International Journal of Innovation and Applied Studies
ISSN 2028-9324 Vol. 5 No. 3 Mar. 2014, pp. 160-191
© 2014 Innovative Space of Scientific Research Journals
http://www.issr-journals.org/ijias/

A Review assessing the "used in the art" Intellectual Property Search Methods and the Innovation Impact therewith

Serge Rebouillat[1] and Damien Lapray[2]

[1]Currently at DuPont Int., CH1218 Geneva, Switzerland

[2]Currently at EPFL, CH1015 Lausanne, Switzerland

ABSTRACT: Is Intellectual property (IP) central to innovation or is innovation central to IP? Univocally, patent valuation starts before drafting. Then IP is a valuation step such as innovation is. Equally "innovation is not the idea, but what you do with it". Then can ideation be engendered by artificial means?
Nearly 60 years have passed since the birth of artificial intelligence and the initial dream of a machine possessing the full-range of human cognitive skills still belongs to science-fiction. However software using artificial intelligence are more and more present in our daily life. This is particularly true in the domain of information retrieval. Today's amounts of data one can access through diverse media necessitate the use of "clever algorithms" to find relevant material. This applies even more to the domain of IP.
Companies' whole innovation strategy relies significantly on the analysis of patents, scientific publications and other IP documents. It is therefore crucial for them to extract in the most efficient way the best of the available information. The purpose of this paper is to give strategists, researchers, business analysts… the possibility to understand what is behind the tools they use.
Furthermore we demonstrate that the current dominant search technologies cannot fully support today's economic challenges and that better solutions are at reach. We also review the transferable technologies and methodologies. This inventory has also for objective to draw attention to some promising directions to follow and advocate for.
For example the semantic and image combined mining, comprising interpretation, ideation and reinvention, is a tremendous opportunity to boost the innovation process via the exploration and exploitation of the IP and NPL (non-patent literature); especially when performed by enlarged teams comprising the engineer, scientist, IP strategist, business model expert…
A significant literature base is reviewed along with examples probing the reality and revealing innovation opportunities.

KEYWORDS: Innovation, collaborative, Collaboratory™, adjacent technology analysis, ATA©, IP strategy, semantic analysis, image analysis, artificial intelligence, reverse engineering, reinvention, neo-retro-innovation, iPad®.

Disclaimer: This article is primarily for educational purposes. Selected cases are strictly illustrative. Neither the author nor the illustrator assumes any liability for any errors or oversights, or for how this article or its contents are utilized or interpreted, or for any consequences resulting directly or indirectly from the usage of it.
This assessment is intended to educate and raise awareness of some of the complex issues that surround the intellectual property in the field of knowledge extraction from the about 80 million patent documents available, and to assist in the development of practical skills for dealing with inventions and innovation in general. It does not seek to provide legal, managerial or technical advice on intellectual property related law as such. For any guidance, legal or any other, seek advice from the appropriate professionals; this study can by no mean substitute for expert legal, technical and managerial advice.
The opinions expressed by the writers in this article do not necessarily represent the viewpoints of the companies the author are employed by.

Corresponding Author: Serge Rebouillat (Serge.Rebouillat@dupont.com)

1 INTRODUCTION AND MOTIVATIONS

Innovation is a growth engine

This study combines and complements elements of the presentation at the Geneva Corporate Innovation Forum of October 6th, 2011 and the plenary lecture given at the International Symposium on Green Chemistry held in La Rochelle on May 21-24th, 2013. It also constitutes a logical continuation of the latest work by Rebouillat [1] (Rebouillat, 2013) untitled "A Science & Business Equation for Collaborative Corporate Innovation" whereby "Business Strategy, IP Strategy and R&D Strategy are the base of an all-in-one Business Model".

In the frequently cited open-innovation context, the business engineer would have to be a Creative-Connoisseur and a Competent-Communicator. But also should conform to almost any profile corresponding to any combinatorial arrangements of the four word roots, 4C©, as depicted on Figure 1.

This reality is taking place and requires the implementation of specific agreements and tools, which can help the sharing of confidential data and the exploitation of public data available by the 100's of million. IP in general and more specifically patents are seen as a monopoly, thanks to a legal situation, which at some point of time became a deterrent; patent is evolving as a mean to boost innovation, thanks to easier access to the 80 some millions of patents reachable by the engineer, scientist, IP strategist, business model expert.

Therefore it might be useful to separate the legal aspect of patent and the technology part; grossly distinguishing the specifications part of the patent document, being a technology and science rich arena for the IP technology strategy expert, and, the claims being the fertile area exploited by the patent lawyer performing the freedom to operate exercise for example.

The appraisal of the patent adjacent technology content via adjacent technology analysis, ATA© [1] (Rebouillat, 2013), led the way to a better characterization of the patent strength. The recent introduction, on top of the patent count per year, of several trademarked, although already existing criteria, such as "Science Strength™" and "Technology Strength™", is a logical step.

The further adoption by patent specialized boards and "patent mining" companies of such criteria led to the recognition of scientifically and technology driven companies such as the one performing business for more than 212 years with a real growth engine based on innovation [2] (www.prweb.com/releases/2012/3/prweb9280362.htm).

Concerning classifications, the world number one company in terms of portfolio count in 2012 is ranked number 3 in the brand value classification; the company ranked number 6 in the brand value classification in 2012 is also ranked number 6 in terms of the portfolio count ranking. Three of the 6 first patent portfolio counts are part of the 10 first valued brands.

Brand and patent count classification seem to get closer as patent portfolio average quality improves with time, i.e. a better alignment of patent strategy with business strategy. The race for "patent count" might be in obsolescence. But the race is still on, slowly shifting towards a "quality rather than quantity" based race.

Undoubtedly patent, IP in general, science and technology in an open innovation context are to lead growth as long as integrated in a larger business model frame of work.

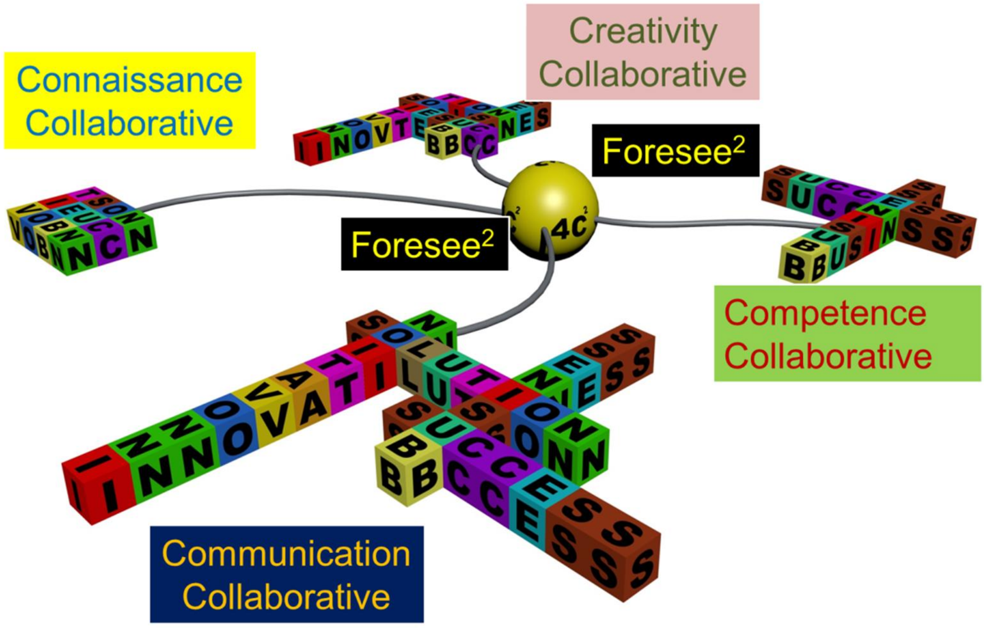

Fig. 1. Foresee Square, 4C²©

In parallel an explosion of applications and publications of various kinds via traditional and e-media is taking place.

From 2008 to 2010 WIPO received about exactly 5,000 new patent applications per day, and Medline received about 110 new articles per day in 2010.

The explosion of design patents in China, more than 500,000 in 2012, i.e. 10x the USPTO or the EPO numbers, has been reported at several occasions by the specialized press. Also described is an “All-time high for activities of the European Patent Office in 2012” which tend to put patent growth in a worldwide context [3] (www.epo.org/about-us/annual-reports-statistics/annual-report/2012/statistics-trends.html).

Within the patent context of novelty, specialized search engines are being made available to provide links to older archived versions of a webpage such as the “way back in time” machines dedicated to electronic archive detection. This is a welcome initiative as underlined below.

To the complexity of finding the right information on time is added the complexity of interpreting; the two are for obvious reason closely linked.

<u>Starting from History review… What’s next?</u>

The concept of “information retrieval” as we know it today was introduced by Calvin Mooers from the MIT in 1950 following the birth of digital computers [4] (Garfield, 1997). The potential benefit of these machines as automated searching systems was evident from the beginning and scientists started developing retrieval and indexing algorithms early on. First applied to scientific and medical literature for military purposes, information retrieval has become nowadays an essential activity of our modern life style. The incredibly large amount of data one can access for example on the web, necessitates the use of more sophisticated tools. According to a survey from netcraft.com in October 2013 the web contained over 189,176,770 active websites, a 942 % increase in 10 years [5] (Netcraft.com, 2013). Companies such as Google Inc., Yahoo! Inc. and others have built their core expertise and reputations on indexing and retrieving information.

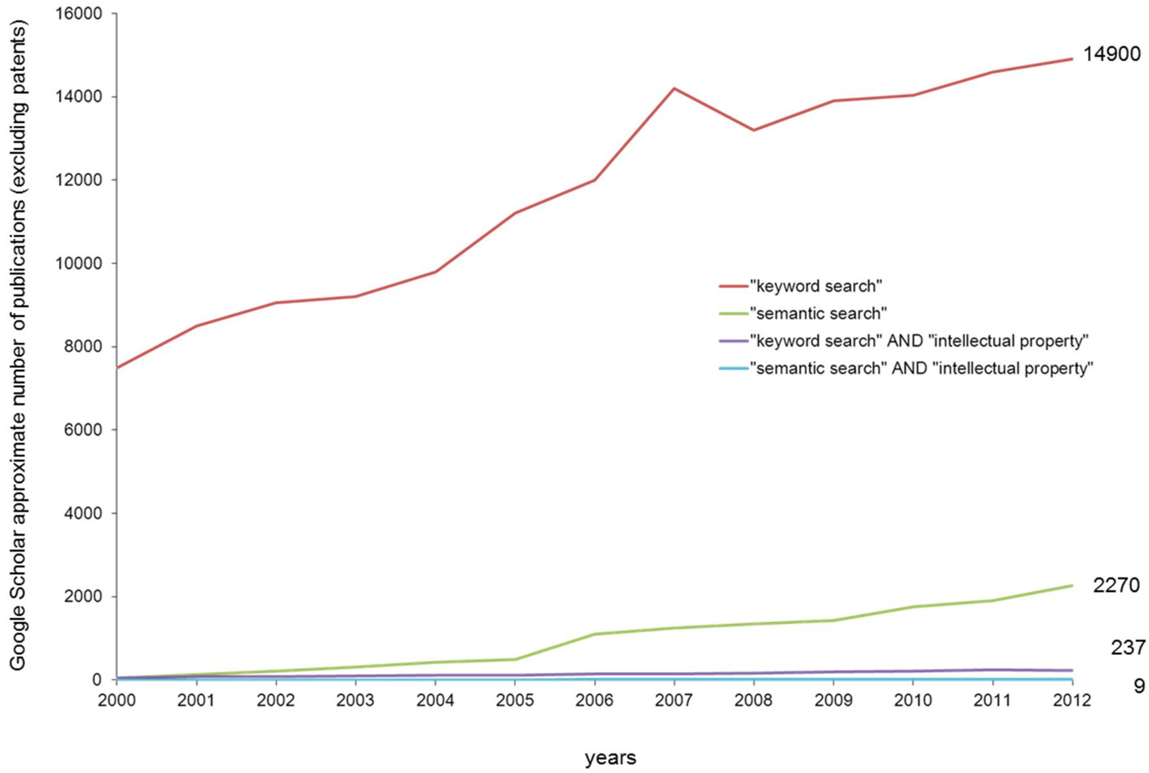

Fig. 2. Search Trends

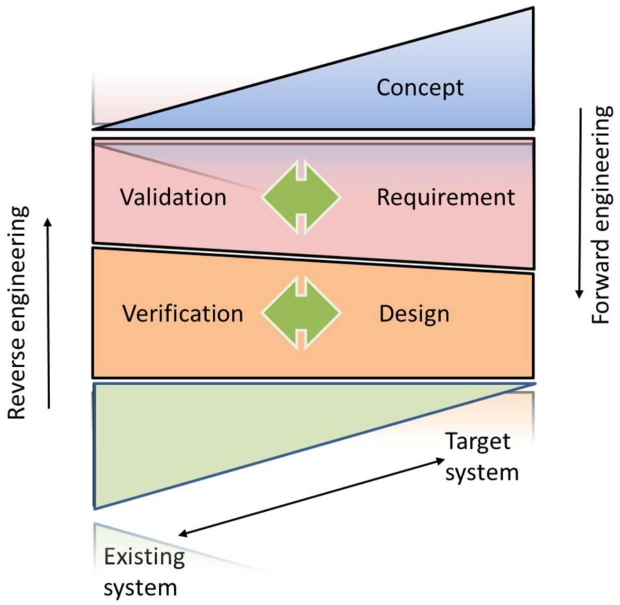

Fig. 3. Reverse Engineering & Reinvention

If the accuracy of search technologies is important for daily requests, it becomes critical when the result directly impacts the chances of survival of a company against an ever growing competition. IP is at the centre of the economical battle and its global analysis through patents, publications, reports, is essential for technological developments and strategic planning. Data have grown enormously in the past few years and traditional search methods may struggle to further cope with the amount and complexity of the documents to be treated. In 2012, the US Patent and Trademark Office granted more than

275,000 patents and Elsevier, the biggest scientific and medical publisher, accepted more than 250,000 articles [6] (Elsevier, 2013); (Figure 2) with more details in section 3.1.. Infringements are also strongly dependent on the ability of search engines to give the right answers to users' queries. In 2012 the number of patent lawsuits filed in US increased by almost 30% compared to 2011 to over 5,000 [7] (PwC, 2013). It is therefore not surprising to see that IP search and companies behind these services are the centre of a lot of attention.

One shall not exclude the perfectly legitimate reinvention methods, as shown on Figure 3. The House-of Quality, the Quality Function deployment (QFD), the Voice-of-Customer (VOC), the Theory-of-Constraints (TOC), the integrated Theory of Inventive Problem Solving (TRIZ), Six Sigma, the 4A method (Accountability-Accreditation-Assessment-Articulation), the Resource-Activity-Results Method (RAR), etc. have paved the way to promote best practices, to boost and to secure the innovation processes and routes to market of multigenerational products and processes.

Most of these approaches, individually or integrated in innovation management processes, have now got Wall Street's attention and became prerequisite to a good design of a portfolio of products. Those methods rely heavily on integrating the most pertinent knowledge.

However there seem to be some expectations divergences between users, software developers and lawyers regarding the finality of IP search. Search companies have for many years favoured a form of search based on keywords, something we practice every day when using Google. Despite some advantages that will be discussed in this review, this technology is not necessary appropriate for patent search and rely mainly on searchers' expertise and on the size and quality of databases.

