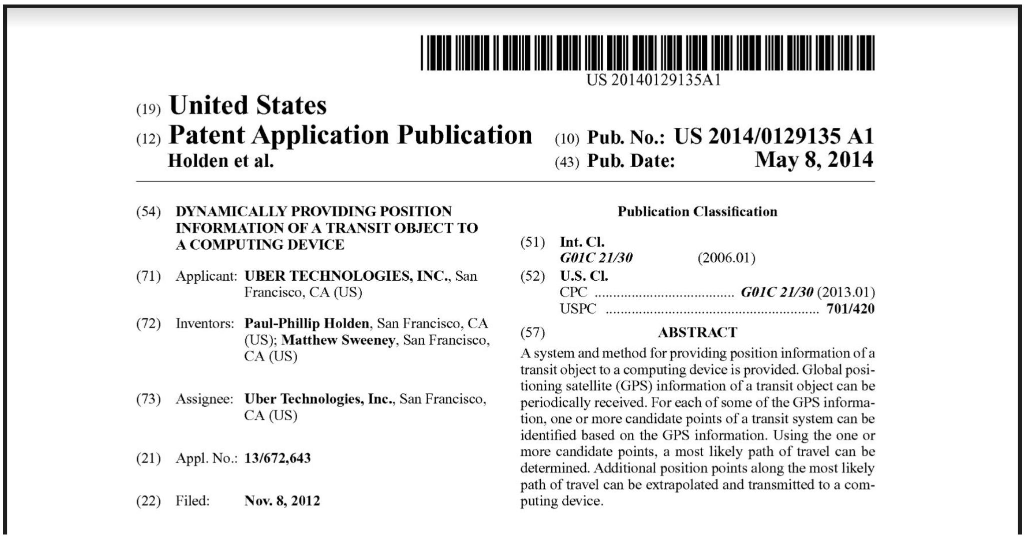
5.2.2 Similarity NOT Relevance

At a high level, topic modeling aims to find structure within an unstructured collection of documents. After learning this "structure," a topic model can answer questions such as: What is document X discussing? How similar are documents X and Y? If I am interested in topic Z, which documents should I focus and read on? Using Topic modeling will not yield a pin point answer/result to a search, rather it will narrow down the collection pool for human intelligence to concentrate on. In this case instead of shuffling around several million documents, you are left with a couple of thousand relevant document to your topic or area of interest, and then you can dig deeper using search tools or even fine tune the search and do a nested topic modeling on top of the results.

Using keyword search or any other kind of advanced text search on a filtered pool of documents that are relevant to your interest, increases accuracy and reduces the hit and miss of searching through a corpus of unstructured document text. It also gives value to the time spent on searching, by making sure that all the results from your refined search indeed comes from the relevant documents related to your area of interest.

5.3 Experimentation

In this experiment, Uber Technologies Patent (See (Holden and Sweeney 2014)) is going to be used as a reference document, the search will be conducted to find out if there is a filed patent with same invention claims, also the search experiment should aim to find out the technology map around the said patent technology landscape. The rest of its bibliographic information can be seen on the figure 13.



US 20140129135A1

(19) **United States**
(12) **Patent Application Publication** (10) **Pub. No.: US 2014/0129135 A1**
Holden et al. (43) **Pub. Date: May 8, 2014**

(54) **DYNAMICALLY PROVIDING POSITION INFORMATION OF A TRANSIT OBJECT TO A COMPUTING DEVICE**

(71) Applicant: **UBER TECHNOLOGIES, INC.**, San Francisco, CA (US)

(72) Inventors: **Paul-Phillip Holden**, San Francisco, CA (US); **Matthew Sweeney**, San Francisco, CA (US)

(73) Assignee: **Uber Technologies, Inc.**, San Francisco, CA (US)

(21) Appl. No.: **13/672,643**

(22) Filed: **Nov. 8, 2012**

Publication Classification

(51) **Int. Cl.**
G01C 21/30 (2006.01)

(52) **U.S. Cl.**
CPC ***G01C 21/30*** (2013.01)
USPC **701/420**

(57) **ABSTRACT**

A system and method for providing position information of a transit object to a computing device is provided. Global positioning satellite (GPS) information of a transit object can be periodically received. For each of some of the GPS information, one or more candidate points of a transit system can be identified based on the GPS information. Using the one or more candidate points, a most likely path of travel can be determined. Additional position points along the most likely path of travel can be extrapolated and transmitted to a computing device.

Figure 13. Uber Patent No US2014/0129135, snapshot from Patent PDF Document (Holden and Sweeney 2014)

Patent US2014/0129135 abstract

The following is an abstract from Holden and Sweeney (2014) "A system and method for providing position information of a transit object to a computing device is provided. Global positioning satellite (GPS) information of a transit object can be periodically received. For each of some of the GPS information, one or more candidate points of a transit system can be identified based on the GPS information. Using the one or more candidate points, a most likely path of travel can be determined. Additional position points along the most likely path of travel can be extrapolated and transmitted to a computing device."

