This document is classified as *PUBLIC*

DELIVERABLE 2.4:

STATE OF THE ART: INFORMATION RETRIEVAL AND EXTRACTION IN PATENTS

by
K.Vanherck (PNO)
Andrea Ciapetti (Innovation Engineering)
Mirko Calvaresi (Innovation Engineering)

Abstract

This document first summarizes the results of a broad literature search into the topic of 'information retrieval' and 'information extraction', more specifically in the field of patent searching and patent analysis. The second part provides an overview of the technologies and the research projects currently in progress in Innovation Engineering in the fields of semantic search and text mining.

Approval status

Author:	Katrien Vanherck Andrea Ciapetti	PNO INNEN	Researcher seconded from PNO to PoliMI Researcher at INNEN
Approved by:	Gaetano Cascini	PoliMI	Project Coordinator

Dissemination

PU

Document details

Issue Date:	**31.01.2013**
Contract n°:	**PIAP-GA-2011-286305**
Project n°:	**286305**

Revision table

Issue	Issue date	Modifications
0.1	28.01.2013	Draft
1.0	31.01.2013	Final issue

Electronic file details

Master file location	FORMAT collaborative space
Filename	FORMAT_TITLE_DATE.doc

0. Index

1. Introduction

This document summarizes the results of a focused state of the art search into the topic of 'information retrieval' and 'information extraction. More specifically, the first part is focused on 'patent searching' and 'patent analysis', the second part overviews the text mining technologies available at INNEN.

1.1. Essential concepts on Text Mining

Text mining is a set of analytic processes for deriving high-quality information from text.
High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning or through the use of authoritative information sources freely available on the Web for annotating and enriching the simple text with references and relations to well-known places, organizations or concepts.

'High quality' information in text mining usually refers to some combination of **relevance, novelty, and interestingness** to be found and discovered in unstructured text. Unstructured information may be defined as the direct product of human communication. Examples include natural language documents, email, speech, images and video. It is information that was not specifically encoded for machines to process, but rather authored by humans for humans to understand. We say it is “unstructured” because it lacks explicit semantics (“structure”) required for applications to interpret the information as intended by the human author or required by the end-user application.

Unstructured information may be contrasted with the information in classic relational databases where the intended interpretation for every field data is explicitly encoded in the database by column headings. Consider information encoded in XML as another example. In an XML document some of the data is wrapped by tags which provide explicit semantic information about how that data should be interpreted. An XML document or a relational database may be considered semi-structured in practice, because the content of some chunk of data, a blob of text in a text field labelled “description” for example, may be of interest to an application, but remain without any explicit tagging—that is, without any explicit semantics or structure.

Unstructured information represents the largest, most current and fastest growing source of knowledge available to businesses and companies worldwide. The web is just the tip of the iceberg. Consider, for example, the droves of corporate, scientific, social and technical documentation including best practices, research reports, medical abstracts, problem reports, customer communications, contracts, emails and voice mails. Beyond these, consider the growing number of broadcasts containing audio, video and speech. These mounds of natural language, speech and video artefacts often contain nuggets of knowledge critical for analysing and solving problems, detecting threats, realizing important trends and relationships, creating new opportunities or preventing disasters.

For unstructured information to be processed by applications that rely on specific semantics, it must be first analysed to assign application-specific semantics to the unstructured content. Another way to say this is that the unstructured information must become “structured” where the added structure explicitly provides the semantics required by target applications to interpret the data correctly.

An example of assigning semantics includes labelling regions of text in a text document with appropriate XML tags that, for example, might identify the names of organizations or products. Another example may extract elements of a document and insert them in the appropriate fields of a relational database or use them to create instances of concepts in a knowledge base.

Typical text mining tasks include several activities or analysis to be performed, such as:

- text categorization,
- concept/entity extraction,
- production of granular taxonomies,
- text clustering,
- sentiment analysis,
- pattern recognition and text annotation or tagging of text from external information source in the Open Linked Data cloud (enrichment)

The task of **text categorization** is to assign to a document one or more main classes or categories, to which the document logically belongs to. The documents to be classified may be texts, images, music, etc. Each kind of document possesses its special classification problems.

Concept extraction is an activity that results in the extraction of topics, that better represent the concepts and the arguments covered in the text of document. This task can be started from a hierarchical taxonomy of general concepts, applicable to the specific context, or can determine automatically the topics basing on term occurrences inside the text.

The task of **text clustering** is to perform an automatic document organization, based on topics extraction for fast information retrieval or filtering. A search engine often returns thousands of pages in response to a broad query, making it difficult for users to browse or to identify relevant information. Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories.

Sentiment analysis aims at determining the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation, or the intended emotional communication, that is the emotional effect the author wishes to have on the reader.