A particularity of patents text is the jargon used to describe objects making it extremely difficult to write accurate queries. In fact, the word "computer" can easily be replaced by "a system having a storage for storing data, an output device to display information, a terminal for entering information, and a component that modifies the input data and controls flow of data between different parts" [8] (Lupu, 2011). The human brain on the contrary to machines is perfectly able to understand that this latter list of words is in fact a computer without the presence of the actual keyword "computer". In order to do so it uses semantics, i.e. each word has a meaning and is used in a context giving only one or a very limited number of possible interpretations. Semantic technologies take their roots in linguistic where each word is more than just a string of letters. Because the human brain is not able to treat the amount of information necessary to perform an exhaustive search, powerful algorithms using artificial intelligence (AI) have to be used [9] (WIPO, 1991). AI is the key to potentially mimic what the brain can do effortlessly but on a much bigger and faster scale. These solutions exist already in search software but their implementation to professional IP search engines has been delayed. Despite numerous publications and attempts by scientists to bring this technology to the application level [10], [11] (Magerman, Looy, & Song, 2009; Spyropoulos & Botsivaly, 2009) only few companies implemented these solutions.

Another activity mastered by humans is image detection. It takes no more than 170ms for the brain to go through the whole process of face recognition [12] (Liu , Harris, & Kanwisher, 2002), one of the most difficult visual tasks to perform. This complex procedure is not trivial to achieve with a computer and has been barely touched in IP search. Because patents and publications are full of diagrams and technical drawings, image recognition technology which exists and is used in other fields [13] (Zhao & Chellappa, 2003), is a necessary step for a search to be complete. The semantic and image combined mining, comprising interpretation, ideation and reinvention, is a tremendous opportunity to boost the innovation process via the exploration and exploitation of the IP and NPL (non-patent literature); especially when performed by enlarged teams comprising the engineer, scientist, IP strategist, business model expert…

The legal aspect of intellectual property is another limitation for a global, collaborative system. Learning is a necessary process if one wants to acquire knowledge and this also applies to AI technologies. However in an open innovation scenario, the absence of interactions between the queries made by one strategist in company A and one in company B make it even more complex to develop the ultimate *in silico* researcher "driven" by an engineer, a scientist, an IP strategist, a business model expert.

One aim of this review is to make it easier for IP search technology users to understand what is behind the black box they use every day; to give an overview of the new tools available inside and outside the field; to assess where we stand and where we go; and to highlight some of the most needed developments and possible future directions.

From an open innovation perspective and framework

Another aim of this review is then to promote and boost innovation beyond the fear associated with the paradigm connected to the word "open" innovation and the ever increasing and contrasting need for trade secret management. Let's underline some principles that may alleviate part of this archetype and help the reader through an appeasing journey, along the proposed review trail.

The IP protection insurance and the innovation strategy are of evergreen vital importance. Close coaching and consultancy beyond administration and legal attention, is ideally performed by the innovation strategist well versed in IP matters, with a broad experience and knowledge of business and technology functions – a rather new and emerging role requiring strong mediation skills. In a joint development the exploration phase is likely, in most case, the most decisive and premonitory one for the future overall success. The exploration phase can be defined as per following [1] (Rebouillat, 2013).

The parties explore the possibility of working together, a mutual one-way or two-way confidentiality agreement being in place. Sample and/or material transfer agreements may be used in this incubation phase.

- Identification of interest areas, business and cultural fit is necessary;

- A clear understanding of what each party brings, such as technology expertise and areas of interest shall be established;

- An open discussion mind-set,

- and an agreement on vision for success shall be reached.

2 INTELLECTUAL PROPERTY SEARCH

2.1 IP MATERIAL: THE FIRST CHALLENGE

Some of the most apparent benefits of patent documentation include:

- description of inventions in a way that is aimed at facilitating their reproduction in practice;
- coverage of material which is, by definition, genuinely new and not earlier available to the public;
- the matter of interest is almost universally categorized by a single international cooperative classification system;
- the documentation access is free of charge;
- the patent document appears in a common format; and
- cross-referencing between documents is generally handy;

Patent information is appreciated for a multitude of purposes. This can rank from a foremost study on technological advances in a particular area (e.g. current status of research into AIDS therapy), or a very specific single examination (e.g. is a patent still in force in a given country?). The strategies and methods used for searching patent information therefore vary extensively, contingent on what the information is to be looked for.

The International Patents Classification, IPC, and its cooperative version CPC, are the universal patents classification systems, which are administered by WIPO, EPO and USPTO. Those classifications can therefore form the foundation of a search of the patent literature and quite a few databases may be explored using them.

Frequently made comments include the following drawbacks:

1. “the IPC has insufficient subdivisions and some of the specific subdivisions are already full;
2. IPC revisions are slow and sometimes incomplete in technological fields of rapid changes;
3. it is not specialized for biotechnology inventions;
4. classification into the IPC is not consistently applied in different Patent Offices;
5. use of the IPC for searching requires detailed understanding of technology and the IPC; and
6. the information classified into the IPC is, of course, limited to patent documents.”

We think it is then important here to clarify what IP search is or should be if one wants to use it not only for document retrieval purposes but as a generator of innovative ideas. Furthermore there seems to be discrepancies between the material uses by professional searchers and what is used for the development of new techniques. When reviewing the literature one quickly notices that a great number of publications [14], [15], [16], [17], [18], [19], [20], [21], [22], [23] (Wang et al., 2011; Yoon et al., 2011a, 2011b and 2012; Trappey et al., 2009; Tang et al., 2012; Jessop et al., 2011; Archibugi et al., 1996; Wanner et al., 2008; Wu et al., 2012) or patents [24], [25] (US5,774,833; US8,161,049B2) related to the development of search tools on IP limit their field of application to patents. There are several good reasons for that. Whatever search one performs the availability of documents is the key factor and patents are freely available through the web. They do also have the advantage of a certain uniformity of format which facilitates some aspects of a search. However the IP domain would suffer a lot if patents, despite their crucial role, would be the exclusive source of inspiration. IP search should incorporate scientific publications, websites, conferences, technical disclosures, defensive filings, brochures and all information that can describe the subject of interest [8], [26] (Lupu, 2011; Codina et al., 2008). Here start the problems, e.g. the total lack of uniformity and sometime the difficulty of accessing such documents. Scientific publications contain a great deal of precious information fundamental for designing future innovations. However scientists and public institutions rarely consider the patenting of knowledge as a natural research step. Moreover, a big part of these publications belongs to publishing companies which, for most, do not provide free access to the manuscripts and their cost can often be a limitation. Even if all these documents would be freely available their retrieval would still be a challenge due to the absence of a global homogeneous indexing system. Some open source search engines, such as FPO, do have a NPL interface that provides scientific articles; this deserves to be underlined. This idea was proposed at the end of the 90s as the semantic web [27] (Greenbaum & Gerstein, 2007), this issue will be discussed later in chapter 7.2. Because of the aforementioned difficulties in broadening the field of IP search, a large part of this review will deal with technologies developed for patent analysis. But we believe that this information apply to all IP documents.

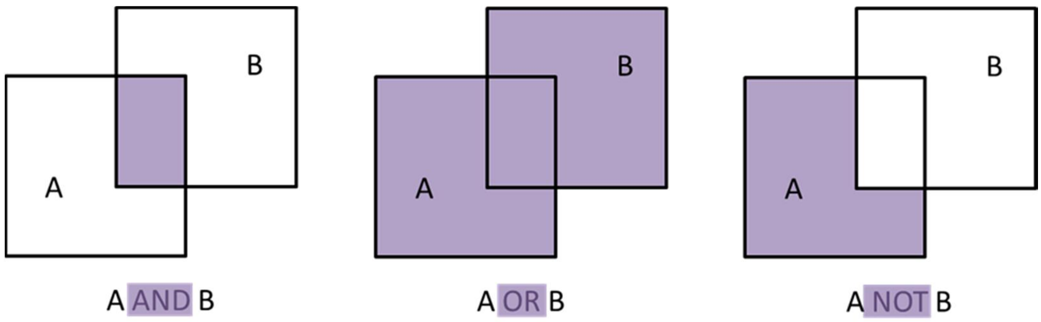

Fig. 4. Boolean Search

2.2 AVAILABLE TOOLS

Before going further into the technical details of IP search it seems important to give an overview of the available technologies. The list provided below does not pretend to be exhaustive but intends to give a broad vision of the current market.

We can divide IP search engines into two broad categories: the free and the paid ones. The free solutions can as well be subdivided into two subcategories which are the systems focusing almost exclusively on patents and those combining it with scientific publications and conference reports. The first category includes the five main institutions granting patents namely the United States Patent and Trademark Office (USPTO), the European Patent Office (EPO), the State Intellectual Property Office of the People's Republic of China (SIPO), the Korean Intellectual Property Office (KIPO) and the Japanese Patent Office (JPO) as well as the World Intellectual Property Organization (WIPO) and data miner engines such as Google Patent Search, The Lens [28] (Jefferson et al., 2013), and Free Patent Online. They are all based on keyword search technologies. The strength of these systems relies on their huge indexed databases, the possibility to look for patent applications and documents are updated regularly. These are interesting tools when one wants to explore quickly a topic or search for a specific patent. The main drawbacks of these software are the restriction of search to patents mostly and the lack of other functionality than the basic Boolean keyword-search and associated tools, with some exceptions, for example The Lens (Figure 4). They can be assimilated to public libraries which offer a logical indexing and search system if one knows what he is looking for. In the category of free solutions, the second subcategory would include engines such as Google Scholar and are pretty similar to the first subcategory in terms of design and functionalities. They do have access to scientific publications, however it is often restricted to their metadata and abstracts and the patent database is much smaller than what the other free technologies offer.

Similar to the free solutions, paid ones can be subdivided into two main subcategories. The first one includes the free-like technologies. Keyword search based, they differ from the free software only from what they offer after the initial search. They are more strategy based, providing statistical tools, similarity measurements and the possibility to visualise in a more friendly way the results of a search. In this subcategory we find software such as DiscoverIP or Delphion. Some software from this subcategory integrate interesting options such as the Examiner's Application Search Tool (EAST) with an image search. However this latter is not available for users.

The second subcategory of paid technologies is based on linguistic tools to search through patents, scientific publications and other relevant documents. We can cite software such as IHS Goldfire, Pantros IP, IPCentury AG, SemIP.com, LexisNexis TotalPatent and TextWise IP.com. They have the big advantage to combine semantic and keyword searches, many statistical tools, innovation oriented analysis trees, clustering and so on. Some of them propose free patent search solutions such as Pantros IP with freepatentsearchsite.com. Despite their relative youth and not yet proven full capabilities they might soon become essential.

To date no global solution exists that would carry the researcher from the initial question up to an answer not requiring users to inconveniently juggle with several tools.

2.3 THE SEARCH SEQUENCE

2.3.1 TEXTUAL INFORMATION

Analysing textual data from IP documents follows the same sequence as any other text mining.

Advanced search tools for patent professionals being still in their infancy [19] (Tang et al., 2012) the first step of the retrieve sequence is very often followed by long and tedious manual experts’ analysis. Emerging technologies are becoming available that could change this process.

The first step corresponds to the selection of relevant documents using an information retrieval engine [29] (Rzhetsky, Seringhaus, & Gerstein, 2009). This step is absolutely critical in the field of IP were being exhaustive is one of the key for success. The rest of the analysis depends almost entirely on it and is, as it will be developed below, still one weak point of the whole process. This is exclusively done using keyword-based search tools. It is important to notice here that most of free technologies do not propose anything after that particular point.

The second step is the named-entity recognition [29] (Rzhetsky, Seringhaus, & Gerstein, 2009). The aim is to scan the text using specialised software and extract various entities (objects, concepts, and symbols) in each sentence. The ideal system would consistently identify an individual entity even though it may have different names and acronyms using machine-learning tools, and dictionaries. This step already necessitates the use of linguistic tools able to catch the meaning of words or use mathematical abstractions. This will be developed in chapter 4.

The third step is information extraction which creates relationships between the extracted entities [29] (Rzhetsky, Seringhaus, & Gerstein, 2009). This step is field oriented; it matches specific requirements and problem of the user using, for example, domain specific ontologies (e.g. genomic, nanotechnologies, chemistry …). An Ontology being: ‘’a taxonomy with multiple, precisely defined links between the items, that represents knowledge as a set of concepts and their relationships’’ [8] (Lupu, 2011). There exists no global ontology; every field has to generate its own. The importance of developing ontologies will be discussed in the chapter 4.2.

The fourth and last step of a search structure is linking a user query and the extracted information. This is where mapping results become relevant to simplify the lecture of the data for a better outcome.

2.3.2 PATENTS METADATA AND CLASSIFICATION CODES

Textual information from the main document is not the only parameter allowing IP analysis. So-called patents' metadata such as title, abstract, publication date, applicants are often used in IP information retrieval. These information are useful to narrow a search to a specific time period, to follow a specific company or inventor or to find prior art documents. In addition patent offices allocate some classification coding to each patent such as the International Patent Classification (IPC), the European Classification (ECLA), the United States patent classification (USPC), the Japanese File Index and F-Term (FI/F-Term) classification and some other minor systems. The IPC is a hierarchical system that replaces or supplements national classification systems since 1971 [8] (Lupu, 2011). Each patent application is allocated a label by the patent examiner based on their technological area. Today it is composed of 8 sections, 129 classes, 639 subclasses, 7,314 main groups, and 61,397 subgroups. It has some potential advantages for information retrieval purposes. It could provide a solution to overcome the problem of heterogeneous terminologies used to describe similar concepts. It is already in a machine-readable format and allows for fast retrieval of patent independently of the language [30] (Lupu, 2013). This code is often used to narrow a search to a homogeneous technical categories, for mapping and collecting prior art publications [31], [32], [33] (Chiu, Hong, & Chiu, 2011; Dirnberger, 2011; Leydesdorff, Kushnir, & Rafols, 2012). Though this seems a very useful piece of information it carries some deadly issues. First of all it has been created by patent institutions for patent examiners to rapidly find if there are similar inventions to the patent under consideration. It is therefore tuned to the need of patent examiners that are not necessary compatible with those of professional search users. Furthermore some examiners might lack sufficient knowledge to appropriately assign a proper IPC code to the patent under consideration. In addition the rapid evolution of technologies and technological terminologies would require a similar evolution speed of classification labels [34] (www.intellogist.com/wiki/IPC_Classification_System). The assignment of the code is subject to human errors that could be potentially overcome by the use of automated patent classification. However these automated methods meet the same challenges as the IP search engines due to the unstructured nature of the data [30] (Lupu, 2013). Finally not every country use the IPC system, and from those using it, such as the US and Japan, they rely more on their own classification codes.

The issues associated with a search using classification codes illustrate once more obvious expectations divergence between the different actors of the IP field and this is detrimental to the overall quality of IP search.

2.3.3 DIGITAL MAPS: VISUALIZATION AND ANALYSIS TOOLS

Information retrieval is not complete without the visualization tools that allow to intuitively understand the large amount of extracted data and to be the foundation of further analysis. It is the topic of a research field called visual analytics [8] (Lupu, 2011). “Appropriate display of clusters points can give the analyst an insight that it is impossible to get from reading tables of outputs or simple summary statistics” [35] (Polanco, François, & Lamirel, 2001). Visualization techniques are numerous and vary depending on the type of data to be displayed [36] (Keim, 2002). Clustering methods are the most extensively studied and have been applied in many scientific fields [37] (Jain, 2010). Clustering techniques are used in neurosciences [38], [39], [40] (Dupret, O’Neill, & Csicsvari, 2013; Kadir, Goodman, & Harris, 2013; Mańko, Geracitano, & Capogna, 2011), bioinformatics [41], [42] (Frijters et al., 2010; Khalid, Yunus, & Adnan, 2010), patent mapping [43] (Woon & Madnick, 2011), face recognition [13] (Zhao & Chellappa, 2003), prosodic modelling [44] (Escudero-Mancebo & Cardeñoso-Payo, 2007) and many others. These techniques can be supervised or unsupervised depending on the presence or absence of initial knowledge injected into the clustering algorithm [35] (Polanco, François, & Lamirel, 2001). Unsupervised methods present more interesting potentials for problem solving or solution finding by grouping data in statistically distinct groups without *a priori* knowledge. There are as many cluster models as algorithms and so far none can be said superior because there are always many possible and correct ways to group data together [37] (Jain, 2010) (Figure 5). Distance and similarity measurements are the typical values used to compare different clusters of data [43], [45], [46] (Choi, Yoon, Kim, Lee, & Kim, 2011; Rodriguez-Esteban, 2009; Woon & Madnick, 2011). In the IP domain these tools have been used in business analytics to visualize citation relationships or patent collections [30] (Lupu, 2013). The expansion of semantic tools in IP search is asking for more visualization and visual analysis techniques to be used to simplify professionals’ task.