One of this patent claim is: A method of providing position information of a transit object to a computing device, the method being performed by one or more processors and comprising:

- periodically receiving Global Positioning Satellite (GPS) information of the transit object, the GPS information including a latitude of the transit object, a longitude of the transit object, and a GPS error amount at a given instance;
- identifying, for each of some of the GPS information, one or more candidate points of a transit system that are within the GPS error amount for that GPS information;
- determining a most likely path of travel of the transit object on the transit system based on the identified one or more candidate points of the transit system;
- extrapolating points along the most likely path of travel; and transmitting a set of extrapolated points to the computing device

More about the patent and its claims as well as other related information can be accessed through uspto database or Google Patent repository for uspto, using patent number (US2014/0129135). This patent document will be used as a main reference document that the search will be performed against, as a test case we also have the advantage of knowing before hand that, this particular patent has been filed in United States of America USPTO patenting office as well as in (PCT) Patent Cooperation Treaty.

5.3.1 Scenario

There are many scenarios of why should an interested person search the existing patent databases, for example an inventor might need to investigate if his invention idea has already been done and patent filed by some other inventors. Data source, the data sets that are used in the experimentation have been obtained from reputable world patent organizations, among them is a free subscription service (USPTO) while the the other two source are paid subscription service products. TEQMINE Analytics Ltd (2017) have given me the permission to let me use the data for my thesis project. TEQMINE Analytics Ltd, is a startup founded in 2013 based in Helsinki, Finland. Specializing in large-scale IPR, patent, providing AI assisted patent similarity search.

The patent data used in this experimentation is obtained from three official patent organizations. USPTO Data Products Available Directly from USPTO. All trademark bulk data products and many patent bulk data products are available online from the USPTO at no charge. Due to their size and complexity, some patent data products are available from the USPTO only on a fee basis. Data products obtained directly from the USPTO are available on the publication date.

PCT 1978 - 2016 = 2,963,029 Full text patent documents
EPO 1978 - 2016 = 4,008,528 Full text patent documents
USPTO 1990 -2016 = 6,706,486 Full text patent documents
About 13,000,000 patents publications

5.3.2 Data preparation

In patent world, a single Patent can mean the difference between breaking or making a company in terms of development or business advantage in the competition scope, this brings the importance of data accuracy and integrity in the preprocessing stage all the way to the presentation of the results to the end user. To facilitate this requirement, crucial pre processing steps and data validation must be taken before the data can be used by the modeling and other analytical tools. If it happens that, there is a patent missing form the database, and this patent has crucial similarities in claims with an invention that is about to be put into research project, a company might incur unnecessary cost of patent filling as well as time spent on the process. This is due to the fact that, the patent examiners will reject the patent due to the luck of novelty because there is an existing patent that was missed during the preliminary prio-art search.

Originally the Patent documents attributes are structured to facilitate patents filing purposes, the filling structure is human readable, and supports traditional filling systems, but they lack the kind of metadata structured that could aid the digital processing of documents. Though in recent years, since 2001, International Patent Offices such as EPO, PCT and USPTO begun implementing XML structure into patent documents. XML Extensible Markup Language is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. Harold (2004) Having patent documents in XML format has made the digital processing and transformation of patents a manageable task.

Older patents from 1978 up to 1990 are still in plain text file format, these file formats does not contain any metadata, digital markup or identifiers that would make it easy to structure data digitally. There are more older patent documents records from 18th century, but they are not digital, they are in a form of scans and many not in OCR standard to be used for digitization easily, older patents ware manually typed using carbon print typewriters, the quality of document text is varying a lot, and many can not easily be OCR read.

Even though latest patent documents are in XML format, they are still a lot of challenges to the quality of data structure to qualify for clean processing. Some Offices pack all the weekly patents into a single XML file, and many contains chemical formulas, product descriptions that are in a format that breaks the XML validation requirements. For this reason, Text manipulation tools and algorithms are needed to extract sections of text into meta structured document before the text can be used for further processing. This behaviour varies from one patent office to another, each office has its own way of packaging the patent documents, so the documents need to be treated in variation and accordingly during pre-processing. By using a carefully designed database structure, the selected meta from sections of a patent document, such as abstract, claims and full text description is extracted and parsed ready for deep learning processes, annotation and structuring.
Then the data is then cleaned as required, making sure that all data content is in the correct encoding format and all the ambiguous characters has been cleaned and removed, this step is very important to the effectiveness and quality of the model that the NLP will be run upon.

Data preparation and structuring is very important and fundamental stage in achieving the successful process of training the model for use on similarity AI (Artificial Intelligent application). The quality and integrity of the data depends on how much effort has been put into guaranteeing that data is clean, reliable and usable. On this experiment project, data had to be brought to a relational database for easy querying and rearanging data into logical structures that supports interlinking of worldwide patent documents to each other. MySQL (2017) has been used as database engine for the experimentation.

Gensim (Gensim 2017) Gensim is a free Python library designed to automatically extract semantic topics from documents, as efficiently (computer-wise) and painlessly (human-wise) as possible. The algorithm is designed to process raw, unstructured digital texts ("plain text"). The algorithms in gensim 5.3.2, such as Latent Semantic Analysis, Latent Dirichlet Allocation 5.2.1 and Random Projections discover semantic structure of documents by examining statistical co-occurrence patterns of the words within a corpus of training documents. These algorithms are unsupervised, which means no human input is necessary – you only need a corpus of plain text documents. One these statistical patterns are found, any plain text documents can be succinctly expressed in the new, semantic representation and queried for topical similarity against other documents.The whole gensim package revolves around the concepts of corpus, vector and model.