Wikipedia is the largest encyclopaedia available online and offline. Its contents are contributed, edited, and updated by hundreds of thousands of active registered users every month. It contains over three million English articles and about 13 million articles in all languages. DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and to make this information available on the Web in a format compatible with the semantic web standard. DBpedia allows you to make sophisticated queries against Wikipedia and leverages this gigantic source of knowledge by extracting structured information from it and by making this information accessible on the Web under the terms of the Creative Commons Attribution-ShareAlike 3.0 License and the GNU Free Documentation License.

The English version of the DBpedia knowledge base currently describes 3.77 million things, out of which 2.35 million are classified in a consistent Ontology, including 764,000 persons, 573,000 places, etc.
The **semantic annotation process** allows to enrich the text included in your document with annotations related to persons, organizations, places and other concepts taken from the DBPedia repository, in order to add additional knowledge to the content and link unstructured information sources to the Linked Open Data cloud through DBpedia. Other information sources, like Freebase, are interlinked to the DBPedia resources, through semantic relations.

1.2. Information retrieval and extraction from patents

The aim of the activities carried out within the Task 2.4 of the FORMAT project was twofold. The first purpose was to find **existing search strategies for retrieving relevant patents**, so as to collect a large but manageable body of relevant patents for a certain topic. More specifically, this search strategy should be compatible with technological questions that are analysed in terms of product or process modelling for technology forecasting. The output of such an analysis will be likely in terms of ‘components’ of a technology, ‘functions’, ‘features’ and ‘evaluation parameters’. **If**

any search strategies exist that are able to manage searching on these types of criteria, they can be useful to be incorporated into the FORMAT methodology. Scientific literature was studied that discusses how to search and retrieve a relevant, complete but manageable patent body, which does not necessarily only include patents in the same technology field but also in fields where similar 'technological questions' can occur.

Next to looking for search strategies, literature was studied about the **analysis of a large patent body in terms of technology** analysis, technological problems, emerging technologies and so on.

Different **sources** were consulted. The important ones include the following:

- Websites and e-learning material of patent government bodies (USPatent, EPO, WIPO, …)
- Peer-reviewed Journals on patents, information management, IP, …
  - World Patent Information
  - Journal of Intellectual Property Rights
  - PaIR 'XX Proceedings of the Xst ACM workshop on Patent information retrieval
  - Advances in Information Retrieval
  - Technology Forecasting & Social Change
  - Expert Systems with Applications
  - Academy of Management Journal
  - …
- Courses and IP websites of universities and knowledge institutes
- Websites of patent-related commercial companies (legal offices, patent search specialists, patent application support, …)

A list of references is given at the end of this document.

2. General Strategies for Patent Searching

2.1. Patent search objectives

The main objective to do a patent search, is to check the patentability of a new invention or to check possible infringements to an existing patent (=prior art search). Other goals include monitoring of technologies, identifying key players in a field, collecting indicators for commercial success of a product, searching for solutions to technical problems etc.:

- Due diligence search;
- Patentability search;
- Infringement search;
- Freedom to use search;
- Document Status search;
- Product/process specifications search;
- Technology search.

The EPO website specifically mentions 'a **technology search**', which is meant to uncover as many relevant patents as possible related to a technical problem. This can be highly relevant to the needs of the FORMAT project, in the case where the output of forecasting should be to innovate by 'solving a technological problem'. Unfortunately, the EPO website only explains what is meant by a 'technology search' and makes no mention of a specific, practical search strategy to help making a technology search. The definition is as follows:

"A technology search is carried out to find technical solutions to a problem. The result usually contains different, often alternative approaches to solving a technical problem. It is not the complete definition of an area of technology which is in the foreground here, but rather the uncovering of as many different solutions as possible. It is only an overview of the available approaches and technologies that is required, although as many relevant documents as possible should be retrieved."

In summary EPO mentions that "an exact description and definition of the subject matter you are searching for" is needed to conduct a 'technology search' and that, when starting from a reference patent document, search criteria can include classification, keywords, citations, time frame and country.

Different authors (see review [38]) mention that different search objectives require different approaches towards searching. A technology survey, on the one hand, will require to retrieve all relevant documents in many technology fields, and therefore requires a high recall rate as well as a high precision rate. However, for an invalidity search, precision is far more important than recall, as one very relevant document can be sufficient to invalidate a new application. It starts from a certain patent application or patent document and focuses on the claims, that need to be invalidated by finding prior art [14].

2.2. Patent search criteria

The main search criteria that are identified on almost all patent database websites, government or commercial, are the following two:

- Classification (IPC, CPC, ECLA…)
- Keywords

Other search criteria, depending on the objective of the search, could be the patent date (e.g. for infringement search) or the assignee name (e.g. to find competitors).