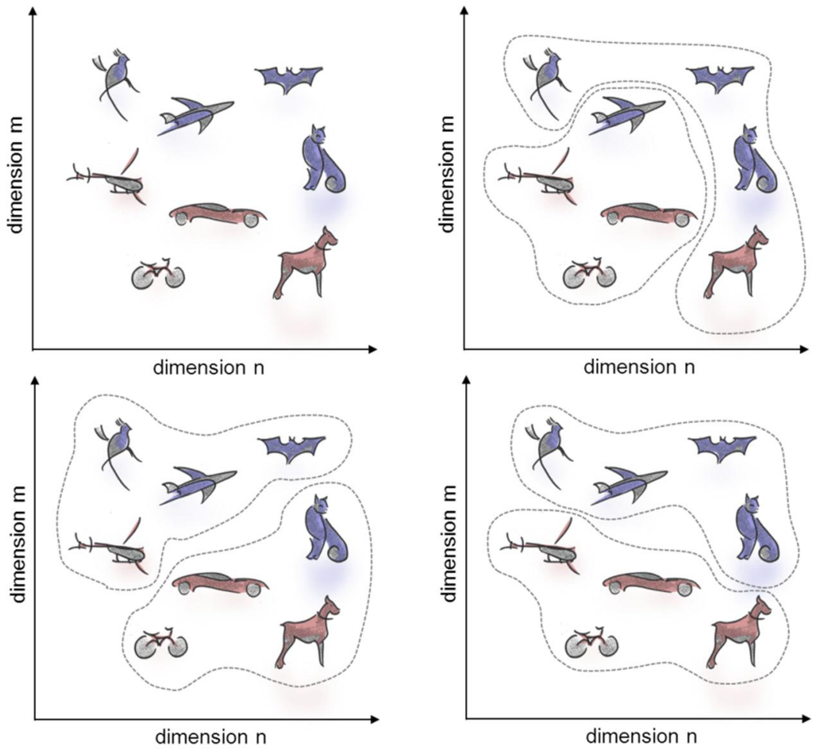

Fig. 5. Multiple Ways of Clustering, with electable parameters such as m,n…

2.4 SEARCH TOOLS EVALUATION ISSUES

The aim of this section is not to detail all possible evaluation methods (a more exhaustive reading can be find in [8] (Lupu, 2011)) but some aspects necessitate our attention. The first one is that IP search engines despite many publications and campaigns on the topic do not come with satisfactory evaluation methods. There are currently two main ways to evaluate the results of a search.

The first one uses "quality measurements" or "effectiveness measurements" which all derived from the initial Cranfield collection tests based on two factors: precision[1] and recall[2] [47] (Bonino, Ciaramella, & Corno, 2010) (Figure 6). Individual attempts have been made to evaluate the effectiveness of keyword search strategies [48] (Xie & Miyazaki, 2013). Despite some methodological discrepancies to obtain these measures, the traditional process is to use a set of predefined queries presented to a search engine using a manufactured corpus of documents. The number of relevant document for each query being known. This generates values for each factor which can then be both summarized in one single value, the F-measure. Many values then derived from this such as the R-precision, the break-even point, as well as the discounted cumulative gain and rank-biased precision family measures, if ranking efficiency is considered. In order to boost the whole field of information retrieval system evaluations campaigns such as the NTCIR, TREC and CLEF have been organised to bring together the experts of the field in order to generate accessible corpus and queries to evaluate any search technology [49] (Piroi et al., 2012) (NTCIR: NII Test Collection for IR systems - NII: National Institute of Informatics (Japan) - TREC: Text Retrieval Conference - CLEF : Conference and Labs of the Evaluation Forum). Though these results might be interesting for keyword-based searches they do not shine any light on more advanced systems using semantic [50] (Hempelmann & Raskin, 2008). The F-measure is also questionable when comparing different applications since a low value might simply reveal a harder task. Moreover, different engines may accomplish different results on different types of text meaning that a high F-measure value might simply be due to a specific entry data format [46] (Rodriguez-Esteban, 2009). The proposed corpuses are also

[1] Precision corresponds to the number of relevant documents present among those retrieved.
[2] Recall is the number of relevant documents retrieved among all relevant documents present in the initial collection.

completely artificial and it is difficult to relate these evaluations to any reality. The assumption that if a system performs well on tests evaluation it will perform well [8] (Lupu, 2011) appears oversimplified considering the complexity of the problem at hand.

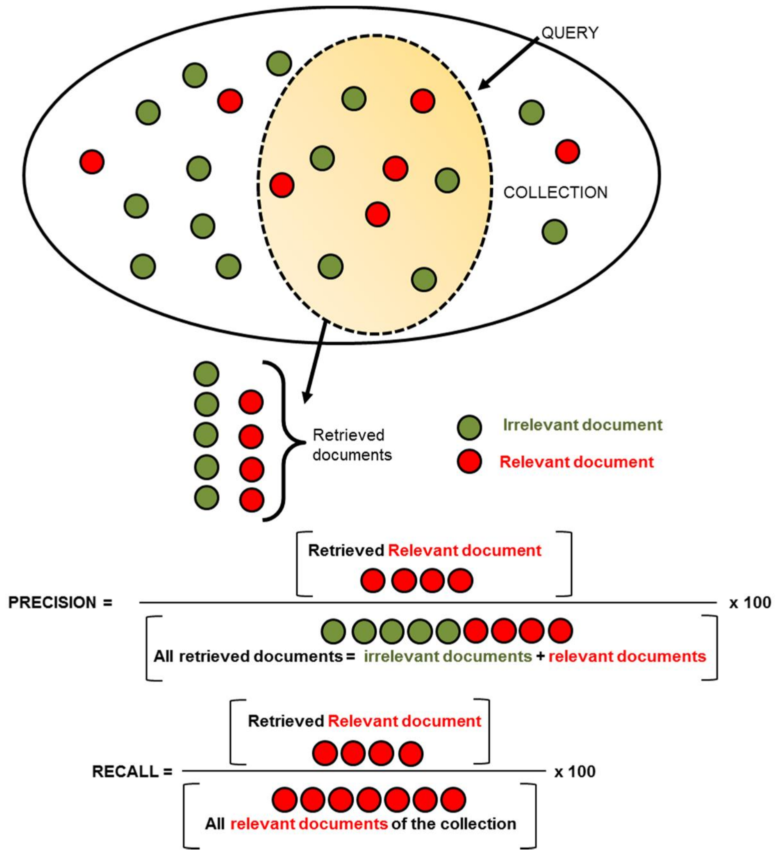

Fig. 6. Precision and Recall Effectiveness Measurement of a Search

The second possibility when evaluating a search system is the expert panel [10], [17], [51] (Magerman et al., 2009; Yoon, Park, & Kim, 2012; Hawizy, Jessop, Adams, & Murray-Rust, 2011). The retrieved results following a query are controlled by experts who validate the relevance of each document. This has several drawbacks such as the relative limited amount of documents that can be assessed as well as the diverging opinions on what is considered to be relevant.

As we mentioned at the beginning of this chapter, there are no satisfactory method and it is often the empirical experience of search professionals that will guide them to use the tool that can provide the more appropriate answers.

3 KEYWORD SEARCH

3.1 THE DOMINANT BOOLEAN SEARCH TECHNOLOGY IN THE IP DOMAIN

The reader can find a detailed description of the Boolean search in ‘An introduction to information retrieval‘ [52] (Manning, Prabhakar, & Schuetze, 2009).

Keyword-based search and in particular Boolean search have dominated for decades free and commercial information retrieval tools including IP search. This information retrieval model is based on queries combining the Boolean operators AND, OR, NOR and NOT (Figure 4 & Figure 2). The model views each document as a set of tokens which for most of the cases correspond to words. In order to query a corpus, documents need first to be indexed. The Boolean inverted index associates each term of the collection with the documents in which they appear. A query such as *Romeo* AND *Juliette* would be process as follow: locate in the inverted index dictionary the term ‘romeo’ (most of keyword search systems are case insensitive); identify what documents the term ‘romeo’ points to and create a posting list; do the same steps for the term ‘juliette’; merge both posting lists; extract the documents that includes both terms. In this kind of index a word does not have any associated meaning even when synonyms are used to improve the performance. Other options have been added to the original Boolean search such as truncation to find words sharing common structures (e.g. word-words-wording); wildcards are expressions that can replace characters or digits; proximity looks for documents where two or more terms occur within a certain distance; fuzzy search helps with misspelling using approximate string matching methods; and synonyms.

Boolean search has been shown to be the fastest technology to search through multiple documents. On the contrary to other technologies it gives clear and fast results, e.g. a document matches or not a query. Nowadays the majority of keyword search engines use ranked technologies to balance the absence of weight for each document in a Boolean search. The ranking of relevant documents is done using tools such as the vector space model [10] (Magerman et al., 2009), probabilistic models and term weighting (tf-idf: term frequency–inverse document frequency) [8] (Lupu, 2011).

3.2 A LEGITIMATE DOMINATION?

Free text query solutions, which have been shown to perform with higher effectiveness [53] (Turtle, 1994), exist since the beginning of the 90s. However as seen above, Boolean search is still the main technology in use nowadays [8] (Lupu, 2011). Despite some obvious advantages for software developers Boolean search is overall not satisfactory. Keyword searches often generate irrelevant results because one term can have different meaning in different contexts and relationship between words is lost in inverted index dictionaries [18] (Trappey, Trappey, & Wu, 2009). Another intrinsic issue of Boolean search is that using AND operators increases precision and lower recall, while using OR operators decreases precision and increases recall, and it is nearly impossible to find the right balance [52] (Manning et al., 2009).

Keyword search engines are for most of them not using any dictionary making the system intolerant to spelling mistakes and will not help the user with inappropriate choice of words as would do Google Search for example. Many patenting organization search engines offer neither case sensitivity technologies nor term proximity. A Boolean system when used alone only reports term presence or absence and display results with nothing more than a chronological ranking [52] (Manning et al., 2009). Keyword searches are adequate for getting information about entities, which are typically nouns or noun phrases. They fail on queries that are looking for relationships between entities. For example, if one wants to retrieve documents containing text of the form “Company X acquired Company Y”, then keywords alone are extremely inadequate [54] (Sarawagi, 2007). Not only these technologies alone are inefficient in finding the right documents, they also rely almost entirely on users’ expertise and databases’ size.

3.3 USER-FRIENDLY OR AN ILLUSION OF SIMPLICITY

At first sight, a search using Boolean operators might looks straightforward but anyone who has used this type of queries knows that the reality is different. First of all it nearly entirely relies on the users’ skills acquired during years of trials and errors [47] (Bonino et al., 2010). Furthermore “regardless of how well the indexer manages to process the patent collection, the ultimate results will also depend on asking the right questions, or, in this case, on generating the most effective queries” [30] (Lupu, 2013) and this is not an easy task.

Here is a query from Yoon and Kim [17] (Yoon et al., 2012) in order to collect granted patents from the USPTO since 2000 about organic photovoltaic cells:

((((((photovoltaic* solar*) adj (cell* batter* device*)) and ((bulk* adjheterojunc*) ppv* phenylenevinylen* tandem* (dye* adj sensitiz*)fluoren* fulleren* PTCBI* PTCDA* PTCDI* H2PC* ZnPc* CuPc*TPyP* TFD* NPD* CBP* PCBM* (conjugat* adj polymer*))) or(((organic* plastic* polymer* (dye* adj sensitiz*)) adj3 (photovoltaic*solar*) adj (cell* batter* device*)) DSSC*))) and (B32B* C07* H01*H05B-033*).IPC.) AND @RD[=20000101\=20101231

This query illustrates two things. First the person who wrote it is an expert in this particular field and second he is accustomed to the USPTO query syntax. Syntaxes between search engines can be very different. For example if one is looking for the word “graphene” in patents’ title, the correct syntax for EspaceNet would be “ti=graphene” whereas for USPTO that would look like “TTL/graphene”. No need to give more examples if such simple query is already that different. It requires the user to be properly trained for one particular engine, and one can only learn few of them if efficiency is the goal. This ultimately impacts knowledge availability with only big corporations able to have the necessary human resources.

Another interesting concept for our present discussion is the principle of ‘query optimization’ [52] (Manning et al., 2009). This is the process of selecting a query that will ease the search process for the system. One way of doing it is to call the different posting lists in a specific order. This is the perfect illustration that to be efficient the system needs the user to formulate expert queries to find the relevant set of documents. Nevertheless it raises an interesting point that the system can be improved since the user can optimize it.

It is then quite understandable that this technology is still the most used one in the IP search domain. It is more convenient and easier to develop and manage to the detriment of the users community.

The keyword based query system demands from the users to formulate their objectives in a very complex and unnatural way. In a society where everything is made to ease the life of users this technology seems pretty obsolete.

In an attempt to summarize for the non-trained in the art, - more details being aforementioned -, the main disadvantages of the overwhelmingly used keywords approach, can be summarized as follow:

- some practical knowledge of the field is necessary including jargon and synonyms;
- some knowledge of how the search engine operates is necessary to yield reliable results. The more flexible and performing-powerful the search engine used, the more essential is the knowledge and experience acquired; and
- acquaintance of the relevant databases, including abstracting and organized-controlled language, is essential to produce a dependable result.

4 SEMANTIC SEARCH

4.1 NATURAL LANGUAGE PROCESSING

Semantic technology is often presented in the literature as a solution developed to overcome the limitations of keyword centred search and to address complex information needs [55], [56] (Dreyfus, 1972; Minsky, 1969). Not only this information is wrong but it gives the false impression that there slow progress is due to the youth of the project. Semantic search is not a novel concept and has been a main branch of AI research since its beginning. The reader might want to read about the early attempts in the semantic treatment of information [55] (Minsky, 1969). The foundation of semantic search is the simulation of human language processing. Linguistic being one of its main field of study. Natural language processing (NLP) is the original component of semantic but nowadays many other theories have emerged. The goal of these tools is to automatically decompose unstructured text data into smaller parts, identifying their elements and relations and putting them into a machine-readable structure for analysis. As opposed to keyword based tools the meaning of each word as well as their relationships is at the centre of the algorithms. The strength of semantic is the ability of the system to perform a search by interpreting the meaning of keywords and extracting object, concepts, symbols and their relationships. NLP uses different tools to break/parse the textual information into meaningful elements such as sentence extraction, tokenisation[3], word stemming[4], part-of-speech tagging[5], lemmatization[6] and named entity recognition[7] [16], [51], [57] (Yoon & Kim, 2011a; Hawizy et al., 2011; Spinakis & Chatzimakri, 2005). In order to "understand" the meaning of a word the system needs specific knowledge regrouped into ontologies. An ontology is "a taxonomy with multiple, precisely defined links between the items that represents knowledge as a set of concepts and their relationships" [8] (Lupu, 2011). The human language has many properties such as synonymy and polysemy that are a challenge for search engines. When using keyword based solutions one would have to enter all possible synonyms and all different meaning of one word in order to grasp the whole information. Semantic is doing that naturally using the context in which a word appears, as a human would do when reading a document. Ontologies are domain-specific because one keyword in a domain might have a different meaning in another one. We think it is important here to highlight the fact that most of the published documents are written in English with a lot of authors using English as a foreign language, which might add confusion around the meaning of a keyword. A French person is for example

[3] Tokenisation is the process of cutting character strings into pieces called *tokens*. These elements can be words, symbols or other meaningful elements [52] (Manning et al., 2009).