Corpus A collection of digital documents. This collection is used to automatically infer the structure of the documents, their topics, etc. For this reason, the collection is also called a training corpus. (McEnery and Wilson 2001) This inferred latent structure can be later used to assign topics to new documents, which did not appear in the training corpus. No human intervention (such as tagging the documents by hand, or creating other metadata) is required.

Sparse Vector In the Vector Space Model (VSM), each document is represented by an array of features. For example, a single feature may be thought of as a question-answer pair:

How many times does the word splonge appear in the document? Zero.
How many paragraphs does the document consist of? Two.
How many fonts does the document use? Five.

The question is usually represented only by its integer id (such as 1, 2 and 3 here), so that the representation of this document becomes a series of pairs like (1, 0.0), (2, 2.0), (3, 5.0). If we know all the questions in advance, we may leave them implicit and simply write (0.0, 2.0, 5.0). This sequence of answers can be thought of as a vector (in this case a 3-dimensional vector). For practical purposes, only questions to which the answer is (or can be converted to) a single real number are allowed. Gensim (2017) the questions are the same for each document, so that looking at two vectors (representing two documents), we will hopefully be able to make conclusions such as "The numbers in these two vectors are very similar, and therefore the original documents must be similar, too". Of course, whether such conclusions correspond to reality depends on how well we picked our questions.

Typically, the answer to most questions will be 0.0. To save space, we omit them from the document's representation, and write only (2, 2.0), (3, 5.0) (note the missing (1, 0.0)).

Gensim does not prescribe any specific corpus format; a corpus is anything that, when iterated over, successively yields these sparse vectors. For example, set((((2, 2.0), (3, 5.0)), ((0, 1.0), (3, 1.0)))) is a trivial corpus of two documents, each with two non-zero feature-answer pairs. (Gensim 2017)

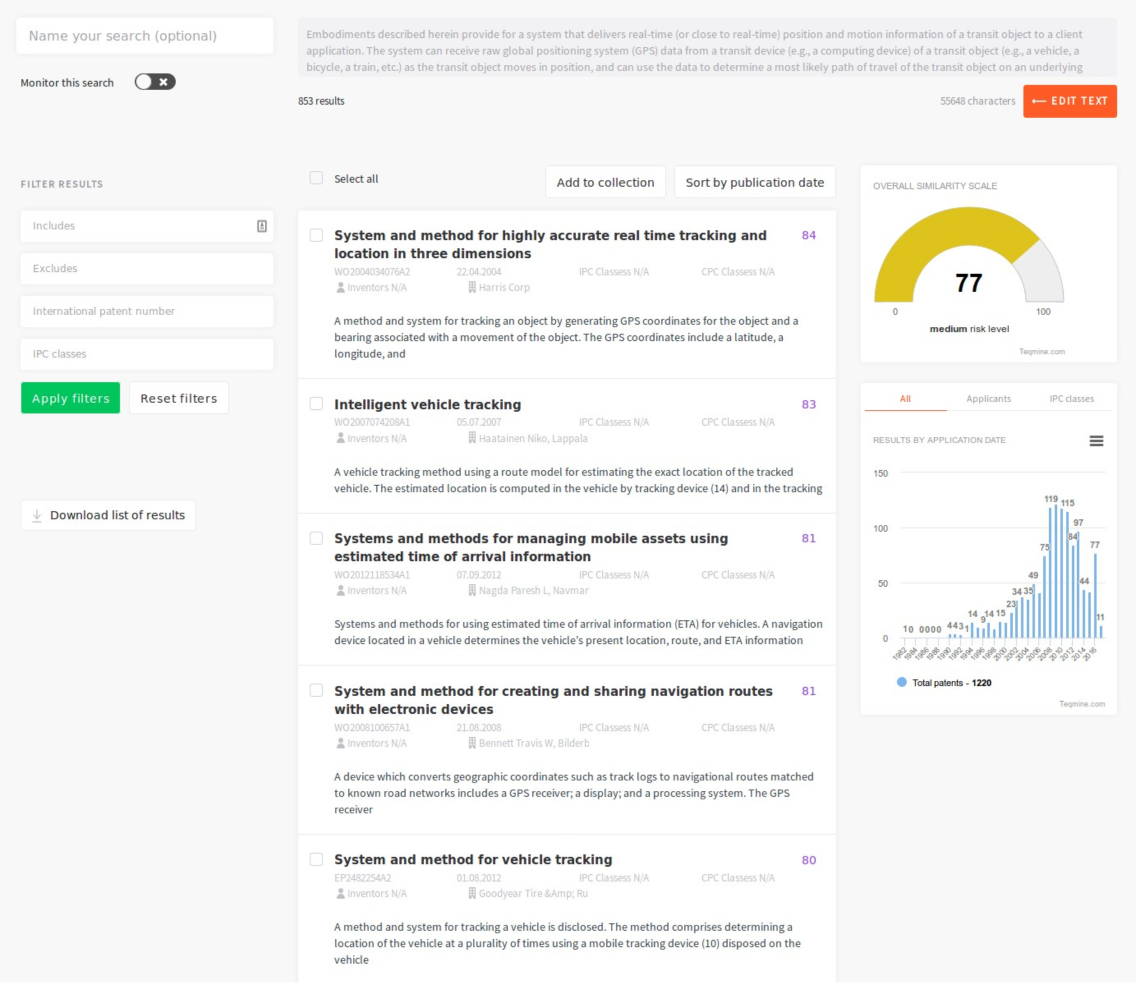
Model We use model as an abstract term referring to a transformation from one document representation to another. In gensim documents are represented as vectors so a model can be thought of as a transformation between two vector spaces. The details of this transformation are learned from the training corpus. The very important factor here is that, by having a trained model from full text patent descriptions, we are provided with the inference process capability to search text using full text as input, users can input full technology description document and the model will facilitate the return of all similar document from the collection, this kind of search results would never be returned by a standard boolean fulltext search, because they do not share any common words with "Human computer interaction."