The traditional search methodology is an iterative process where the searcher first enters the main keywords with Boolean operators (and try searching with all alternative technical terms or possible synonyms of the keywords) to identify the relevant classifications. Then a classification search is made to find the most relevant patents, which will help to identify new keywords and so on. Also, an analysis of backward and forward citations is usually included.

In most cases the initial patent body is reduced to a relevant patent body by doing a rigorous analysis of the patent content. Sometimes mentioned is the use of patent clustering (by keyword strings). Also, the relationship between concepts, the size of the patent family, the length of the first claim, the number of forward citations, ordering the results by the 'best match'… are mentioned as good indicators for relevance.

The **main criteria to assess the success of a patent search** [38] are:

- Precision: The proportion of relevant documents in the set of retrieved documents,
- Recall: The proportion of retrieved documents that are relevant.

2.3. Step-strategies for Prior Art Search

2.3.1. USPTO 7-step patent search

As most proposed search strategies, the 7-stap approach educated by the US Patent&Trademark Office is based on 'classification' and 'keyword' searches:

1) Brainstorm keywords related to the purpose, use and composition of the invention;
2) Look up the words in the Index to the U.S. Patent Classification to find potential class/subclasses;
3) Verify the relevancy of the class/subclasses by using the Classification Schedule in the Manual of Classification;
4) Read the Classification Definitions to verify the scope of the subclasses and note "see also" references;
5) Search the Issued Patents and the Published Applications databases by 'Current US Classification' and access full-text patents and published applications;
6) Review the claims, specifications and drawings of documents retrieved for relevancy;
7) Check all references and note the "U.S. Cl." and "Field of Search" areas for additional class/subclasses to search

2.3.2. EPO 3-step patent search

The European Patent Office suggests a 3-step search for prior art.

1) Finding the right keywords: Spend some time thinking of key words or search terms which best describe your idea;
2) Product searching: You need to find out what is already on the market that is similar to your idea (prior art) and that tackles the same problem (competing art);
3) Patent searching: combine keyword and ECLA classification search.

In step 1, it is critical to choose the right keywords, considering that patent literature is very specific. Keywords cannot be too obvious or too simple. Often very specialized, technical terms or very new, recently invented terms are needed to find the relevant patents. Step 2 involves searching online, in scientific and commercial literature, talking to experts, … In fact, the 3rd step of this process actually comprises the patent search and is made up of substeps related to keyword and classification searches, namely:

Keyword search
- Using your list of keywords, prepare search strings of up to four keywords;
- Evaluate relevance of results list

ECLA search
- Check classification of most relevant found patent;
- Search patents in that classification;
- Refine result list by combining keywords;
- Examine result list;
- Repeat with new relevant patents.

2.3.3. WIPO patent search

WIPO does not indicate a search-step-algorithm but does indicate the main criteria for patent searching as the following:

- Keywords (+boolean operators);
- Inventor and Applicant names;
- Patent classification;
- Other criteria (reference numbers, filing date, country, family, …).

2.3.4. Search methodologies suggested by commercial firms in the patent field

Many commercial firms active in the field of patents and IP management indicate a basic form of their search strategies on their website. Two are mentioned below as an example. The main criteria of these searches always comprise a combination of keyword and classification searches and some form of patent indicators such as ‘citations’ to expand the search.

Example 1: https://www.acclaimip.com/

1) Start by keyword and search in CLASS of most relevant resulting patent (IPC , USPC, ECLA, F);
2) Query by example to find best relating patents to a strong patent, based on keywords or classification;
3) Cluster results to find most relevant keywords, concepts or classes;
4) Use OR and then order by relevance to all keywords;
5) Use patent data value indicators (e.g. forward citations, length of claim 1, size of patent family);
6) Iterate searches combining all techniques (zoom in zoom out);
7) Chart your results (by date, by name of inventor, …).

Example 2: http://www.intellogist.com/

1) Understand the search;
2) Full-text search to quantify the scope of the art;
3) Identify related patent documents to determine more specific terms related to art in the field;
4) Narrow the search body with the most relevant classes and subclasses from the appropriate classification area(s) of interest;
5) Search all relevant art within each chosen subclass;
6) Iterate (4) and (5) to identify additional references;

7) After exhausting (6), examine key central references for classes and subclasses not originally considered and repeat with respect to each new subclass;
8) Return to the full-text searching body and search the art for more recently identified keywords;
9) Search the remaining body of art using keywords found from central references, client notes, Examiner suggestions, etc.;
10) Perform a forward and backward citation search on each centrally relevant reference found during the search.

2.4. The PSA approach (Problem-Solution)

The PSA approach is used by EPO to assess the patentability of new inventions by doing a prior art search, and thus to decide upon the approval or denial of a patent application.