[4] Stemming is used to gather certain words under the same *stem* and reduce the number of indexed elements. The stem 'search' is the root word for 'searching', 'searcher' [10] (Magerman et al., 2009).

[5] Part-of-speech taggers classify words as nouns, verbs and others [52] (Manning et al., 2009).

[6] Lemmatization groups all grammatical inflections of a word under a common *lemma*. The lemma 'have' includes 'has' and 'had' [52] (Manning et al., 2009).

[7] Named-entity recognition techniques are used to find and classify domain specific *terms*. A term can be a medical disease, chemical name, person, and others [29], [46] (Rodriguez-Esteban, 2009; Rzhetsky et al., 2009).

profusely using “eventually” instead of “possibly” because “possibly” corresponds to “éventuellement” in French. The context is therefore crucial to understand textual information.

Despite its relatively long history, semantic has only been recently introduced in search engines. Other conceptual models are available instead of NLP such as object or entity relationship modelling [58] (Wen, Zeng, Li, & Lin, 2011). The notion of function, defined as the action changing the features of any object, is another part of semantic research from TRIZ. The functions are also associated with Subject-Action-Object (SAO) models [15], [16], [17] (Yoon & Kim, 2011a, 2011b; Yoon et al., 2012).

According to Eetu Mäkelä [59] (Mäkelä, 2005) semantic search has five main research directions: “augmenting traditional keyword search with semantic techniques, basic concept location, complex constraint queries, problem solving and connecting path discovery”. The efficacy of semantic search has been demonstrated for keyword search increase performance [59] (Mäkelä, 2005), new technological opportunities detection [15] (Yoon & Kim, 2011b), technology landscape visualization [43] (Woon & Madnick, 2011), similarity measurements [60], [61] (Moehrle & Gerken, 2012; Moehrle, 2010), technological trends identification [16], [17], [45] (Yoon & Kim, 2011a; Yoon et al., 2012; Choi et al., 2011), distance analysis [15] (Yoon & Kim, 2011b) and novelty monitoring [62] (Gerken & Moehrle, 2012) among others.

4.2 Ontology: general challenges and IP specificities

A machine needs specialized knowledge to “understand” textual information. Ontologies are the knowledge base of semantic technologies and are domain specific as seen previously. It is therefore critical to create ontologies for all existing domains including IP [63] (Hare, 1998). Biomedical and chemistry domains are both ahead of many others in term of semantic search and therefore available ontologies [20], [46], [51], [64] (Jessop, Adams, & Murray-Rust, 2011; Rodriguez-Esteban, 2009; Hawizy et al., 2011; Gurulingappa, Müller, & Hofmann-Apitius, 2011). The fact that some ontologies are provided by organisations such as the Royal Society of Chemistry demonstrates that a community effort is necessary to develop such dictionaries. Only domain experts are qualified to generate such databases and this ground work is one major limitation for semantic search to be used more broadly. The IP domain and in particular patent documents propose an even greater challenge. “It is well settled that a patentee can act as his own lexicographer by choosing his own definition of a term and using it as he wishes so long as he remains consistent in its use and makes the meaning clear” [65] (Wetherell & Mehok, 2005). Not only this makes the whole ontological process a nightmare but it does generate irrational activities in the course of a patent’s life. It is important to remind the reader that during infringement analysis the claims of the patents under examination are compare to one another using similarity measurements. This aberration is analogous to the measure of similarity between a French written document and a German one using an English dictionary.

A collaborative ground work from domain experts is a necessity if we want ontologies to ever become the foundation of search engines. It is important to notice here that they might not be the appropriate response to the current issues in data mining.

4.3 Latent Semantic Analysis

An alternative to NLP is Latent Semantic Analysis (LSA) also called Latent Semantic Indexing. “This is a method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text” [66] (Landauer, Foltz, & Laham, 1998). This method does not rely on ontological knowledge to find and assimilate new concepts. Using artificial neural networks it creates its own conceptual lexicon using algebraic methods based on word frequency. The corpus of documents is the only knowledge necessary for the system to work. This unsupervised learning method produces concepts that “do not have any textual or intuitive expression of their own, but they are defined by vectors that relate them to words of the initial vocabulary”. Concepts are mathematical abstractions obtained by different sets of comparisons [67] (Fernández Sánchez, 2009).

LSA is used to do semantic with a non-semantic tool, avoiding many inherent problems but reducing the potential power of its application [50] (Hempelmann & Raskin, 2008). The aim is to deduce meaning from non-meaningful language properties using statistical tools such as co-occurrence of words or other structural aspects [43] (Woon & Madnick, 2011). LSA is employed in language processing to help with text categorization, indexing, essay grading, image auto-annotation, auto-annotation, and automatic cross-language retrieval. Google recently implemented semantic in its search engine with the hummingbird update. Though it is difficult to find the information on what is exactly behind this improvement the patent title suggests the use of LSA [68] (US 8,538,984). Such technology has been shown to outclass Boolean models and seems to give an interesting alternative to the complexity of NLP [69] (Chen, Martin, Daimon, & Maudsley, 2013). However it has not been yet proven that LSA concepts correspond to any reality [67] (Fernández Sánchez, 2009).

4.4 INADEQUATE INITIAL DOCUMENTS SELECTION

It is of common use to run semantic analysis on pre-selected set of documents. The selection process is often done using keywords centred search systems. We cannot stress enough that one purpose of semantic solutions is to get rid of the inadequate use of keywords to build a better understanding of specific fields of research. By pre-selecting documents using inappropriate tools researchers greatly reduce the usefulness of semantic to a simple analytical instrument. In the IP domain, considering everything that has been discussed so far in this review, keyword search engine should be used only when searching for a specific patent or publication. If the goal is bigger it is then a necessity to use the full potential of linguistic to generate innovative paths, map existing domains, compare technologies and so on. Human communication, and more importantly, problem formulation is made using natural language and not a list of keywords. It is therefore natural to question a system using queries written in natural language. The technology to process such queries exists and is already used but the IP industry is slow to make the change. The relatively low number of existing ontologies or thesaurus is obviously a limitation as described in the previous chapter.

There are processes in place [1] (Rebouillat, 2013) that request ideation to start with a 300 word expression, defining the problem to solve, the proposed solution and a path to practice. The 300 count is not a magic number but corresponds to a size that avoids distraction and temptation to express several rather intricate ideas with some googling intrusions; possibly removing the idea from its original inception. Such ideation expression is then used as natural language inputs in ad hoc engines.

4.5 PROBING THE VARIOUS APPROACHES

In an illustrative attempt, of very limited scope, amplitude and depth, to compare some of the various search approaches (Figure 7) and to underline their complementary and/or usefulness, three cases were studied and are reported here.

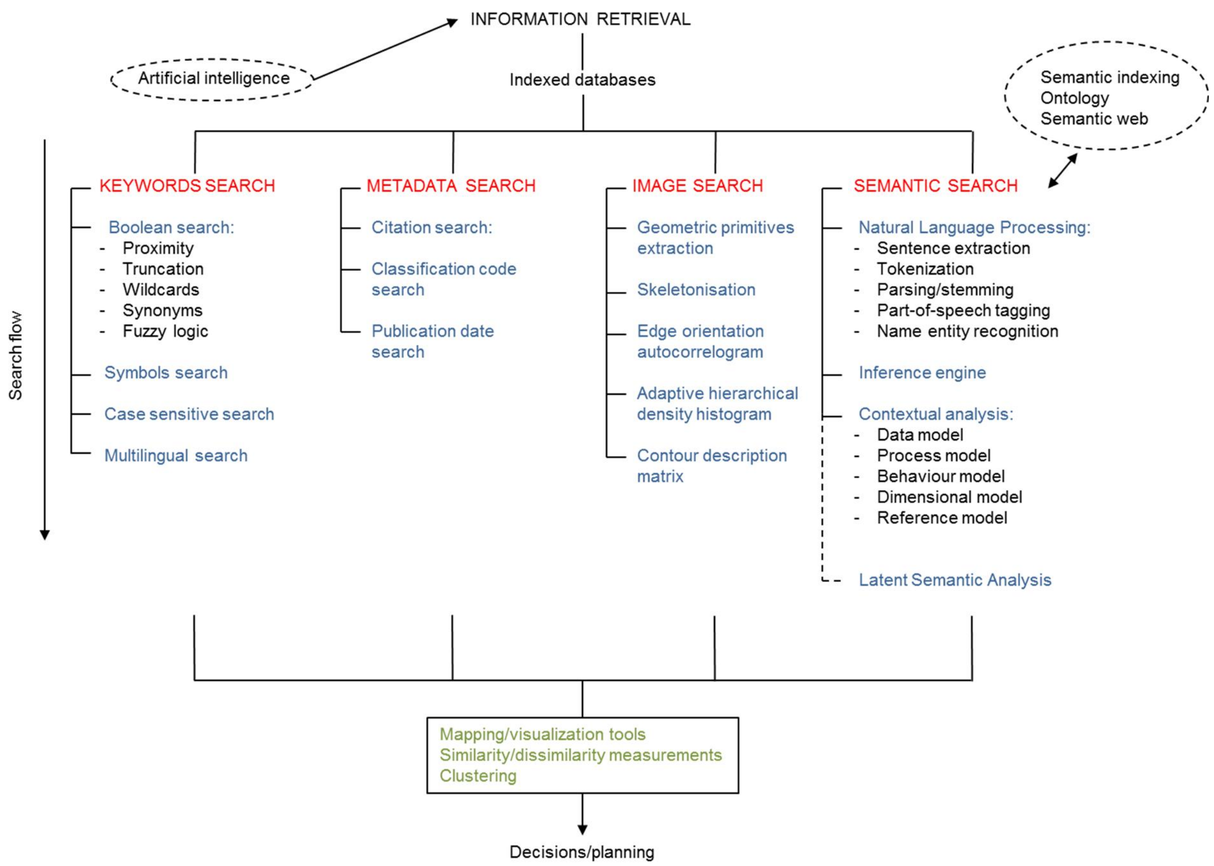

Fig. 7. Various Search Approaches

During the course of a patent examination, a preliminary patent search report on patentability will be coordinated by the ad hoc patent office, at the national level or at the international level. For example in the case of a broader (PCT) application via the WIPO, or in the case of a European application via the European Patent Office. The PCT is an international Patent Cooperation Treaty, directed by the World Intellectual Property Organization (WIPO), involving more than 140 Paris Convention countries; therewith enabling broader and simpler patent application procedure with a well-defined priority date applicable across the member states.

The patent search report on patentability provides, earlier on, to the assignee, often referred to as the applicant, the list of relevant documents that may affect the patentability of the considered patent applications.

Two categories of cited documents in the search report have special and critical relevance. The X category covers the document of particular relevance against which the claimed invention can neither be considered novel nor inventive. The Y category classifies documents which are of particular relevance against the claimed invention which cannot be considered to involve an inventive step when the document is combined with one or more such classified documents, such combination being obvious to the person skilled in the art.

The search report is performed by a series of structured patent preliminary examination steps involving patent examiners who rate the application against the above classification criteria. The related patent examiner hierarchy is very specialized and experienced in the domain and is believed to operate via IPC/keyword searches mostly. The expert in the field examiner has means to cross check her/his findings and corroborate with various adjacent cases.

First case study: "Continuous flow biodiesel processor, US20070175092"

The search report on patentability, available via the Patenscope® dedicated patent database, cites two X references dated 1928 and 1973, and 4 Y documents dated between 1987 and 2000.

The search was performed on the 4th of February 2008. We chose to conduct a semantic search using the entire patent document and limiting documents output to those appearing prior to that date.

The semantic search yielded no X references but only two Y references, one classified in 3rd position, of all relevant documents found, with a relevance of 88 over 100 (100 being the maximum relevance achievable, generally meaning the exact same document) and one ranked 19 with a relevance of 80 over 100. Worth observing that the X documents are all available in the selected database; included the one dating back to 1928. The semantic analysis did not found those X documents as being relevant.

We clearly outline here the importance of the field expertise favourable to the human expert examination and co-involvement in any automated computerize process. The latter can also misjudge. The bioscience domain of concern being in its establishment phase may also suffer from the lack of coverage and classification in the IPC current version.

Second case study: "US20130317853 - WO/2013/177493A1 - Device selectively storing and presenting critical medical information."

For that particular case the "recordation" of the PCT search history is provided with a completion date of 20/8/2013.

As listed below 6 patent classes, 2 patent databases, 4 NPL databases, 21 keywords have been used for that search; this is hardly conceivable by non-patent search experts.

"Field of Search/Classification Information:

IPC(8) Classification(s): G06Q 50/00, G08B 23/00 (2013.01) USPC Classification(s): 705/3; 128/903; 600/300,340/573.1

Database(s) Searched (Patent and Non-Patent Literature (NPL), Including SubDatabases and Files Searched) and Search Terms Used: MicroPatent (US-G, US-A, EP-A, EP-B, WO, JP-bib, DE-C.B, DE-A, DE-T, DE-U, GB-A, FR-A); DialogPRO; Google; Google Scholar; Medline/PubMed: worn, wearable, wristwatch, pendant, broach, bolo, wristband, bracelet, jewelry, medical, drugs, allergies, diseases, illnesses, microprocessor, cpu, microchip, rom, ram, portable, hand-held."

Performed on the 23rd of August 2013 the search report available in the Patenscope® database cites 2 X classified documents and 4 Y classified ones.

A semantic analysis using the entire patent document was conducted limiting the output to a date prior to the completion of the international search. In rather unfavourable positions 69 and 73, were identified the two X classified documents with a relevance of 75 and 74 over 100, being the maximum relevance obtainable. No Y cited documents were found within a reasonable relevance level, which is generally set at a maximum 70-75 by experts in the field. Below that level of 70-75, documents are generally considered as non-relevant but only of documentary value representing the general state of the art.

The above demonstrates the complexity of the search in the patent field and the rather distant results obtained from a semantic search versus the human expert approach interacting with multiple databases.

Nonetheless, in both cases some significant references were found, which would have been unlikely, probably impossible, based on a single keyword approach or sole patent class search methodology.

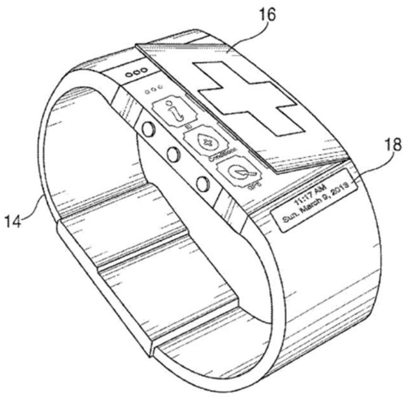

Fig. 8. Device Selectively Storing and Presenting Critical Medical Information

Figure 8 above provides a good representation of the article subject of the second case. One would wonder whether an image based search, possibly combined with some interactive semantic, would have not provided a faster and more precise recall; at least the searcher would have an already quite inspiring idea of the competitive adjacent article to look for and/or to compare with.

<u>Third case study:</u> Neo-Retro-Innovation, way back to future innovation?

We propose to further test the information retrieval means at disposal to the trained in the art of searching, non-expert in the field.