5.3.3 Searching for similarity

In a conventional search procedures a patent searcher would have a search strategy, in the procedures several steps has to be followed to ensure a discovery of an acceptable result quality from the patent database in question. The search process would include patent classification search, to narrow down the technology classification of the targeted or expected results to come from. Then several rounds of keyword search will be conducted to narrow down the search pool until a reasonable number of results is achieved. This whole process includes the actual reading of some important sections of the description text and claimed claims.

In comparison, with the method that is being utilized in this experiment, a full text description of a patent document Holden and Sweeney (2014) is used as input text for search to the search application running on top of the generated model. Running the Patent document full text through the LDA 5.2.1, produces the similarity index value for each document in the model from corpus in relation to the patent used to search against. The index number is from 0 to 100. For efficiency and relativity purposes, for this test it was set that the top 1000 patent records will be return for more human filtering.

5.3.4 Similarity Search Results



Embodiments described herein provide for a system that delivers real-time (or close to real-time) position and motion information of a transit object to a client application. The system can receive raw global positioning system (GPS) data from a transit device (e.g., a computing device) of a transit object (e.g., a vehicle, a bicycle, a train, etc.) as the transit object moves in position, and can use the data to determine a most likely path of travel of the transit object on an underlying

853 results

55648 characters

□ **System and method for highly accurate real time tracking and location in three dimensions** 84
WO2004034076A2 22.04.2004 IPC Classess N/A CPC Classess N/A
Inventors N/A Harris Corp

A method and system for tracking an object by generating GPS coordinates for the object and a bearing associated with a movement of the object. The GPS coordinates include a latitude, a longitude, and

□ **Intelligent vehicle tracking** 83
WO2007074208A1 05.07.2007 IPC Classess N/A CPC Classess N/A
Inventors N/A Haatainen Niko, Lappala

A vehicle tracking method using a route model for estimating the exact location of the tracked vehicle. The estimated location is computed in the vehicle by tracking device (14) and in the tracking

□ **Systems and methods for managing mobile assets using estimated time of arrival information** 81
WO2012118534A1 07.09.2012 IPC Classess N/A CPC Classess N/A
Inventors N/A Nagda Paresh L, Navmar

Systems and methods for using estimated time of arrival information (ETA) for vehicles. A navigation device located in a vehicle determines the vehicle's present location, route, and ETA information

□ **System and method for creating and sharing navigation routes with electronic devices** 81
WO2008100657A1 21.08.2008 IPC Classess N/A CPC Classess N/A
Inventors N/A Bennett Travis W, Bilderb

A device which converts geographic coordinates such as track logs to navigational routes matched to known road networks includes a GPS receiver; a display; and a processing system. The GPS receiver

□ **System and method for vehicle tracking** 80
EP2482254A2 01.08.2012 IPC Classess N/A CPC Classess N/A
Inventors N/A Goodyear Tire &Amp; Ru

A method and system for tracking a vehicle is disclosed. The method comprises determining a location of the vehicle at a plurality of times using a mobile tracking device (10) disposed on the vehicle