The approach involves analysing **which technical problem is solved** by an invention as well as **how it is solved**, and finding out **whether this solution is 'inventive' or not** by searching the relevant patent literature.

Of course, in this case the searcher already very practically knows the technical solution before starting the search for relevant patents. The method is not as useful to search for an 'unknown' technical solution to an existing technical problem, as what is needed in a 'technology search' or for innovation purposes.

Actually, the PSA approach is **not a patent search strategy, but a patent analysis strategy**. The body of patents in which the PSA search is made, is established by 'normal' classification and keyword searches that retrieve as many relevant patents as possible in the specific technological field that the patent application would fall into. If no results are found in this patent body, more patents are retrieved from other technological fields that are related to the invention.

3. Advanced search functionalities in patent databases

Many databases offer search functions that can be adapted to be very useful in terms of the FORMAT project. A brief summary is indicated below.

3.1. Public databases:

A short summary is given on the main search formats available in public databases (EPO, USPTO, WIPO). It is clear that while these databases offer 'structured searches' (search in specific patent fields, Boolean operators, wildcards and truncates) there are no means of using text mining tools such as parsing or other more advanced search mechanisms.

3.1.1. EspaceNet

The freely available **EspaceNet** patent search engine offers a 'simple' and a 'smart' search. The 'simple' search is mainly based on keywords. The 'smart' search offers a 'classification search and an 'advanced' search, where you can enter keywords for specific patent fields (title, abstract, inventor, reference numbers, date, classification). Booleans, wildcards and truncations are recognized.

EPO also offers subscription databases for specific purposes.

- The **World Patent Index** is specifically meant for Patent Information Experts and offers similar but more extensive search fields compared to Espacenet (numbers, dates, classifications, names, citations, keywords). With keywords, if offers indexes that allow you to specifically search either in the full text, or only in titles or abstracts, and in a specific language. However, there is no option available to search specifically in other patent parts such as 'claims'.

- **European Patent Bulletin** is mainly useful to follow up on the legal or procedural status of patents or patent families. The search fields are grouped into 4 categories: bibliography (~espacenet), examination (finding info on granting procedure), opposition (finding info about opposition procedures) and other (other legal or procedural information).

- **Espace EP** (A and B) is based on the Mimosa Retrieval software which allows to search EPO, US Patent and Japan Patent office databases in ten languages and can be used as a archiving system.

http://worldwide.espacenet.com/advancedSearch?locale=en_EP
http://www.epo.org/searching/subscription/gpi.html
http://www.epo.org/searching/subscription/bulletin.html
http://www.epo.org/searching/subscription/ep.html

3.1.2. PATENTSCOPE

The WIPO PATENTSCOPE search offers some additional structured search fields compared to EPO. Mainly interesting is the ability to search specifically not only within a title or an abstract, but also within 'English claims' or the 'description' part of patent documents. It allows to enter keywords in up to 13 structured search fields connected with Boolean operators.

http://patentscope.wipo.int/search/en/structuredSearch.jsf

3.1.3. USPTO

The US Patent and Trademark Office offers different online search tools.

- **PatFT** is the Patent Full Text and Image database and allows to perform a keyword search in full text, in certain patent fields (inventor, abstract, claims, classification, …), also in TIFF image patents.

- **AppFT** is similar to PatFT but also allows to search in patent application that have not yet been granted.

- **PAIR**, the Patent Application Information Retrieval system, offers a safe, simple, and secure way to retrieve and download information regarding patent application status.

http://www.uspto.gov/patents/process/search/index.jsp

3.1.4. Japanese Patent Office

The JPO offers two databases for patent searches through keywords search with Boolean operators, but the search is conducted in full text. It is not possible to indicate specific patent fields to search keywords in other than publication data, inventor or IPC class.

http://www4.ipdl.inpit.go.jp/Tokujitu/tjftermena.ipdl?N0000=114
http://www19.ipdl.inpit.go.jp/PA1/cgi-bin/PA1INIT?1359109277491

3.2. Other databases:

The WIPO website offers a link to more advanced free-to-use or subscription databases:

- LexisNexis: TotalPatent
- Minesoft: PatBase
- Questel: Orbit
- Thomson Reuters: Thomson Innovation
- WIPS: WIPS Global

While these, and other databases existing online such as Delphion, offer very advanced tools related to the retrieval, view and analysis of search results, the search itself is often still based on keyword and classification searches comparable to the public databases. However, **more advanced search mechanisms that can be very useful in the FORMAT methodology** are included such as:

- Semantic search (recognizing contextual meaning of keywords)
- Saved search queries and automatic updates of search results
- Customized search fields or intelligent search fields (‘key concepts’, ‘object of the invention’, ‘advantages over prior art drawbacks’, ‘examples’, …)
- Search for non-patent cited literature, relevant business information, etc.