Using a recent definition of a PC-Tablet, provided by Henderson [70] (Henderson, 2009), this third case study aims at finding whether with the information available back within a selected period of time, such as 1975-1995, references corresponding to that definition were available at that time. The definition of interest is formulated in the following fashion:

“As the name suggests, a tablet PC is a small computer about the size of a notebook (not to be confused with a “notebook PC,” which is a small, light laptop). The user can write on the screen with a stylus to take notes (for similar functionality, see graphics tablet), draw, and make selections with stylus or fingertip. If the user writes on the screen, software converts the writing to the appropriate characters and stores them in a file (see handwriting recognition). As with some PDAs, there may also be a system of shorthand “gestures” that can be used to write more quickly. Alternatively, the user can type with stylus or fingertips on a “virtual keyboard” displayed on the screen (see touchscreen). A more versatile and natural interface is becoming available: “multitouch,” pioneered by the Apple iPhone and Microsoft Surface, can recognize multiple motions and pressure points simultaneously. This allows the user to, for example, flick the finger to “turn a page” or use a pinching motion to “pick up” an object. Applications for tablet PCs include many PDA-type applications (see personal information management and PDA), field note taking, inventory, and other tasks that require a device that is not encumbering. Because of its compactness, a tablet PC can also be a good reader for ebooks (see e-books and digital libraries). Tablet PCs generally follow common specifications developed by Microsoft, and often use Windows XP Tablet PC Edition or, later, Windows Vista, which has built-in support for tablet PCs. These operating systems include support for sophisticated handwriting recognition that can be “trained” by the user and that can store handwritten input in special data formats. Voice recognition is also supported. A “convertible” tablet PC is a hybrid in which the tablet is attached to a base containing a keyboard. The display can be used vertically (laptop style) or rotated and folded down over the keyboard for tablet use.”

Using commercial tools with a semantic analysis feature, a search using the above PC tablet definition yields a list of related references, appearing within the selected period of 1975 to 1995, and classified from most to least relevance level. We leave other details for a paper more focused on those search criteria.

18 references were found ranking from an acceptable relevance level of 76 to a moderate relevance level of 67.

In position 2 of the 18 reference list appears the Agulnick, Todd patent dated 13 September 1994, (Application Date: 31.10.1990), untitled "Control of a computer through a position-sensed stylus" published under United States Patent 5,347,295.

A glance at one of the patent drawing, Figure 9, would draw the attention of the observer towards a tablet of nowadays; visually an iPad® like almost.

Additionally a Google keyword search using the word sequence "wiki tablet history" provide the link, "http://en.wikipedia.org/wiki/History_of_tablet_computers", which relates to the development history of the tablet computers. The proposed Wikipedia article cites, in position 82 out of 103 references, the exact patent by Agulnick and Todd mentioned above.

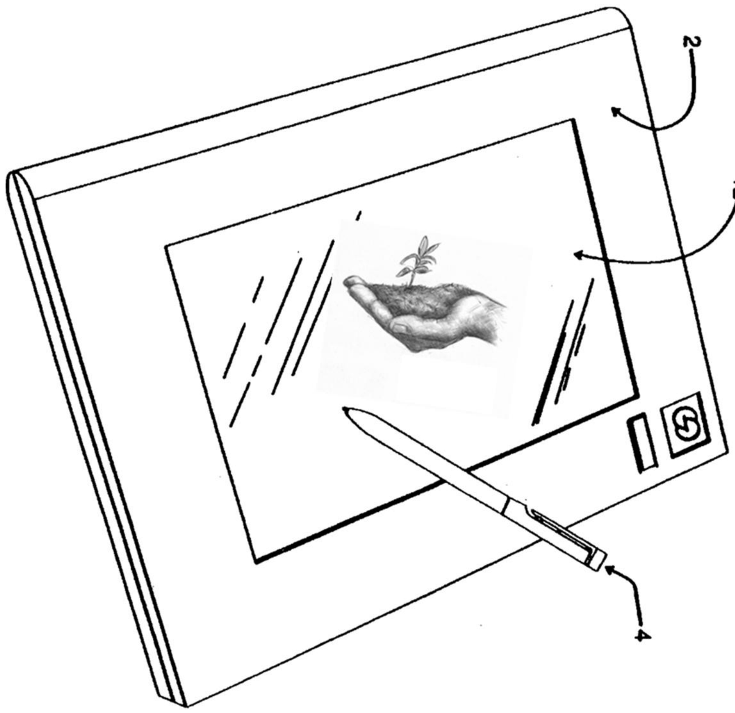

Fig. 9. Control of a Computer through a Position-sensed Stylus with Application Date: 31.10.1990

Wikipedia list of references comprises only 4 patents, all retrievable in free databases as long as the number is known. Only one of those references was found from the semantic analysis, although the most pertinent one by far; especially given its inspiring picture that may have led to earlier discoveries in the field, should the image be findable beyond a pure IP context and the examination of the specific entire document.

The "Telautograph" was subject to patents dated back in 1888 and 1942, which are referred to in the Wikipedia reference list and clearly state the object of the two inventions. For the latter, 1942, "My invention relates to improved means and systems whereby writing or tracing movements of a body such as a pen, pencil or stylus in tracing a picture or design may be reproduced at a distance.". And for the first one, 1888, "The invention relates… to the act of writing at a sending-station operates to reproduce it at the receiving-station".

The dawning (Figure 10) associated with that latter invention, 1888, United States Patent 386,815, is eye-popping given the clear interpretation by drawing of the fact that the movement of a pen is reproduced at distance via electro-magnetic components; this back in 1888.

- The above study shows that semantic is a quick and a valuable improvement to patent searching, requiring minimum training. There are improvements which would be welcome as indicated in the rest of the paper.
- The current case study also underlines that retro-innovation is conceivable, using contemporary information. The possibly most inspiring information are discoverable back 20 years before a breakthrough technology is appearing on the market. The 1994 patent could have been or has been inspirational to current tablets, ipad® and the likes. The 1888 Telautograph drawings are premonitory, prophetic or clairvoyant provocateurs to many remote actionable devices of today.
- Finally, image analysis combined with semantic support is definitely a promising route to enhance innovation pace. The proliferation of camera life video recording such as dashcams could promote the development of *in silico* IP image interpretation and semantic ideations and/or reverse.

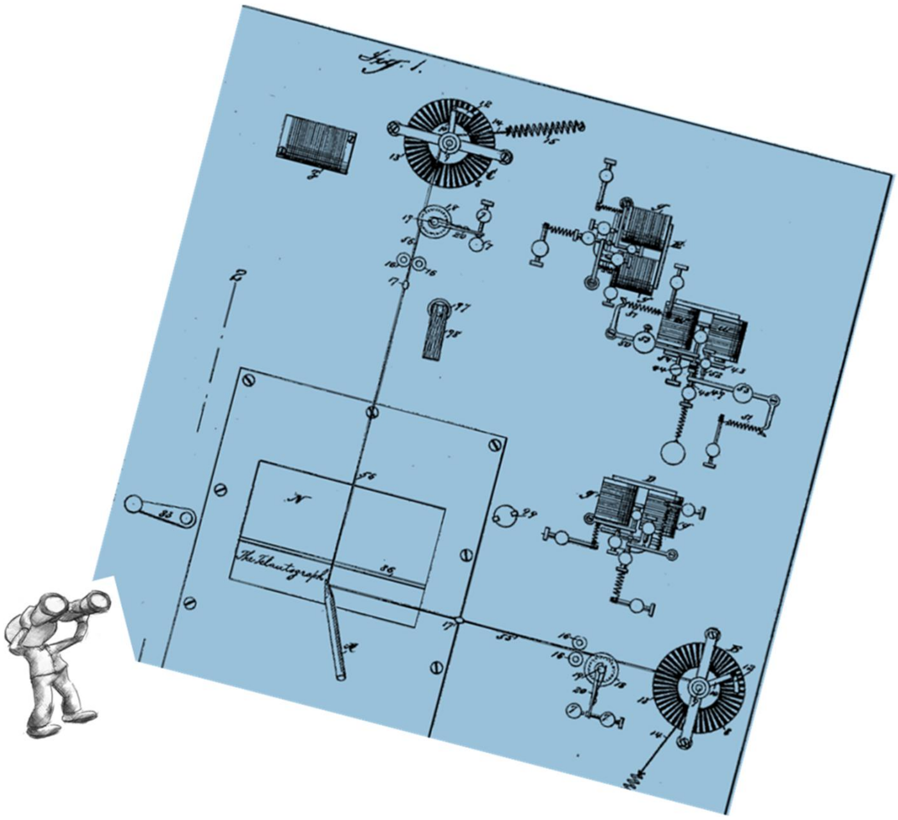

Fig. 10. The Telautograph (1888)

5 IMAGE RETRIEVAL SYSTEMS

5.1 A MISSING LINK

The adage "a picture is worth a thousand words" suits particularly well to the IP domain. In a patent; also true for scientific publications; drawings, diagrams, plots convey an often critical amount of information that is largely ignored by IP search engines [71] (Bhatti & Hanbury, 2012). Considering the terminology problems inherent to patent, image retrieval technologies seem a natural and crucial step in IP search. It is thus surprising that it did not yet reach a satisfactory level. However some tools exist and have been used on patents for image retrieval. The Optical Structure Recognition Software (OSRA) was applied to recover chemical information [20] (Jessop et al., 2011) and the Content-Based Image Retrieval (CBIR) for drawing retrieval [30] (Lupu, 2013). In addition some rare attempts have been made to associate several techniques including image search in the IP domain [26] (Codina et al., 2008). This is obviously not a trivial task but some other domains seem to have done better in this matter.

Semantic Image Analysis, SIA

There is no need to further insist on this aspect and the necessity to add it to the patent/IP search toolbox; the sooner the better at the condition to take care of its harmonization with legal aspects. The lawyer's "shyness", vis-à-vis the interpretation of drawings and the claimable matter related to those, is well-recognized.

It could be the subject matter of an entire chapter. Let's first make sure that the principle is clearer than the title, semantic image analysis (SIA), may pretend, and let's use for that purpose a very simple analogy associated with children recognition and learning process.

The SIA, using a symbolic neural architecture, is schematically associated with Figure 11; one of the key references on the matter has been proposed by Kollia [72] (Kollia, 2010).

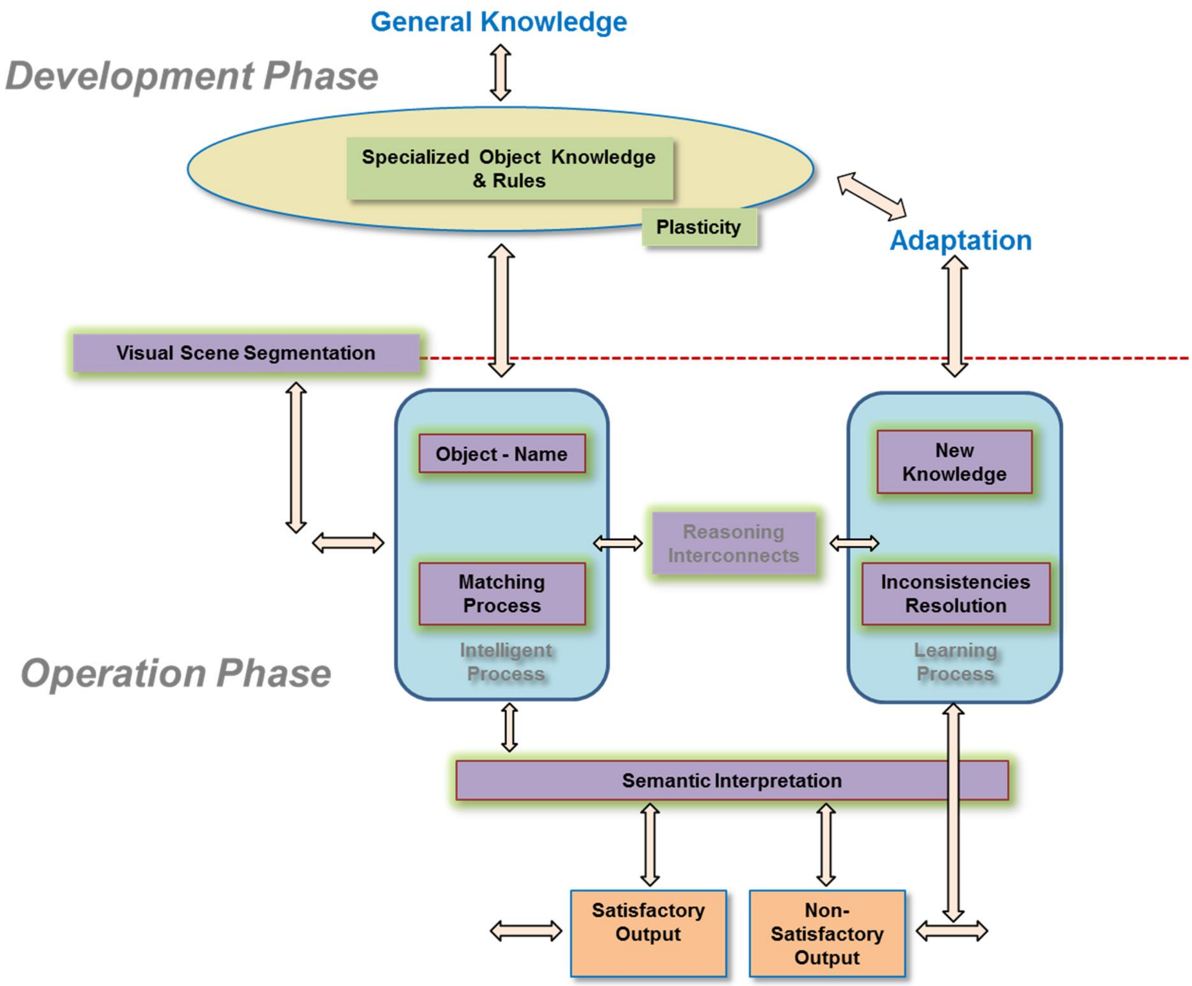

Fig. 11. The Semantic Image Analysis, using a Symbolic Neural Architecture

In order to understand the whole process let's describe what could be the way a child is visually interpreting its environment. The first step would be to segment what he sees into objects depending on colours, intensity, depth, size, etc. (= feature extraction and classification). The next step would be to associate a name to each detected object (= semantic interpretation) using the knowledge he gained from his parents (= formal knowledge). However not every shape an object can take has been previously encountered and explained. The formal knowledge therefore contains some conflicting information (e.g. a lamp does not always possess a lampshade). His brain (= reasoning engine) will use the new information (connectionist learning) to adapt and modify (= knowledge extraction) his general knowledge. Using this strategy and new inputs from his parents, the child will slowly become expert in the interpretation of his surroundings. By similar means the semantic image analyser segments and classifies pictures using a formal knowledge. This latter contains inconsistencies that will be corrected by the knowledge adaptation through the reasoning engine. Follows a correct or improved interpretation of new pictures presented to the analyser.

It shouldn't be of any surprise if the child chooses to put a hat on a lamp that is missing a shade by design, or by some sort of necessity. The child will, by this token, appreciate the fuller purpose of a shade when the recreated object fails to meet its primitive observable function. A lot more can be learned from Chomsky's paramount contribution [73] (Chomsky, 2013) in the language-cognition field.

5.2 WHAT CAN WE LEARN FROM OTHER DOMAINS?

Image retrieval is not restricted to the IP domain and has made huge progresses in other fields of research [71] (Bhatti & Hanbury, 2012). One field in particular, which shares several elements with information retrieval as defined here, has been really successful during the past few years. Face recognition technology necessitates a high level of image analysis and encounter many challenges. Similar to IP image search, face recognition using AI needs to analyse an image and compare it with a database. However the analysis has to be often done online while screening live and poor video signals. The success of these techniques might be attributed to the large number of applications for this technology. They are present in the field of advanced video surveillance and passport controls, but also TV parental control, video games, human-computer-interactions and many others [13] (Zhao & Chellappa, 2003). Face recognition sequence follows a similar path than image search, namely face detection, feature extraction and face recognition. Though they do not confront exactly the same issues, there are no reasons for image search technologies to be behind if a process as complex as facial identification is possible. Other technologies are used for hand-drawn sketches recognition [74], [75], [76] (Brieler & Minas, 2010; Fernández-Pacheco, Albert, Aleixos, & Conesa, 2012; Sezgin & Davis, 2008), to identify children who are at risk of handwriting difficulties [41] (Khalid et al., 2010), to create a 3D model from a sketch [77] (Olsen et al., 2009) among others. These various successes illustrate well the feasibility of image search in the IP domain.