Figure 14. TEQMINE Patent Similarity results

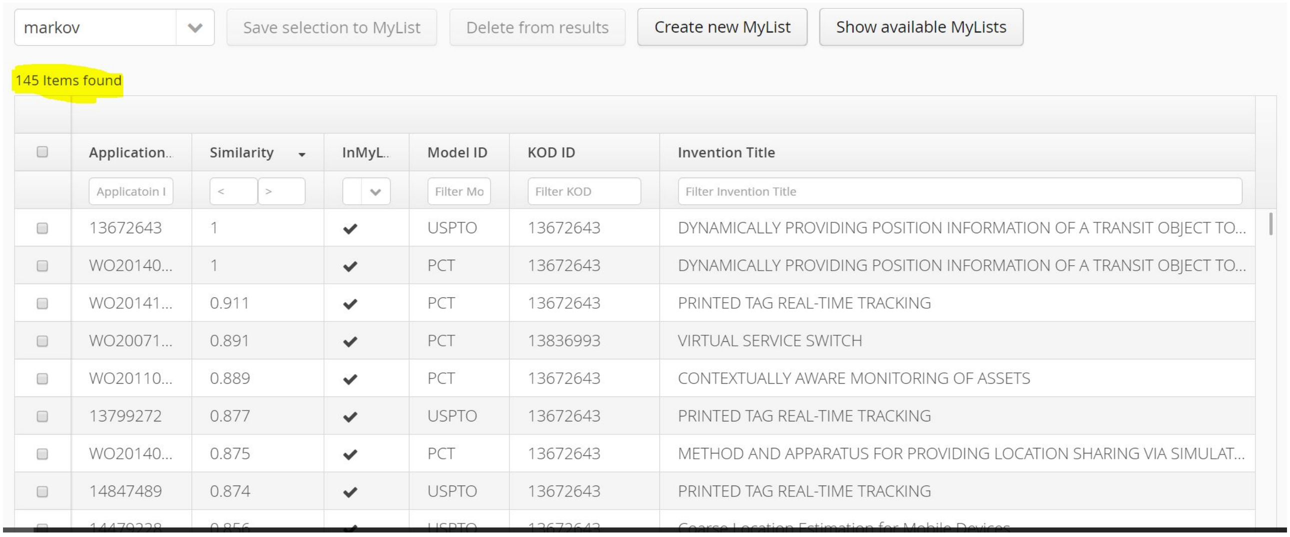
As seen from 14, there are about 843 records return from about 13,000,000 patents. These are the most similar patents comparing to the searched technology description text used for similarity indexing. These patents are not necessarily similar on a word to word bases, but rather their context they are explaining about the same thing, different wording might have been used, but the context within is the same. This is a strong advantage of using unsupervised AI assisted machine learning similarity searching, a keyword search would hardly return results based on context. Even though the meaning is the same between these two patents, the keyword search would not see any match, and thus return a negative answer of not finding any matching documents, and resulting in missing out the very crucial patents that in essence are very similar to the one being searched against.

The results from the similarity model can now be used by a human being to further dive in and evaluate the results, the fact that the user will be dealing with a filtered list of the most relevant documents, adds value to time, accuracy and efficient for the user finding the needed information. With the help of the Similarity Tool from TEQMINE, i was able to sort the results according to different criteria, i.e publication date, application date, assignees, publication office just to mention a few. There is a generic search and advanced search functionality to aid a more control in finding information from the results. The search can focus on a specific content segment of the patent such as, description only, claims only, abstract or patent title. Also there is boolean search capability to further reinforce a complicated search query based on the results.

This search method, does not completely eliminate the human input into searching and refining the results from document pool, but rather it narrows down the scope of where the human input would be more productive and relevant for the search results. Filtering out the most important and similar patents form millions of them, to just about a few thousands, reduces the search and filter cycles for the user, thus improving search result quality, cut down on time spent in searching, and it gives power to non search experts to be able to perform complicated queries with easy and high accuracy.

For example: Using keyword "Markov" on the results list description text, returns about 145 patents with mention of Markov on the description.

There is a tremendous advantage of having these similarity results in terms of analytics and data that can be extracted and queried to produce information that give more details



markov

145 Items found

□	Application..	Similarity	InMyL..	Model ID	KOD ID	Invention Title
	Applicatoin I	< >		Filter Mo	Filter KOD	Filter Invention Title
□	13672643	1	✓	USPTO	13672643	DYNAMICALLY PROVIDING POSITION INFORMATION OF A TRANSIT OBJECT TO...
□	WO20140...	1	✓	PCT	13672643	DYNAMICALLY PROVIDING POSITION INFORMATION OF A TRANSIT OBJECT TO...
□	WO20141...	0.911	✓	PCT	13672643	PRINTED TAG REAL-TIME TRACKING
□	WO20071...	0.891	✓	PCT	13836993	VIRTUAL SERVICE SWITCH
□	WO20110...	0.889	✓	PCT	13672643	CONTEXTUALLY AWARE MONITORING OF ASSETS
□	13799272	0.877	✓	USPTO	13672643	PRINTED TAG REAL-TIME TRACKING
□	WO20140...	0.875	✓	PCT	13672643	METHOD AND APPARATUS FOR PROVIDING LOCATION SHARING VIA SIMULAT...
□	14847489	0.874	✓	USPTO	13672643	PRINTED TAG REAL-TIME TRACKING
□	14479228	0.856	✓	USPTO	13672643	Coarse Location Estimation for Mobile Devices

Figure 15. Search results of keyword "Markov" from description of documents

and views of the invention in a broader perspective. Having a generic traditional search, by classification or keyword search yields results that matches or have a mention of that specific keyword in its text regardless of its context to the topic or the subject relating to the search. This means the results cannot be used to gauge the similarity or relationship among them or against the search term. Similarity results gives back a list of records that has similar context and relate in context with the subject in question. This brings the possibility to find other dimension of relationship between the results. Using similarity service tool from TEQMINE we see an example of analytics possible to generate from the results.