4. Scientific Literature on Patent Searching

For some of the purposes of technology forecasting, the search criteria will be generally quite ‘broad’ and a combination of keyword and classification searches will result in a very large patent body to be analysed. As mentioned for the ‘technology search’ by EPO, indeed the goal should be to retrieve as many patent documents as possible (that are relevant to the specific technological problem that needs to be solved, that contain all emerging technologies relevant to the field, that are used to obtain historical data on technology evolution, …).

A preliminary questionnaire has revealed that for most companies, an important constraint for a technology forecasting tool is the time within which the forecast should be completed. In this case, a more structured patent search, involving semantics or a strategy towards a pre-selection of the most relevant keywords can help to reduce the analysis time needed.

The most useful papers that have been found related to these topics are summarized below.

4.1. Finding the relevant keywords

Li *et al.* [8] conducted a study on how to extract the significant-rare keywords for an efficient patent analysis, stating that **if the keywords are not chosen properly, results can bring up too many or too few patents**. Because patents are rarely written by the technological experts, but rather by their patent engineers or attorneys, different terminologies are used for similar technological features. They state some more reasons for varying terminology. Experts in different domains use dissimilar words, developing technologies don’t have a standard name yet and translation can cause terminology to differentiate even further. And in many cases, the strategy is to protect an invention without making it too easy to find, resulting in complex and fuzzy keywords. A standard keyword search can therefore still leave many things to chance. The authors have developed a snowball-rolling-process (based on TF and TF-IDF) that selects the significant-frequent and significant-rare terms in a patent body.

Jeon *et al.* [31] discuss how to use patent information in order to select the most relevant technology partners to start an innovation project. They state that patent information can be specifically important when the technologies required are distributed over various sources and become increasingly complex or when technology fusion is an important part of collaborative innovation. Main steps in their patent analysis process are the data collection and pre-processing. **The technological need should be clearly defined and then analysed by experts in all relevant technology fields, to reformulate the need in terms of similar technological problems in related fields.** This way, a structured set of keywords and synonyms can be used as a basis for the patent search. Once a patent body is identified, they are **analysed in terms of co-occurrence of keywords that represent the problem-solution relationship** of the technological need. The most relevant patent documents can be identified through the value of a cosine similarity indicator, which identifies the similarity of co-occurrence vectors between the technological need and the patent documents.

OuYang and Weng [39] describe the New Comprehensive Patent Analysis approach to analyse a body of patents and identify the key patent and patent families relevant for new product development. In the approach, there are 5 main phases, of which the first consists of identifying the appropriate patent database by conducting a patent search. The patent search strategy is based on **information about market trends, technological evolution, and main competitors in the field, as well as on expert interviews (management and R&D).** These sources result in a selection of technological keywords and confirmation of relevant classes (IPC).

In summary, due to the complexity of technological terminology use in patents, most authors indicate that the relevant keywords for any search query are defined by relying on the knowledge of experts in the relevant technological fields, as well as on other sources of technological information. In most cases, an initial search is made with a limited set of keywords. More relevant keywords can be found in the first resulting patents and can be used to expand or narrow the search query.

4.2. Searching in the most relevant parts of a patent

Xie & Miyazaki [13] study the effectiveness of the keyword search strategy to find patent literature related to cross-disciplinary or integrated technologies or products that are emerging and cannot be clearly defined through patent classes. By a case study of 'automotive software' in the USPTO patent database, they analysed the effectiveness of using keyword search by searching different parts of the patent documents: (i) title and abstract, (ii) title, abstract and claims and (iii) description. The effectiveness was evaluated by assessing two 'errors': the number of times of retrieving an irrelevant patent and not retrieving a relevant patent. Their results indicate that **keyword searches should be conducted specifically in patent claims to identify and extract more information in patents**, while the description is too 'noisy' and not as useful to identify relevant patents. They also highlight the importance of a precise selection of the right keywords.

Tseng and Wu [33] studied the main search tactics of patent engineers. One important conclusion is that, before starting a search, the main keywords are identified by scanning existing information and already known patents, as well as by **discussing with colleagues, seniors, technicians etc. that have knowledge of the technology concerned**.

Tiwana and Horowitz [34] propose a **text mining supported search** to identify the 'problem solved concept' in a patent. Their main conclusion is that the 'problem to be solved' is usually found in the first sentences describing the **'background of the invention'**, which is usually the first paragraphs in the 'description'. How the invention solves this problem can be found in the **'summary'** of the invention. They propose a text mining methodology that can extract these relevant parts of the patent and prove that a **keyword search into 'finding the technological problem and how the invention solves it' can be improved** in terms of recall and precision if the search can be conducted specifically in these most relevant parts.