6 WHAT IS BEHIND THE AI BLACK BOX: I WISH I KNEW!!

Before going deeper in the subject of this paragraph we think it is important to clarify one concept, e.g. artificial intelligence is *artificial*. It might sound obvious to the reader but it is in fact an important aspect of a search engine or AI in general. Despite a lot of publications that claim otherwise for many years now [55], [78] (Davies, 2011; Minsky, 1969), the abilities of AI is very far from performing the full range of human brain cognitive skills [79] (Schlinger, 2012). The purpose of this declaration is not to minimize the progresses done since the birth of AI but to highlight the gap that still needs to be crossed to have a human brain *in silico*, if required. Neuroscience is still a juvenile discipline and the mysteries of the human brain have been barely explored and even less understood. This is why this intelligence is artificial, e.g. based on brain activity theories that might only partially reflect reality [56], [80] (Dreyfus, 1972; Schlinger, 1992).

Artificial intelligence is the core of search engines [81] (Liao, Chu, & Hsiao, 2012) but what we would like to assess in this section is how clever these algorithms really are and how much they simplify users' task.

According to the Oxford English dictionary intelligence is "the ability to acquire and apply knowledge and skills". Intelligence is rational, a goal-driven behaviour. A child will develop and learn from its interaction with the environment as well as other individuals. Search engines such as those using LSA technology are self-learning through documents ingestion. George, the AI from Pantros IP, gets around 40,000 new documents per week but what about its interaction with other intelligent entities. The ultimate goal is to have a machine able to reason with the user. However this can only be achieved if there is the output results are understood and will receive feedbacks to improve the whole process. Users need to assess AI judgements reliability. Tait and Diallo proposed the concept of *searcher trust* that "would require explanations of the internal processing mechanism leading the displayed results to be accessible and comprehensible to the searcher" (chapter 20 [8] (Lupu, 2011)). This interaction does not exist and professionals have to blindly trust results and analysis. It is well known that when comparing relevant documents from two professionals only 40% will overlap [8] (Lupu, 2011). But how is it with search engines? What are those relevant documents that I am presented with? Should I use them like that or would it be reasonable to reformulate a query and start again? These are logical question one can expect from two professionals questioning each other's judgements. What about this new player that is AI? Is it to assume that its judgments are not to be questioned?

7 MOST NEEDED EVOLUTIONS AND FUTURE DIRECTIONS

7.1 POSITIONING USERS AT THE CENTRE OF THE SEARCH PROCESS

One apparent but striking aspect of IP search is the appearing lack of common efforts to facilitate the professional searchers' task. As seen in the previous chapters, professionals generally have to juggle with multiple search engines, interfaces, document collections, query syntaxes, patent institution classification, languages in an ever growing competitive industry. Nothing seems to be done to simplify the different steps of the whole process.

The counter argument is that this is probably a perfectly justified position coming from the non-expert in the field. The professional, as the Doctor with the "antediluvian" stethoscope, needs that proximity and wants to develop a feeling for where she/he is heading to. The number of steps to reach a never predictable objective is a joy for most of the professionals who constantly learn from a journey during which they don't want to be GPS guided at the risk and anxiety therewith of no longer been able to read a map.

The often heard "don't worry about your engine selection, pick one and go" illustrates how the field is organised. Innovation is the foundation of industry progress with IP laws to protect and propagate knowledge. The industrial applicability erected as the third patentability pillar and the associated requirement is not negotiable. In order to help companies and communities at large to use their maximum creative potential, efforts must be made to simplify the access, retrieval, analysis and use of all available information.

We believe that positioning the user at the centre of the IP search development is critical to promote innovative ideas. It is then decisive to favour the adoption of automated *in silico* search to assure that the professional "by no means" feels at the merci of her/his GPS and ends-up in a dead-end of a dirt bike road; just at a few 100's yards from the final destination.

7.2 THE SEMANTIC WEB INITIATIVE: OTHERS FAILURES CAN BE INSPIRING

Semantic Web was proposed in 2001 by Tim Berners-Lee, inventor of World Wide Web. His motivation was to index semantically the whole information present on the Web and make it meaningful for computer programs [82] (Berners-Lee, Hendler, & Lassila, 2001). Unfortunately, more than 10 years later it still did not happen. Nevertheless it is still active with a collective effort under the lead of the World Wide Web Consortium (W3C) which encourages the inclusion of semantic content to web pages. The idea of indexing the whole knowledge following standardized rules is of great interest for the IP domain. Every web document would have its attached machine-readable description. There would be no need to process text anymore; knowledge would be the entity machines would work with. This brilliantly simple idea is far from being a reality for many reasons but one raised issue merits all our attention. "The Semantic Web will never work because *it depends on businesses working together, on them cooperating*" [83] (Downes, 2007). This simple fact illustrates a fundamental issue that also needs to be addressed by the IP community.

Only a collaborative and standardized movement will make today's developments worth the efforts. What is needed is what misses for the Semantic Web to succeed, meaning "simple, well-defined, measurable, widely understood tasks; specific practices and approaches for performing them; tools, resources, and industry support" [84] (Kiryakov et al., 2004). The initiative may not come from relatively independent scientists working in their labs, rather remote from the business and industry prime objectives in that field. Industry and business support may be necessary to boost the initiative wherever it comes from. IP search entities might lead the way to a real evolution, if not a revolution, in the treatment of relevant information. One should not stop at a high-level vision but it should be combined with high-level standards with legal approval and integration, of course.

7.3 NEW SEARCH TECHNOLOGIES REQUIRE NEW EVALUATION TOOLS

The whole field of IP search is slowly moving away from keyword only based technologies and the evaluation has to progress towards new ways of measuring a search engine ability. Effectiveness has been for many years the main value evaluations would take into account as seen in chapter 2.4. This might be a useful thing to consider when retrieving "relevant documents" is the main goal but is of poor value when one wants to go further into the potential of this technology. IP search users want more from their software than simply finding results, they want answers. Professionals use these tools to help them planning and generating innovative ideas. There is therefore a need for a new appropriate way of evaluating search tools.

One major potential of search technologies especially in the IP domain is the ability to predict technological trends [16], [17], [45] (Yoon & Kim, 2011a; Yoon et al., 2012; Choi et al., 2011). We propose here a new evaluation test based on a simple idea: if the software would have been used at a particular moment in the past, could the technological breakthrough of that time period have been anticipated?

Here the test corpus would contain all available prior art documents from a major technological event. It should not only focus on documents' relevance but on being exhaustive and should include any kind of data and not only patents. The main idea is to move away from a first keyword-based selection of data that reduces the collection size only to "relevant documents". Such large collection already exists for patents but the tests are usually done on a small subset. CLEF-IP 2010 had 2.6 million patents in XML format but tests were only done on subsets of 500 to 2,000 documents. Considering storage and computing abilities of modern super-computers, handling big corpus is not anymore a technical limitation. This type of test corpus would have, to our opinion, several major advantages. The first one is the possibility to finally create what the Semantic Web intends to do but on a smaller scale than the whole web. Coordinating efforts on these data would be useful to come up with standardized index parameters for the IP domain. On the other hand working on very large dataset might implicate different algorithms since it is known that the behaviour of AI programs vary when acting on small or big collections. In the proposed model there would be no pre-selection of documents but the system would work on everything available in the corpus. The idea here is that new innovative concepts and breakthrough might not be detectable using only topic related data. The final measurement of this kind of evaluation would be prediction abilities and visualization capability.

Of course this rather broadly conceptualized idea would be perfected using noise reduction technology involving measurement of the noise impact and its value, possibly. One may "soundboard" the idea by referring to the work published by Nature in February 2009 on the "Detection of influenza epidemics using search engine query data".

7.4 AUTONOMY THROUGH DYNAMIC AND INTERACTIVE SYSTEMS

The currently available tools in the field of information retrieval even when using technologies more solution-oriented stick to a unidirectional kind of user-machine interaction. The user enters its query, the machine retrieves, processes and gives a result. IP domain is in need of adaptive software able through interactions with the users to learn and evolve to more autonomous entities. Once more it is not necessary to go very far to see examples of technologies that could be the future of IP search. Recently Apple launched a new application named Siri. Siri is the first commercially available AI that uses human-like ways of interaction [85] (Weaver, 2012). Unlike traditional search engines that necessitate screen and keyboard, Siri interacts with spoken language. It is then possible to query a system using natural language instead of lists of keywords. Ambient intelligence is a new evolution of AI that "refers to a digital environment that proactively, but sensibly, supports people in their daily lives" [86] (Ramos, Augusto, & Shapiro, 2008). Here the AI interprets its environment through sensorial sources, retrieves and represents the associated knowledge then plans, takes actions and learns from it. Its applications spread from smart homes to health monitoring and assistance [87] (Cook, Augusto, & Jakkula, 2009).

"Although these systems tend to be currently used as classification assistants, the ultimate goal for researchers in the field is to provide fully unsupervised systems, for instance to build pre-classification tools or to classify large volumes of patents in batch mode" (chapter 20 [8] (Lupu, 2011)). If one goes even further, automated systems will be able to plan, act, adapt, and be autonomous. This is already a reality in some field of research in biology. King and colleagues published in 2009 an article describing a robot scientist named Adam [88] (King et al., 2009). Adam formulates and tests hypothesis without human intellectual intervention. The IP domain would benefit from automated hypothesis particularly considering the multi-domain literature and concepts to manipulate [89] (Evans & Rzhetsky, 2010). The whole methodology of innovation is in need of a revolution and tools are available, one needs to put them together.

7.5 WHAT CAN WE LEARN FROM TRIZ?

The "theory of inventive problem solving" (from Russian TRIZ) has been developed by a Russian patent examiner Genrich Altshuller. The concept on which lie this theory is the following: "existing methods of inventing are so bad that they should be replaced" [90] (Altshuller, 1999). The innovative process tends to follow the same trial and errors schema throughout the ages. Despite people making some great inventions this is neither efficient nor satisfactory. Most of the time this rather limits the chances to solve "inventive" problems. Nowadays trial and errors remains a method which is still rather extensively used for scientific research, although most of the time accompanied by a design of experiments (DOE) which improves convergence. To overcome these issues the industry introduced, for a long time already, methods to enhance creativity and networking such as brainstorming. Yet the outcome still leaves some room for improvement. TRIZ, already in place for some time, tries to go further and proposes to define a methodology for invention based on knowledge. Because TRIZ approaches problems from an algorithm point of view, it is more repeatable, more predictable and more reliable as opposed to other methods [91] (http://www.triz-journal.com/archives/what_is_triz/). The purpose of this section is not to praise TRIZ in particular but to attract the attention of the reader on the afterwards of information retrieval and analysis. TRIZ is one of the first "real" attempts to rationalize the innovation procedure. Innovation methodology needs to remain subject to careful scientific examination. It is now part of few search engines such as IHS Goldfire. We believe that these tools are a necessary evolution for the whole industry to continue its mutation to the next level of creativity in an open collaborative environment making use of the tremendous information available in various forms, such as images in patent literature and problem solving motivations at disposal in such literature and in patent specifications in particular.

From an academic stand point the above is even more legitimate and has to be mentioned in a review; additional what is more natural from a researcher standpoint to dream of a research recipe? To the tribute of TRIZ creators, the genesis of that method started from a clear understanding that problem solvers are good creators. Several successful business entrepreneurs developed computerized social networks from a problem solving approach; most of the time "improving and transferring" rather than discovering.

Opposite to some academic research framework, keeping a honeycomb search approach with lack of adjacent communication networks between the cells, industry and corporate research more specifically, have developed some level of hierarchy in their organization; leaving all research results more or less accessible to the researcher at large.

Some of those structures have naturally or by design evolved towards research crews with well identifiable profiles such as:

- The legendary **gate-keeper** focusing year after year on the maintenance and evolution of trade secrets and patent fortress therewith.
- The **coaching style technology master** (CSTM), capable of creating year after year evolving generational technology packages with awareness of the routes to market and the business models. The CSTM is well versed to communication, nurturing and empowerment of less experienced innovator generations or to help redirect already in place more experienced researchers.
- **Freelance profiles**, sometime perceived as "unmanageable" intrapreneurs (conversely entrepreneurs), take part to such controlling organization and are likely to bring non-core innovative solutions. They come generally from the problem solver side of the research organization; generally originating from the sometime less appreciated problem spotters, i.e. "trouble makers".

Earlier on, back in the 1990 [1] (Rebouillat, 2013), was developed and further adapted the Z-process, as a business development process, which starts from the VOC (Voice Of the Customer) analysis. Such VOCs rarely yield lack of clarity regarding product improvement opportunities from a customer standpoint; the latter is generally eager to sell the best products then requiring the best solutions from her/his suppliers and her/his research approach.

The Z-process includes the following steps:

- Market Survey
- Target Definition – Research and market Targets
- Speculative Research and Discovery Process
- Process Development
- Pilot Scale Manufacture
- Full Scale Manufacture
- Launch Phase
- Establishment Phase

The details of that process is out of the scope of this article, nonetheless one would wonder how a trial-and-error defiance would survive in such an organized research and business development, with rigorous time table, critical objective tasks, and CTQs; i.e. Critical To Quality parameters representing the product or service characteristics as defined by the customer/ user.

Rebouillat advocates the discipline of adopting an ATA© [1] (Rebouillat, 2013) as a process yielding the integration of IP and NPL in the business model and roadmaps therewith. The key ATA© associated “search” phases imply to:

- Conduct an Xpress inventory of science and technology portfolio at disposal, being core or adjacent;
- Reconsider Unexploited assets;
- Identify Unexplored areas;
- Formulate Ideal and dreamed approaches;
- Express Perceived Limitations & Frustrations;

Sequences, which involve 8 to 12 participants, 4 hour session repeated 4 times with a convergence mind-set, generally insure awareness and cross team interactions without being over time consuming.

Once more this approach will unlikely tolerate a lack of knowledge of the available prior art, core or adjacent. One may as well reconsider, in an open innovation perspective, the need to avoid giving dominancy to internal vs. external comparative analysis in favour of an external vs. internal focus; having the advantage of promoting a modest profile more permeable to others’ contributions.

This does not mean that the information search, retrieval and exploitation, is handy, since improvement is definitely needed in this area as seen all along this work.

8 LEGAL ISSUES

IP is a legal concept and issues will naturally rise from a global use of intelligent search technologies. Two main points merit our immediate attention. The first one is related to the manipulation of data. The biggest challenge is to accompany the move of IP search from an internal task where confidentiality stays under a certain level of control, to external entities that would need a full access of in-house information (Chapter 20 [8] (Lupu, 2011)). Outsourcing data processing to businesses that most likely will manipulate information coming from direct competitors is a serious problem.

The second issue is AI itself and the creation of IP through human-machine partnership. Already in 1991 the WIPO had a symposium entitled “The intellectual property aspect of artificial intelligence” [9] (WIPO, 1991). The issue of ownership of IP from AI work has been recently raised again with the launch of Siri from Apple. This software is the first of many and will necessitate a re-examination of how we grant IP rights of AI creations as well as how we determine and assign liability [85] (Weaver, 2012). In anticipation of such events people have proposed to give a new legal status to AI and recognised them as individuals [78] (Davies, 2011). Law has to be part of the whole process not to restrict creativity or innovation but to accompany it to another level.