Figure 15 shows a graph of results illustrating number of patents per year in relation to the context of the search document. This information takes the results into a very high level and provide a valuable insight on the trend of the invention or the patent that has been searched against. Here the graph shows that the patenting trend is growing since 2007 there has been an increase in patents or inventions around the technology and in year 2014 there was a spike in patents related to the subject, and it kept going on till 2016. This information is a key indicator that the invention or patent in question is build around a technology that many companies are patenting on. It provides clear vision for the company or inventors in making decisions on what way to move on with invention research, patenting or investing in research and development projects.

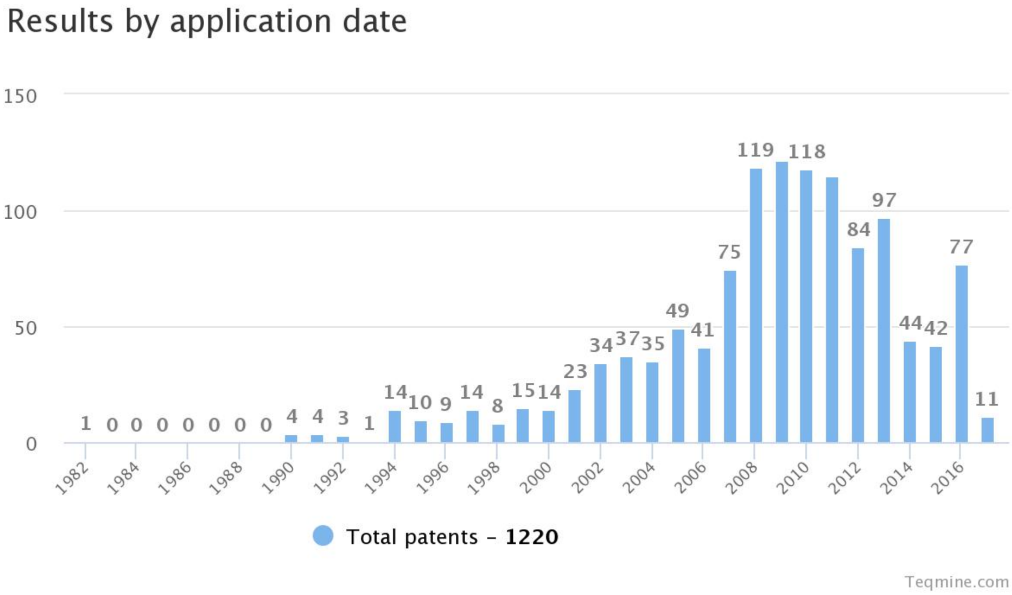


Figure 16. Graph showing number of Patents per year similar to search document

Figure 17 gives an insight into number of inventors by city, this might not be information for an inventor or a company, but it might serve well the organization that needs statics to evaluate which cities has more invention about a certain technology. For strategic planning as well as companies positioning themselves on the technology map, it is important that the management understands the landscape and players in the field, this kind of report is a valuable indicator to pin point potential competitors, it can also help to scout for potential companies to buy or partner with.

5.4 Limitations

This study has several limitations. Most importantly, Semantic annotation and Big data cover a large collection of techniques and algorithms, of which only few have been addressed in detail here. Secondly, this thesis has examined Semantic annotation and Big data only in the context of patent information processing, an area with very specific features and requirements, and with characteristics that do not always allow comparison to other application domains. Other, smaller, limitations include the language specificity of the demonstration. We used English language input text, but perhaps the results could be different in other languages. We focused only at information contained in textual format, but images are of great

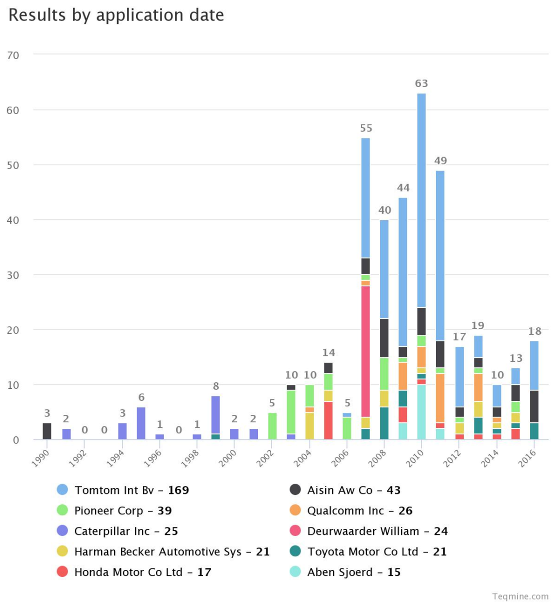


Figure 17. Graph showing top Inventors from the search results

importance for interpreting patent information and beyond of this study.

6 Conclusion

This thesis has analyzed how Semantic Annotation and Big Data methods can be used to improve patent information processing. This is an important issue, because patent information is used for decision making in technology, business, law, government, and other arenas, and accurate, timely and effective information processing is fundamental to reach decision are right. The global collection of patent information captures effectively the evolution real-world technologies, as patents as legal documents include highly accurate information about new-to-the-world inventions and technological ideas, and new patent publications total about 3 million per year currently.