4.3. Identifying experts in the field by patent analysis

The paper of Jeon *et al.* [31], discussed in 4.1, specifically analyses patents by co-occurrence vectors of relevant keywords to be able to **identify the most suitable partner for collaborative innovation**, which are the applicants of the most relevant patent documents resulting from the analysis.

Ernst [30] similarly **uses patent information to identify the main players in a technological field**, as well as the **key inventors** employed by those companies. The analysis is based on indicators such as the patent activity (number of patents of a player in a field), the technology share, the cooperation intensity, the citation frequency, patent strength and so on. They conclude that patents are a valuable source of information to identify and assess external sources for knowledge generation.

4.4. Semantics in Patent Searching

Semantic solutions work on word meaning rather than on word occurrence of frequency counts [38], which can allow providing more advanced search and filtering functionalities in patent seaching.

Semantic solutions are based on knowledge bases [38] such as:
- Taxonomy = a hierarchy of relevant entities (parent-child, part-of, instance-of, …)
- Thesuarus = a set of related terms (hierarchical, equivalence, associative, …)
- Ontology = set of representational primitives such as classes, attributes, relationships

It should be remarked that, compared to the use of semantics in patent searching, the literature on the use of semantics in patent analysis, i.e. the step after patent search and collection, is far more abundant. In most of these cases, the patent search is conducted in the more traditional way, by finding the most relevant keywords or classes through expert knowledge and applying an iterative search.

4.4.1. Relevance of semantics in Patent Searching

Azzopardi *et al.* [9] conducted a survey on the search behaviour of patent users and their requirements for a search system. Their main conclusion is that patent searchers in general have a strong **preference towards search functionalities that provide fine grained control** over the search process, using operators with a **clear semantic and interpretation**.

Also, Tseng and Wu [33] performed a survey of main search tactics of patent engineers (see also 4.2) and provide some recommendations for a patent search system, such as the **automatic suggestion of other, relevant search vocabulary (context-sensitive)** and an automatic correction of errors in keywords or search commands.

As Ryley *et al.* [37] remark, a **fundamental drawback of keyword searching** (with strings of keywords and Boolean operators) to retrieve patents, is the inability of the process to identify relevant documents that, for a variety of reasons, simply do not contain those keywords. They state that there is little to be done about this, other than employing subject matter experts who attempt to create queries with every possible synonym that could describe an invention.

Many other authors make the same assessment. The increased complexity of patent research has resulted in a field of research related to ‘semantics’ and ‘information retrieval’ specifically applied to patent searching, aiming to develop more advanced search mechanisms that improve the recall rate as well as the relevance of retrieved documents.

4.4.2. Application of semantics in Patent Searching

4.4.2.1. Enhancing search functionalities

Koch *et al.* [35] recognize that the complexity of patent information inevitably results in an iterative search strategy, expanding or refining an initial query based on the first results. They have built a patent search and analysis system, PatViz. In the patent search phase, this system allows to integrate different search facilities in one query, including full text (keyword) search, metadata search (e.g. classification), image similarity search, semantic search and document similarity search. In the semantic search, **new context-sensitive keywords are suggested in a dropbox menu, based on previous search terms and the content of search results**.

4.4.2.2. Document retrieval mechanisms

Ryley *et al.* introduce **Latent Semantic Indexing** as a mechanism to improve patent retrieval. This is an advanced document retrieval technique based on combining the vector space model (VSM) and singular value decomposition (SVD). The VSM turns a document into a vector, defined by the terms in the document and the frequency with which they appear. The relevancy between a query and the document can be determined by the distance or the angle between the query and document vectors. VSM relies upon literal word matching, which is a weakness. SVD can enhance VSM by **automatically ascertaining word relationships such as synonyms or other correlations**.

Chen *et al.* [36] propose a method that combines a semantic, content-based search and a syntactic, form- or word-based search. The semantic search, based on **Differential Latent Sematic Indexing**, reduces the patent database to a content-relevant search space and the syntactic search pins down the most relevant documents in this space by a syntactic search on words. Differential LSI adapts LSI by differentiating between external vectors (between documents) and internal vectors (between two parts of the same document).

Takaki *et al.* [41] apply **'associative document retrieval'** to patent searching, which is based on searching for similarity with a reference document. In the reference document, all relevant subtopics are identified by compositional element extraction. For each subtopic, a query is made to retrieve similar documents, based on a set of terms extracted from the subtopic text. The results of all queries are then ranked according to the weighted relevance of the subtopics to the reference document.

4.4.3. Existing semantic solutions for patent searching

Many tools exist that make use of some form of semantics for patent searching or analysis.