In this review we have clearly identified the key improvement in the information retrieval related to patent literature and NPL. The semantic image analysis (SIA) is necessary and has also to be adopted from a legal standpoint. There is no secret about the discomfort of some patent agents when dealing with images and drawings in patents; this in spite of guidance strongly encouraging applicants to include drawings and the risk of rejection for incompleteness.

Furthermore the guide for the preparation of patent drawings, available on the USPTO site, is a 135 page document with series of rules, which add to the complexity for the non-represented applicant at least.

The reluctance may be also attributed to the misinterpretation potential and the creation of new jurisprudential boundaries with the proliferation of images, versus annotated drawings, in patent applications.

“The Eye Alone Is the Judge” in the address of Rebecca Tushnet at Georgetown University Law Centre Conference in 2012, summarizes the most recent view on the matter in the following terms:

“But what does it mean for the eye to be the judge in a legal system organized around words? How can the report of an eye be turned into a verdict, and further into a reviewable judgment?”

A picture is worth a thousand words, which refers to the notion that a complex idea can be conveyed with just a single still image, is by far not yet the adage of the patent legal environment.

9 Conclusion

Being an IP researcher is becoming more and more complex with the explosion of data that need to be reviewed. Not only it is necessary to be an expert in different domains, to be aware of the diversity of the available literature, there is also a need to find the best strategy to retrieve the information. This later being at the centre of the whole process and is not an easy task.

Many obstacles, some of them depicted in this review, face strategists and IP researchers. Nevertheless companies are still able to generate new patents and move always forward the limits of innovation. Furthermore using automated search tools is still relatively young and needs to find its own pace. New technologies are becoming available and soon will become an essential part of any corporation that desires to grow and protect its interests at the same time. These new developments to be fully beneficial need to be more user-friendly and focused on slowly transferring the burden of data gathering to *in silico* researchers. And this will have to go through a simplification of the human-machine interactions. An exciting revolution faces the whole industry and IP search using AI is a key component.

We have underlined two elements that may be central to a near future success in the field.

One is the adjacent technology analysis, ATA©, which means that the position or distribution of patents (or IP in general) in an enlarged business environment may be more important than the value of the core technology indicators. Recurrent Xpress inventories of the core technology offering and reflection on unexplored, unexploited and ideal solutions, well balanced by the honest recognition of the frustrations and missing new capabilities, constitute the essence of the method. ATA© is the opportunity to stay in touch with the IP world from a technology standpoint as well as from a legal one without excessive lawyering.

The second element outlined in this review is the urgent need for SIA, in other words put the image at a central place in the IP search process and promote the IP via pictorial description. This in order to harmonize information and get closer to the non IP world made of pictures and motions rather than 1000's of words. The legal dimension is to be integrated naturally.

There is a tremendous opportunity to remove the fear associated with the open-innovation concept. The IP and NPL search harmonization can be enabling via semantic and image analysis. Finally, why not adopting a more holistic approach of data usage in the context of IP by integrating larger amount of data without doing screening from keywords and relevance standpoint? Instead elect for a well proportionated Semantic Web like approach. *A long way to go and an exciting journey onward!* And more papers to come on the influence of such approaches on the process of innovation.

We highly recommend the additional reading, of about 56 references, provided at the end of the reference list. More details on related specific matters are available in those.

Acknowledgements

The authors thank:

Benoit Steffenino for his thoughtful comments on figures and outstanding illustrative interpretations of some initially basic draft drawings; those greatly enhance this work and provide pleasurable reading.

Imre Horvath, eminent business strategist; for numerous business and IP strategy interactions and casual discussions on the matter.

Miroslawa Lapray, for informal conversations yielding useful improvements.

More about the authors:

Serge Rebouillat [Rebooya], Dr. Ing., Dr-ès-Sciences, Certified Prof., Ind. Energetic, Chem/Bio Eng., Rheology, IP/Mediation/Innovation & Strategy.

Damien Lapray, Ph.D., Neural Network Analysis.

References

[1] S. Rebouillat, "A Science & Business Equation for Collaborative Corporate Innovation . Business Strategy, IP Strategy, R&D Strategy: an all-in-one Business Model. A review with a Bio-Technology & Green Chemistry Focus." *Int. J. Innov. Appl. Stud.*, vol. 4, no. 1, pp. 1–19, 2013.

[2] D. (PRWEB) WILMINGTON, "A Record Year for DuPont Innovation," 2012. [Online] Available: http://www.prweb.com/releases/2012/3/prweb9280362.htm (2012).

[3] "Statistics and trends." [Online] Available: http://www.epo.org/about-us/annual-reports-statistics/annual-report/2012/statistics-trends.html.

[4] E. Garfield, "A tribute to Calvin N. Mooers, a pioneer of information retrieval," *Sci.*, 1997.

[5] Netcraft.com, "October 2013 Web Server Survey," 2013. [Online] Available: http://news.netcraft.com/archives/2013/10/02/october-2013-web-server-survey.html (2013).

[6] Elsevier, "Elsevier for authors," 2013. [Online] Available: http://www.elsevier.com/journal-authors/home (2013).

[7] PwC, "2013 patent litigation study," 2013. [Online] Available: http://www.pwc.com/en_US/us/forensic-services/publications/assets/2013-patent-litigation-study.pdf (2013).

[8] M. Lupu, *Current challenges in patent information retrieval*, ISBN: 978-3-642-19230-2, vol. 29. 2011.

[9] WIPO, "The intellectual property aspect of artificial intelligence," in *WIPO worldwide symposium*, 1991.

[10] T. Magerman, B. Looy, and X. Song, "Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications," *Scientometrics*, vol. 82, no. 2, pp. 289–306, Jun. 2009.

[11] B. Spyropoulos and M. Botsivaly, "Improving the interpretation of Patent Data by the employment of the Semantic-Web formalism in the specific context of Biomedical Technology," *IFMBE Proc.*, pp. 257–260, 2009.

[12] J. Liu, A. Harris, and N. Kanwisher, "Stages of processing in face perception: an MEG study.," *Nat. Neurosci.*, vol. 5, no. 9, pp. 910–6, Sep. 2002.

[13] W. Zhao and R. Chellappa, "Face recognition: A literature survey," *Acm Comput. Surv.*, vol. 35, no. 4, pp. 399–458, 2003.

[14] W. M. Wang and C. F. Cheung, "A Semantic-based Intellectual Property Management System (SIPMS) for supporting patent analysis," *Eng. Appl. Artif. Intell.*, vol. 24, no. 8, pp. 1510–1520, Dec. 2011.

[15] J. Yoon and K. Kim, "Detecting signals of new technological opportunities using semantic patent analysis and outlier detection," *Scientometrics*, vol. 90, no. 2, pp. 445–461, Oct. 2011.

[16] J. Yoon and K. Kim, "Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks," *Scientometrics*, vol. 88, no. 1, pp. 213–228, Apr. 2011.

[17] J. Yoon, H. Park, and K. Kim, "Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis," *Scientometrics*, vol. 94, no. 1, pp. 313–331, Sep. 2012.

[18] A. J. C. Trappey, C. V. Trappey, and C.-Y. Wu, "Automatic patent document summarization for collaborative knowledge systems and services," *J. Syst. Sci. Syst. Eng.*, vol. 18, no. 1, pp. 71–94, Mar. 2009.

[19] J. Tang, B. Wang, Y. Yang, P. Hu, and Y. Zhao, "Patentminer: Topic-driven patent analysis and mining," *KDD'12*, 2012.

[20] D. M. Jessop, S. E. Adams, and P. Murray-Rust, "Mining chemical information from open patents.," *J. Cheminform.*, vol. 3, no. 1, p. 40, Jan. 2011.

[21] D. Archibugi and M. Planta, "Measuring technological change through patents and innovation surveys," *Technovation*, vol. 16, no. 9, pp. 451–468, 1996.

[22] L. Wanner, R. Baeza-Yates, S. Brügmann, J. Codina, B. Diallo, E. Escorsa, M. Giereth, Y. Kompatsiaris, S. Papadopoulos, E. Pianta, G. Piella, I. Puhlmann, G. Rao, M. Rotard, P. Schoester, L. Serafini, and V. Zervaki, "Towards content-oriented patent document processing," *World Pat. Inf.*, vol. 30, no. 1, pp. 21–33, Mar. 2008.

[23] C. Wu and C. Yao, "Constructing an Intelligent Patent Network Analysis Method," *Data Sci. J.*, vol. 11, no. November, pp. 110–125, 2012.

[24] M. A. Newman and T. L. Works, "Method for syntactic and semantic analysis of patent text and drawings," US 5,774,833 (1998).

[25] A. Williams, "System and method for patent evaluation using artificial intelligence," US 8,161,049 B2 (2012).

[26] J. Codina, E. Pianta, S. Vrochidis, and S. Papadopoulos, "Integration of Semantic, Metadata and Image Search Engines with a Text Search Engine for Patent Retrieval.," *SemSearch*, vol. 334, 2008.

[27] D. Greenbaum and M. Gerstein, "Semantic web standards: Legal and social issues and implications," in in *Revolutionizing Knowledge Discovery in the Life Sciences*, 2007, pp. 413–433.

[28] O. a Jefferson, D. Köllhofer, T. H. Ehrich, and R. a Jefferson, "Transparency tools in gene patenting for informing policy and practice.," *Nat. Biotechnol.*, vol. 31, no. 12, pp. 1086–93, Dec. 2013.

[29] A. Rzhetsky, M. Seringhaus, and M. B. Gerstein, “Getting started in text mining: part two.,” *PLoS Comput. Biol.*, vol. 5, no. 7, p. e1000411, Jul. 2009.

[30] M. Lupu, *Patent Retrieval*, ISBN 978-1-60198-648-1, vol. 7, no. 1. 2013, pp. 1–97.

[31] L. Leydesdorff, D. Kushnir, and I. Rafols, “Interactive overlay maps for US patent (USPTO) data based on International Patent Classification (IPC),” *Scientometrics*, pp. 1–23, 2012.

[32] T. Chiu, C. Hong, and Y. Chiu, “Using IPC-Based Clustering and Link Analysis to Observe the Technological Directions,” in in *Semantic Methods*, 2011, pp. 183–197.

[33] D. Dirnberger, “A guide to efficient keyword, sequence and classification search strategies for biopharmaceutical drug-centric patent landscape searches - A human recombinant insulin patent landscape case study,” *World Pat. Inf.*, vol. 33, no. 2, pp. 128–143, Jun. 2011.

[34] Intellogist, “IPC classification system.” [Online] Available: http://www.intellogist.com/wiki/IPC_Classification_System.

[35] X. Polanco, C. François, and J. Lamirel, “Using artificial neural networks for mapping of scienceand technology: A multi-self-organizing-maps approach,” *Scientometrics*, vol. 51, no. 1, pp. 267–292, 2001.

[36] D. a. Keim, “Information visualization and visual data mining,” *IEEE Trans. Vis. Comput. Graph.*, vol. 8, no. 1, pp. 1–8, 2002.

[37] A. K. Jain, “Data clustering: 50 years beyond K-means,” *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

[38] M. Mańko, R. Geracitano, and M. Capogna, “Functional connectivity of the main intercalated nucleus of the mouse amygdala.,” *J. Physiol.*, vol. 589, no. Pt 8, pp. 1911–25, Apr. 2011.

[39] D. Dupret, J. O’Neill, and J. Csicsvari, “Dynamic reconfiguration of hippocampal interneuron circuits during spatial learning.,” *Neuron*, vol. 78, no. 1, pp. 166–80, Apr. 2013.

[40] S. Kadir, D. Goodman, and K. Harris, “High-dimensional cluster analysis with the Masked EM Algorithm,” *arXiv.org*, pp. 1–10, 2013.

[41] P. I. Khalid, J. Yunus, and R. Adnan, “Extraction of dynamic features from hand drawn data for the identification of children with handwriting difficulty.,” *Res. Dev. Disabil.*, vol. 31, no. 1, pp. 256–62, 2010.

[42] R. Frijters, M. van Vugt, R. Smeets, R. van Schaik, J. de Vlieg, and W. Alkema, “Literature mining for the discovery of hidden connections between drugs, genes and diseases.,” *PLoS Comput. Biol.*, vol. 6, no. 9, pp. 1–11, Jan. 2010.

[43] W. L. Woon and S. Madnick, “Semantic distances for technology landscape visualization,” *J. Intell. Inf. Syst.*, vol. 39, no. 1, pp. 29–58, Oct. 2011.

[44] D. Escudero-Mancebo and V. Cardeñoso-Payo, “Applying data mining techniques to corpus based prosodic modeling,” *Speech Commun.*, vol. 49, no. 3, pp. 213–229, Mar. 2007.

[45] S. Choi, J. Yoon, K. Kim, J. Y. Lee, and C.-H. Kim, “SAO network analysis of patents for technology trends identification: a case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells,” *Scientometrics*, vol. 88, no. 3, pp. 863–883, Jun. 2011.

[46] R. Rodriguez-Esteban, “Biomedical text mining and its applications.,” *PLoS Comput. Biol.*, vol. 5, no. 12, p. e1000597, Dec. 2009.

[47] D. Bonino, A. Ciaramella, and F. Corno, “Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics,” *World Pat. Inf.*, vol. 32, no. 1, pp. 30–38, Mar. 2010.

[48] Z. Xie and K. Miyazaki, “Evaluating the effectiveness of keyword search strategy for patent identification,” *World Pat. Inf.*, vol. 35, no. 1, pp. 20–30, Mar. 2013.

[49] F. Piroi, M. Lupu, A. Hanbury, W. Magdy, A. P. Sexton, and I. Filippov, “CLEF-IP 2012: Retrieval Experiments in the Intellectual Property Domain,” *CLEF-IP Initiat. EU*, 2012.

[50] C. Hempelmann and V. Raskin, “Semantic Search: Content vs. Formalism,” 2008.

[51] L. Hawizy, D. M. Jessop, N. Adams, and P. Murray-Rust, “ChemicalTagger: A tool for semantic text-mining in chemistry.,” *J. Cheminform.*, vol. 3, no. 1, p. 17, Jan. 2011.

[52] C. D. Manning, R. Prabhakar, and H. Schuetze, *An introduction to information retrieval*, no. c. 2009.

[53] H. Turtle, “Natural language vs. Boolean query evaluation: A comparison of retrieval performance,” *West Publ. Co.*, 1994.

[54] S. Sarawagi, “Information extraction,” *Found. Trends Databases*, vol. 1, no. 3, pp. 261–377, 2007.

[55] M. Minsky, *Semantic Information Processing*. 1969.

[56] H. L. Dreyfus, *What computer can’t do - The limits of artificial intelligence*. 1972.

[57] A. Spinakis and A. Chatzimakri, “Analysis of Biotechnology Patents,” *Knowl. Min.*, vol. 290, no. 2005, pp. 281–290, 2005.

[58] K. Wen, Y. Zeng, R. Li, and J. Lin, “Modeling semantic information in engineering applications: a review,” *Artif. Intell. Rev.*, vol. 37, no. 2, pp. 97–117, May 2011.

[59] E. Mäkelä, “Survey of semantic search research,” pp. 1–18, 2005.

[60] M. G. Moehrle, “Measures for textual patent similarities: a guided way to select appropriate approaches,” *Scientometrics*, vol. 85, no. 1, pp. 95–109, May 2010.

[61] M. G. Moehrle and J. M. Gerken, "Measuring textual patent similarity on the basis of combined concepts: design decisions and their consequences," *Scientometrics*, vol. 91, no. 3, pp. 805–826, Mar. 2012.

[62] J. M. Gerken and M. G. Moehrle, "A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis," *Scientometrics*, vol. 91, no. 3, pp. 645–670, Jan. 2012.

[63] C. Hare, "Towards an Ontology of Intellectual Property," *Am. J. Econ. Sociol.*, 1998.

[64] H. Gurulingappa, B. Müller, and M. Hofmann-Apitius, "Information Retrieval Framework for Technology Survey in Biomedical and Chemistry Literature.," *TREC*, no. Xml, pp. 1–4, 2011.