The central work of this thesis has been to consider how the processing of patent information could be enhanced with semantic annotation or with other Big data methods. In detail, this thesis has explored how the following methods can be applied: Semantic annotation, Semantic Web, Ontology, and LDA. Semantic annotations is the method used to tag information in order to provide context to words. Semantic web serves the same function, but in the context of the Internet. Ontology is, in short, a principle to define existing things. LDA was employed to demonstrate how a patent information search could be conducted in a very large collection of patent records.

To test LDA, a database consisting of original data from different Patent Offices was established. This included the collection of about 12 million full-text patents in XML-format, which was then saved into a database. In the second step, the description of these full-text patent records were fed into the LDA modeling. When established, a method to query the said model with any natural-language-text was created, but with the specific objective of supporting people in their efforts to identify similar patents to a patent of interest. This was done by conducting the patent similarity search with an example patent, for which 6,000 results were obtained. In addition to the similarity results, semantic annotation and other big data techniques were used to create contextual information for the reference patent. The information was relayed over the Internet and graphical user-interface.

The experiment demonstrated several advantages over traditional or human-based processing of patent information. The LDA based search did not require sophisticated understanding of what makes patents relevant to the search objective. The whole search process can also be automated, and thereby offer the opportunity to eliminate redundant work and to accelerate discovery. Creation of automated contextual information, such as density of patent landscape, number of patents per year, key technology classes, and listing of most important inventors and firms, additional important information can be created with little or no additional work.

Semantic annotation and other Big data methods to enhance the processing of patent information offer several advantages over the traditional and current methods, as discussed above, to store and retrieve patent information. In addition to the advantages listed above, the accuracy of results can be maintained and even improved as the data amount increases beyond of being anymore possible to handle with human effort.

The methods discussed here are only some of the available methods to enhance the processing of patent information, and other types of method would include self-organizing maps, advances in statistical NLP, and other advanced machine-learning methods.

Bibliography

W3C. 2017. ”Semantic Web Activity Statement”. Visited on May 6, 2017. https://www.w3.org/2001/sw/Activity.

WIPO. 2015. *World Intellectual Property Indicators 2015* [**inlang**English]. Chapter Highlights, 23. Visited on September 4, 2017. http://www.wipo.int/edocs/pubdocs/en/wipo_pub_941_2015-part1.pdf.

———. 2016a. ”World Intellectual Property Organization”. Visited on October 4, 2016. http://www.wipo.int/patents/en/faq_patents.html.

———. 2016b. ”Global Patent Applications Rose to 2.9 Million in 2015 on Strong Growth From China; Demand Also Increased for Other Intellectual Property Rights”. Visited on October 4, 2017. http://www.wipo.int/pressroom/en/articles/2016/article_0017.html.

———. 2016c. ”Reasons for Patenting Your Inventions”. Visited on October 9, 2016. http://www.wipo.int/sme/en/ip_business/importance/reasons.htm.

———. 2017. ”IP and Business: Patent Information: Buried Treasure”. Visited on October 8, 2017. http://www.wipo.int/wipo_magazine/en/2005/01/article_0003.html.

Amber Stubbs, James pustejovsky. 2012. *Natural language Annotation fro Machine learning.* O’Reilly Media Inc. ISBN: 9781449332693. https://www.safaribooksonline.com/library/view/natural-language-annotation/9781449332693/.

Balagopalan, A. 2012. *Improving Topic Reproducibility in Topic Models.* University of California, Irvine. ISBN: 9781267246806. https://books.google.fi/books?id=mtPlnQAACAAJ.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. ”Latent Dirichlet Allocation”. *J. Mach. Learn. Res.* 3 (): 993–1022. ISSN: 1532-4435. http://dl.acm.org/citation.cfm?id=944919.944937.

Bouchoux, D.E. 2016. *Intellectual Property: The Law of Trademarks, Copyrights, Patents, and Trade Secrets.* Cengage Learning. ISBN: 9781305948464. https://books.google.fi/books?id=Scu5DQAAQBAJ.

Chen, Edwin. 2017. "Introduction to Latent Dirichlet Allocation". Visited on October 4, 2017. http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/.

Clarivate Analytics. 2017. "Patent information: Definitionof a patent". https://support.clarivate.com/DerwentInnovation/s/article/ka1390000004bBIAAY/Patent-information-definition-of-a-patent?language=en_US.

EPO. 2016. "European Patent Office". Visited on December 7, 2016. https://www.epo.org/about-us/office.html.

Espacenet. 2017. "US2003173072A1". Visited on May 6, 2017. https://worldwide.espacenet.com.

Gensim. 2017. "Introduction to Gensim". Visited on October 8, 2017. https://radimrehurek.com/gensim/intro.html.

Google Inc. 2017. "Google patents advanced search". Visited on October 1, 2017. https://www.google.com/advanced_patent_search.

Harold, E.R. 2004. *XML 1.1 Bible.* Bible. Wiley. ISBN: 9780764569302. https://books.google.fi/books?id=4qg5GOvqd-cC.

Holden, P.P., and M. Sweeney. 2014. *Dynamically providing position information of a transit object to a computing device.* US Patent App. 13/672,643. https://www.google.com/patents/US20140129135.