Interesting existing tools, listed by Bonino *et al.* [38], that make use of semantics in the field of patent searching are:

Patent Café
The aim of using semantics is to improve the recall in patent search. It builds a semantic index of the patent database by Latent Semantic Analysis (based on the statistical Principle Component Analysis), allowing to directly exploit semantic correlations between the available patents [38, http://www.pantrosip.com/].

PATexpert
PATexpert supports interactive (feedback-oriented) search. Both patent documents and search queries are described in terms of ontology concepts. This can support search queries such as: 'find documents similar to document X as far as feature Y is concerned'. It offers a number of search engines based on different search criteria: full-text retrieval (keywords + Boolean operators), **semantic retrieval (search criteria such as material of object, functionality of component, …)** similarity search (find semantically similar documents) and image retrieval (association of figures with text). [35, 45]

IntelliPatent
Semantic indexing is used to cluster and filter search results, allowing to refine a first, coarse search in an iterative fashion. The semantic indexer is domain-specific and the user can adapt the ontologies and classification criteria [38, http://www.intellipatent.eu/].

Invention Machine
Amongst other functionalities, a semantic search engine can capture relationships at the basis of analysis tasks such as FMEA [38, http://invention-machine.com/goldfireinnovator.htm].

IPInQuest
http://www.ipcentury.de/

SemIP InfoPatent
infoPatent deploys semantic processing algorithms based on so-called "invention clouds", i.e. term vectors containing a description of the essence of a patent. These vectors are generated by identifying those parts in the full-texts or abstracts that most clearly describe the invention from a technical point-of-view. Each patent is described by one or more invention clouds. To conduct a similarity search, invention clouds can be built for the textual search query and compared to those of the stored patents.
http://www.infoapps.de/en/patent-searching-tool-sem-ipcom/

Other software providers, using text mining and semantics in the field of patent searching, are listed by Lee *et al.* [43], as shown in the table below, taken from their paper:

Table 2
Features of the existing systems for patent analysis.

System	Main functions				Main techniques	Main applications
	S	A	M	I		
Aureka	●	●	●		Web-based, simple statistics, data mining, text mining, citation analysis	IP strategy
IPMap	●				Web-based, simple statistics	IP strategy
PatentMatrix®		●	●		claim analysis	IP strategy
Spore® Search	●		●		Semantic-based statistical search	IP search and mapping
Invengine™	●				Web-based, semantic analysis	IP search
M-Cam Door	●				Linguistic genomic algorithm	IP risk management
CIA™ Database	●				Web-based, simple statistics, data mining, citation analysis	R&D and business strategy

Note: S: search (collection), A: analysis, M: mapping (visualization), I: interpretation

5. Patent Analysis related to technology analysis and forecasting

In general, patents contain both structured data (title, inventor, classification, year…) and unstructured data (description, claims…). The structured data is relatively easy to analyse by use of bibliometrics. The unstructured data can be analysed by a combination of more advanced techniques based on text mining.

The study into the analysis of unstructured data in patents has led to the development of various useful patent analysis techniques.

5.1. Selecting Key Patents

Jeon *et al.* [31] studied the use of text mining for analysing patents aiming at selecting the key players in a certain technology. Relevant keywords were selected by technology experts and transformed into action-object **co-occurrence vectors**, representing e.g. the problem-solution relationship. For each patent in the body as well as for the search query (the technological need for the search, what do we need to find?), the frequency of co-occurrences at the sentence-level was analysed. A cosine similarity index was then set-up to define the similarity in co-occurrences between the search query and each patent in the body. **The patents with the highest similarity index are the key patents related to the technological need of the searcher**.

OuYang and Weng [39] apply a New Comprehensive Patent Analysis approach as a support for New product Design in the field of Mechanical Engineering. In the five-phase approach, the first two phases consist of finding and collecting a relevant patent body. In the third phase, the key patents are selected by **computing and statistically analysing the 'key patent priority number'**, based on the technological value added (functionality and variability), the application potential (based on 9 criteria) and any other innovativeness (energy efficiency, security, environment). The resulting key patents are further analysed by tracking their citations, selecting the patents that are related to key technologies for new product design and analysing the technological performance of these patents by TRIZ.

Jeong and Kim [52] developed **methods to extract and to calculate the problem and the solution similarities between patents**. They state that the background section is most relevant for identifying the technological problem, as it contains the field of the invention, state of the art, the problems of exist system and the objectives of the invention. The claims section was utilized to extract the solutions, as it contains the principle, the component, the process, the material, and the detailed conditions of the invention. They then process these parts of the patents into vector pairs to be able to calculate frequencies and similarities. By assessing the similarities between problems and between solutions of patents, they identified 4 patent clusters (similar in neither, similar in both or similar in either of the two).