[65] J. R. Wetherell and M. Mehok, "United States: Claim Construction: The Dictionary/Specification Dichotomy," 2005. [Online] Available: http://www.mondaq.com/unitedstates/x/32125/Patent/Claim+Construction+The+DictionarySpecification+Dichotomy (2005)

[66] T. Landauer, P. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse Process.*, vol. 25, no. 2, pp. 259–284, 1998.

[67] M. Fernández Sánchez, "Semantically enhanced Information Retrieval: an ontology-based approach," 2009.

[68] A. Mahabal, N. Takahiro, Z. Garrett, and K. Inoue, "Synonym identification based on co-occurring terms," US 8,538,984 (2013).

[69] H. Chen, B. Martin, C. M. Daimon, and S. Maudsley, "Effective use of latent semantic indexing and computational linguistics in biological and biomedical applications.," *Front. Physiol.*, vol. 4, no. January, p. 8, Jan. 2013.

[70] H. Henderson, *Encyclopedia of Computer Science and Technology*, vol. 17, no. 8. 2009, pp. 38–39.

[71] N. Bhatti and A. Hanbury, "Image search in patents: a review," *Int. J. Doc. Anal. Recognit.*, Dec. 2012.

[72] I. Kollia, N. Simou, A. Stafylopatis, and S. Kollias, "Semantic Image Analysis Using a Symbolic Neural Architecture," *Image Anal. Stereol.*, vol. 29, no. 3, pp. 159–172, Nov. 2010.

[73] N. Chomsky, Lecture at the University of Geneva Grandes Conferences. "The Language-Cognition Interface", International Congress of Linguists, Geneva July 2013.

[74] T. M. Sezgin and R. Davis, "Sketch recognition in interspersed drawings using time-based graphical models," *Comput. Graph.*, vol. 32, no. 5, pp. 500–510, Oct. 2008.

[75] D. G. Fernández-Pacheco, F. Albert, N. Aleixos, and J. Conesa, "A new paradigm based on agents applied to free-hand sketch recognition," *Expert Syst. Appl.*, vol. 39, no. 8, pp. 7181–7195, Jun. 2012.

[76] F. Brieler and M. Minas, "A model-based recognition engine for sketched diagrams," *J. Vis. Lang. Comput.*, vol. 21, no. 2, pp. 81–97, Apr. 2010.

[77] L. Olsen, F. F. Samavati, M. C. Sousa, and J. a. Jorge, "Sketch-based modeling: A survey," *Comput. Graph.*, vol. 33, no. 1, pp. 85–103, Feb. 2009.

[78] C. R. Davies, "An evolutionary step in intellectual property rights – Artificial intelligence and intellectual property," *Comput. Law Secur. Rev.*, vol. 27, no. 6, pp. 601–619, Dec. 2011.

[79] H. D. Schlinger, "What Would It Be Like to Be IBM's Computer, Watson?," *Behav. Anal.*, vol. 1, no. 1, pp. 37–44, 2012.

[80] H. D. Schlinger, "Intelligence: Real or artificial?," *Anal. Verbal Behav.*, vol. 10, no. 1985, pp. 125–33, Jan. 1992.

[81] S.-H. Liao, P. Chu, and P. Hsiao, "Data mining techniques and applications–A decade review from 2000 to 2011," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 11303–11311, Sep. 2012.

[82] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Sci. Am.*, 2001.

[83] S. Downes, "Why the Semantic Web Will Fail," 2007. [Online] Available: http://halfanhour.blogspot.ch/2007/03/why-semantic-web-will-fail.html (2007).

[84] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff, "Semantic annotation, indexing, and retrieval," *Web Semant. Sci. Serv. Agents World Wide Web 2*, vol. 2, no. 1, pp. 49–79, Dec. 2004.

[85] J. Weaver, "Siri Is My Client: A First Look At Artificial Intelligence And Legal Issues," *Mondaq*, 2012. [Online] Available: http://www.mondaq.com/unitedstates/x/176106/Telecommunications+Mobile+Cable+Communications/SIRI+Is+My+Client+A+First+Look+At+Artificial+Intelligence+And+Legal+Issues (2012).

[86] C. Ramos, J. C. Augusto, and D. Shapiro, "Ambient Intelligence—the Next Step for Artificial Intelligence," *IEEE Intell. Syst.*, vol. 23, no. 2, pp. 15–18, Mar. 2008.

[87] D. J. Cook, J. C. Augusto, and V. R. Jakkula, "Ambient intelligence: Technologies, applications, and opportunities," *Pervasive Mob. Comput.*, vol. 5, no. 4, pp. 277–298, Aug. 2009.

[88] R. D. King, J. Rowland, S. G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. N. Soldatova, A. Sparkes, K. E. Whelan, and A. Clare, "The automation of science.," *Science (80-.).*, vol. 324, no. 5923, pp. 85–9, Apr. 2009.

[89] J. Evans and A. Rzhetsky, "Philosophy of science. Machine science.," *Science (80-.).*, vol. 329, no. 5990, pp. 399–400, Jul. 2010.

[90] G. Altshuller, *The Innovation Algorithm*. MassachusettsTechnical Innovation Center, Inc. ISBN-10: 0964074044, ISBN-13: 978-0964074040, 1999.

[91] K. Barry, E. Domb, and M. S. Slocum, "TRIZ - What Is TRIZ?," *The TRIZ Journal*. [Online] Available: http://www.triz-journal.com/archives/what_is_triz/

RECOMMENDED ADDITIONAL READING

Anonymous. (n.d.). Patent Search Strategies : Keywords or Classifications ? http://patex.ca/pdf/publications/Patent Search Strategies.pdf

Archibugi, D., & Planta, M. (1996). Measuring technological change through patents and innovation surveys. *Technovation*, *16*(9), 451–468.

Bekmambetov, T. (2013). Visualisation of natural language text. Patent WO 2013/032354 A1

Benson, C. L., & Magee, C. L. (2012). A hybrid keyword and patent class methodology for selecting relevant sets of patents for a technological field. *Scientometrics*, *96*(1), 69–82. doi:10.1007/s11192-012-0930-3

Bogal-Allbritten, E. (2013). Decomposing notions of adjectival transitivity in Navajo. *Natural Language Semantics*, *21*(3), 277–314. doi:10.1007/s11050-012-9093-2

Cable, S. (2013). Beyond the past, present, and future: towards the semantics of "graded tense" in Gĩkũyũ. *Natural Language Semantics*, *21*(3), 219–276. doi:10.1007/s11050-012-9092-3

Cohen, K., & Hunter, L. (2008). Getting started in text mining. *PLoS computational biology*, *4*(1), 1. doi:10.1371/journal.pcbi.0040020

Corporation, F. F. (2013). New Patent For Decision Making Using Artificial Intelligence Issues With Financial Firebird Corporation Ready To Launch The First Public Use To Help People In Debt.

Deerwester, S., & Dumais, S. (1989). Computer information retrieval using latent semantic structure. *US Patent …*.

Fattori, M., Pedrazzi, G., & Turra, R. (2003). Text mining applied to patent mapping: a practical business case. *World Patent Information*, *25*(4), 335–342. doi:10.1016/S0172-2190(03)00113-3

George, B. R. (2013). Which judgments show weak exhaustivity? (And which don't?). *Natural Language Semantics*. doi:10.1007/s11050-013-9098-5

Ginsberg J. et al., Detecting influenza epidemics using search engine query data. Nature 457, 1012-1014 (19 February 2009). doi:10.1038/nature07634

Govada, S., & Kandekar, S. (2008). Patent analysis of RFID technology. In *MIS 580 – Knowledge Management Project*.

Greenbaum, D., & Gerstein, M. (2007). Semantic web standards: Legal and social issues and implications. In *Revolutionizing Knowledge Discovery in the Life Sciences* (pp. 413–433).

Guha, R., McCool, R., & Miller, E. (2008). Semsearch. *5th European Semantic Web Conference*.

He, J., Jifeng, X., Zhilei, R. E. N., Youxi, W. U., & Xindong, W. U. (2012). Misleading Classification. *Sci China Inf Sci*, *55*(12), 1–16.

Hehenberger, M., Ph, D., & Coupet, P. (1998). Text mining applied to patent analysis, 1–13.

Hersh, W. R., & Hickam, D. H. (1995). An evaluation of interactive Boolean and natural language searching with an online medical textbook. *Journal of the American Society for Information Science*, *46*(9), 719–719.

Isa, D., Blanchfield, P., & Chen, Z. (2009). Intellectual Property Management System for the Super-Capacitor Pilot Plant. *IC-AI*, 37–42.

Jiang, Y. (2014). analysis on the application of artificial intelligence technology in modern physical education. *Information Technology Journal*.

Lahm, A., & de Rinaldis, E. (2007). PatentMatrix: an automated tool to survey patents related to large sets of genes or proteins. *Source code for biology and medicine*, *2*, 4. doi:10.1186/1751-0473-2-4

Lamirel, J., & Shehabi, S. Al. (2003). Intelligent patent analysis through the use of a neural network: experiment of multi-viewpoint analysis with the MultiSOM model. *Proceedings of the ACL-2003 Workshop on Patent Corpus Processing*, 7–23.

Lange, N., Bishop, C. M., & Ripley, B. D. (1997). Neural Networks for Pattern Recognition. *Journal of the American Statistical Association*, *92*(440), 1642. doi:10.2307/2965437

Lucio-Arias, D., & Leydesdorff, L. (2007). Knowledge emergence in scientific communication: from "fullerenes" to "nanotubes." *Scientometrics*, *70*(3), 603–632. doi:10.1007/s11192-007-0304-4

Meżyk, E., & Unold, O. (2011). Machine learning approach to model sport training. *Computers in Human Behavior*, *27*(5), 1499–1506. doi:10.1016/j.chb.2010.10.014

Mende, M. (2010). Method for semantic processing of natural language using graphical interlingua. *WO Patent 2,010,051,966*.

Mika, P., Elfring, T., & Groenewegen, P. (2006). Application of semantic technology for social network analysis in the sciences. *Scientometrics*, *68*(1), 3–27.

Miron, A., Gensel, J., & Villanova-Oliver, M. (2010). Semantic Analysis for the Geospatial Semantic Web. In *Advances in Knowledge Discovery and Management* (pp. 287–306).

Mustafa, J., Khan, S., & Latif, K. (2008). Ontology based semantic information retrieval. *2008 4th International IEEE Conference Intelligent Systems*, 22–14–22–19. doi:10.1109/IS.2008.4670473

Newman, M. A., & Works, T. L. (1998). Method for syntactic and semantic analysis of patent text and drawings.

Osborne, G., & Chin, S. (2002). Artificial intelligence system for genetic analysis. *US Patent 2012/0177261 A1*

Palangkaraya, A., Webster, E., & Jensen, P. (2011). Misclassification between patent offices: Evidence from a matched sample of patent applications. *Review of Economics and Statistics*, *93*(August), 1063–1075.

Papić, V., Rogulj, N., & Pleština, V. (2009). Identification of sport talents using a web-oriented expert system with a fuzzy module. *Expert Systems with Applications*, *36*(5), 8830–8838. doi:10.1016/j.eswa.2008.11.031

Patinkin, S. (2009). Method, apparatus, and system for clustering and classification. *US Patent 7,574,409*.

Poltorak, A. (2007). Apparatus and method for identifying and/or for analyzing potential patent infringement. *US Patent 7,296,015*.

Pyle, D. (2003). Data Mining, 273–273. doi:10.1016/B978-155860653-1/50011-7

Russo, D. (2011). Global Product Development. (A. Bernard, Ed.). doi:10.1007/978-3-642-15973-2

Rzhetsky, A, Seringhaus, M., & Gerstein, M. (2008). Seeking a new biology through text mining. *Cell*, *134*(1), 9–13. doi:10.1016/j.cell.2008.06.029.Seeking

Scellato, G., Calderini, M., Caviggioli, F., Franzoni, C., & Ughetto, E. (2011). *Study on the quality of the patent system in Europe* (pp. 1–194).

Smucker, M. D., Allan, J., & Carterette, B. (2007). A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. In *CIKM'07*.

Srikrishna, D., & Coram, M. (2013). Finding relevant documents. *US Patent 20,130,031,088*.

Sternitzke, C. (2009). Patents and publications as sources of novel and inventive knowledge. *Scientometrics*, *79*(3), 551–561. doi:10.1007/s11192-007-2041-0

Trappey, Amy J C, Trappey, C. V, & Wu, C. (n.d.). A Semantic Based Approach for Automatic Patent Document Summarization, (10608).

Trappey, Amy J.C., Trappey, C. V., & Wu, C.-Y. (2009). Automatic patent document summarization for collaborative knowledge systems and services. *Journal of Systems Science and Systems Engineering*, *18*(1), 71–94. doi:10.1007/s11518-009-5100-7

Vermeir, D., Laenens, E., & Dierick, J. (1988). Applying AI techniques for patent information retrieval. *World Patent Information*, *10*(1), 26–36.

Wang, W. M., & Cheung, C. F. (2011). A Semantic-based Intellectual Property Management System (SIPMS) for supporting patent analysis. *Engineering Applications of Artificial Intelligence*, *24*(8), 1510–1520. doi:10.1016/j.engappai.2011.05.009

Wanner, L., Baeza-Yates, R., Brügmann, S., Codina, J., Diallo, B., Escorsa, E., … Zervaki, V. (2008). Towards content-oriented patent document processing. *World Patent Information*, *30*(1), 21–33. doi:10.1016/j.wpi.2007.03.008

Weaver, J. (2012). Siri Is My Client: A First Look At Artificial Intelligence And Legal Issues. *Mondaq*. Retrieved from http://www.mondaq.com/unitedstates/x/176106/Telecommunications+Mobile+Cable+Communications/SIRI+Is+My+Client+A+First+Look+At+Artificial+Intelligence+And+Legal+Issues

Wen, K., Zeng, Y., Li, R., & Lin, J. (2011). Modeling semantic information in engineering applications: a review. *Artificial Intelligence Review*, *37*(2), 97–117. doi:10.1007/s10462-011-9221-2

Williams, A. (2012). System and method for patent evaluation using artificial intelligence. *US Patent 8,161,049*.

Wilson, J. (2010). Ontology and the regulation of intellectual property. *The Monist*, *93*(3), 450–463.

Wu, CC, & Yao, C. (2012). Constructing an Intelligent Patent Network Analysis Method. *Data Science Journal*, *11*(November), 110–125.

Wu, Chun-yi, Trappey, A. J. C., & Trappey, C. V. (2013). Concurrent Engineering Approaches for Sustainable Product Development in a Multi-Disciplinary Environment. In J. Stjepandić, G. Rock, & C. Bil (Eds.), *Concurrent Engineering Approaches for Sustainable Product Development in a Multi-Disciplinary Environment*. London: Springer London. doi:10.1007/978-1-4471-4426-7

Wyner, A., & Mochales-Palau, R. (2010). Approaches to text mining arguments from legal cases. In Semantic Processing of Legal Texts. Springer.

Zhang, Y., Zhou, X., Porter, A., Gomila, J., & Yan, A. (n.d.). Triple Helix innovation in China's dye-sensitized solar cell industry: hybrid methods with semantic TRIZ and technology roadmapping. *Scientometrics*.

Zhou, X., Han, H., Chankai, I., Prestrud, A., & Brooks, A. (2006). Approaches to text mining for clinical medical records. *Proceedings of the 2006 ACM symposium on Applied computing - SAC '06*, 235. doi:10.1145/1141277.1141330

Zhuhadar, L., & Nasraoui, O. (2010). Multi-language ontology-based search engine. In proceeding of: ACHI 2010, The Third International Conference on Advances in Computer-Human Interactions, Saint Maarten, Netherlands, Antilles, February 10-16, 2010.