Information resources management association. 2016. *Big Data: Concepts, Methodologies, Tools, and Applications.* O'Reilly Media Inc. ISBN: 9781466698406. https://www.igi-global.com/book/big-data-concepts-methodologies-tools/140960.

Intellectual Property Law, American Bar Association. Section of. 2010. *What is a Patent?* ABA Section of Intellectual Property Law. ISBN: 9781604428056. https://books.google.fi/books?id=IsJJ0ehiC4EC.

Jolly, Adam, and Jeremy Pholpott. 2009. *European Intellectual Property management: Develping, managina and protecting your company's intellectual property.* Kogan Page.

Landes, William M., and Richard A. Posner. 2003. *The Economic Structure of Intellectual Property Law. Harvard University Press.* where published: Harvard University Press.

LexixNexis. 2017. "lexisNexis". Visited on October 4, 2017. http://intl.lexisnexisip.com/.

Lupu, Mihai, Katja Mayer, John Tait, and Anthony J. Trippe. 2011. *Current Challenges in Patent Information Retrieval.* 1st. Springer Publishing Company, Incorporated. ISBN: 9783642192302.

Mario Cimoli, Benjamin Coriat, and Annalisa Primi. 2009. "Intellectual pproperty and industrial development". 1.

McEnery, T., and A. Wilson. 2001. *Corpus Linguistics: An Introduction.* Edinburgh University Press Series. Edinburgh University Press. ISBN: 9780748611652. https://books.google.fi/books?id=nwmgdvN%5C_akAC.

MySQL. 2017. "MySQL". Visited on October 8, 2017. https://www.mysql.com/.

NOLO. 2016. "Intellectual Property Law firm". Visited on October 4, 2016. http://www.intellectualpropertylawfirms.com/resources/intellectual-property/patents/the-usual-costs-filing-a-patent.htm.

Ontotex. 2016. "Semantic Annotation". Visited on January 3, 2016. http://ontotext.com/products/ontotext-semantic-platform/semantic-annotation/.

Oren, Eyal, Knud Hinnerk M Oller, Simon Scerri, Siegfried Handschuh, and Michael Sintek. 2016. "Value of patent information". Visited on October 12, 2016. http://www.siegfried-handschuh.net/pub/2006/whatissemannot2006.pdf.

PCT. 2016. ”PCT_WIPO”. Visited on December 7, 2016. http://www.wipo.int/pct/en/.

Pustejovsky, J., and A. Stubbs. 2012. *Natural Language Annotation for Machine Learning.* Oreilly and Associate Series. O’Reilly Media, Incorporated. ISBN: 9781449306663. https://books.google.fi/books?id=QtzmqamXxx4C.

Ramsay, G. 1870. *Ontology, Or, Things Existing.* Walton. https://books.google.fi/books?id=dSs-AAAAYAAJ.

Researh and Development Magazine. 2017. ”Research and Development Magazine”. Visited on October 4, 2017. https://www.rdmag.com/.

SAS Institute Inc. 2015. ”Big Data History and Current Considerations”. Visited on December 25, 2015. http://www.sas.com/en_us/insights/big-data/what-is-big-data.html.

Semantics, Cambridge. 2017. ”Semantic web”. Visited on September 28, 2017. https://www.cambridgesemantics.com/blog/semantic-university/intro-semantic-web/.

Slimani, Thabet. 2013. ”Semantic Annotation: The Mainstay of Semantic Web”. *Researchgate* 1:4–5. doi:10.7753/IJCATR0206.1025.

Staab, S., and R. Studer. 2013. *Handbook on Ontologies.* International Handbooks on Information Systems. Springer Berlin Heidelberg. ISBN: 9783540247500. https://books.google.fi/books?id=uTwDCAAAQBAJ.

Tang, Jie, Duo Zhang, Limin Yao, and Yi Li. 2009. ”Automatic Semantic Annotation Using Machine Learning”. *IGI Global* 1:1. doi:10.4018/978-1-60566-028-8.ch006.

Technologies, Gridlogics. 2016. ”Patent analysis organization”. Visited on October 8, 2016. http://patentanalysis.org/worldwide-patent-application-filing-trends/.

TEQMINE Analytics Ltd. 2017. ”TEQMINE Patent Similarity Service”. Visited on October 8, 2017. http://teqmine.com.

USPTO. 2016a. ”USPTO”. Visited on May 6, 2017. https://www.uspto.gov/.

———. 2016b. ”USPTO Patent office”. Visited on December 7, 2016. http://www.uspto.gov.

———. 2016c. ”USPTO Patent searching”. Visited on February 14, 2016. http://www.uspto.gov/video/cbt/ptrcsearching/.

———. 2017a. ”How to conduct a preliminary U.S. Patent search A step by step strategy”. Visited on May 4, 2017. https://www.uspto.gov/video/cbt/ptrcsearching/.

———. 2017b. ”Types of Patents”. Visited on April 19, 2017. https://www.uspto.gov/web/offices/ac/ido/oeip/taf/patdesc.htm.

———. 2017c. ”USPTO Patent Full-text and image database”. Visited on May 6, 2017. http://patft.uspto.gov/netahtml/PTO/search-bool.html.