5.2. Patent mapping

Kim *et al.* [46] describe the use of patent mapping for selecting emerging technologies in a technological field. After finding a relevant patent body using a selection of keywords, a 'keyword existence matrix' is formed and patents are clustered by the appearance of keywords in the patent text using the k-Means algorithm. **Semantic networks of keywords** are then formed according to which clusters contain which combination of keywords. A patent map can then be built, combining the keyword networks and the bibliographic information of the patents on the nodes of each network. This map can be used to track the advance of an emerging technology or gain insight into the evolution of technologies in a field.

Lee *et al.* [43] propose an approach for creating and utilizing keyword-based patent maps for use in new technology creation activities. The approach consists of three steps: (i) identifying keyword vectors in patents by text mining, (ii) reducing the number of keywords by principal component analysis and creating a 2-dimensional map and (iii) **identifying and validating the technological vacancies on the map**. By criticality and trend analysis, vacancies are defined as emerging or declining technology fields and as important or meaningless.

A similar patent mapping approach is proposed by Son *et al.* [50] based on generative topological mapping. It follows the steps of collecting a relevant patent body, extracting keywords by text mining and expert opinions, and preprocessing the patent data by creating keyword vectors. Then, a patent map is created from the keyword vectors by a GTM algorithm. Patent vacuums are automatically and objectively identified as gaps in the map. By inversed mapping, the vacuums are transformed to the original keyword vectors.

Choi and Park [47] recognize that it is difficult to identify and understand the detailed structure of technology development at the level of individual patents. They propose an approach to trace a 'patent development path'. They develop a **'patent citation matrix'**, wherein columns and rows are individual patents and the numbers in the matrix define the citations among patents. Other relevant numbers include the number of forward and backward citations for each patent and the 'forward citation node pair (FNCP) as the product of the number of forward citations of a patent pair. Using the citation matrix, they **trace a development path forward from an origin patent (not citing earlier patents) to a terminus patent (not (yet) being cited)**.

5.3. Subject-Action-Object parsing in patent analysis

Park *et al.* [51] propose a new approach to identification of promising patents for technology transfer by adopting TRIZ evolution trends as criteria to evaluate technologies in patents. The **Subject–Action–Object (SAO)-based text-mining technique** is used to deal with big patent data and analyse them automatically. SAO structures can reflect the technological key concept in a patent document and can be used to directly compare patents with the rule base of 'reasons for jumps' (RFJ) of TRIZ trends that are expressed in Action–Object (AO) structure.

Similar to Park *et al.* [51], Yoon and Kim [53] propose a method of detecting new technological opportunities. They analyse **only the claims** section of patents by using SAO-based semantics (making use of parsing tools) and measure semantic similarities between patents. They then build a **patent map or 'dissimilarity matrix'** and **detect outliers** to identify unusual or distinctive patents in a given technology area, that can be reviewed by experts.

A nice overview of other papers employing the SOA-semantic technique is given at p. 651 by Gerken and Moehrle [54]. They propose a 4-step **process to identify novelties in patents**: (i) identifying and extracting semantic structures, (ii) defining their relevance by a domain- or situation-related linguistic analysis, (iii) measuring similarity and (iv) calculating the novelty of an invention from the similarity matrices.

Further relevant experiences worth to be mentioned refer to the Pat-Analyzer project [55, 56], described by the developers of the PatExpert project as the "most advanced language technology technique available to date for patent processing". The main outcomes are the development of new techniques and algorithms for patent analysis and comparison with the following capabilities through the analysis of patent text:

- identify the components of the invention;
- classify the identified components in terms of detail/abstraction level and their compositional relationships in terms of supersystem/subsystem links;
- identify positional and functional interactions between the components both internal and external to the system;
- identify the most relevant components of each patent for a given project according to a ranking criterion, which combines the detail level of the description with components' occurrences in patent claims and with the Inverse Document Frequency, i.e. the "rarity" of each synset of the Thesaurus.

5.4. TRIZ classification of patents

Tong *et al.* [48] and Cong and Tong [49] conducted studies into how to classify a patent body according to TRIZ contradictions, so to facilitate patent searching and patent analysis for TRIZ users. They process patent bodies by document indexing (term frequency), word stemming and feature selection. They concluded in a first study that the performance of their classification system was dependant on the specific TRIZ class and that the accuracy generally decreased after adding the 'summary' of the patent to the analysis. In a second study, they conclude that 33 Inventive Principles are 'distinct' that can be easily defined by text information, while 7 are 'obscure', as there is little typical text information available and will require more advanced textual classification methods.

6. Recommendations for the FORMAT project

For the methodology to be developed in FORMAT project, in many cases the goal of a patent search will be to **find a 'solution for a technological problem'**. In such a case, it is possible that analogous solutions could be found in a completely different technology field. The FORMAT methodology should be able to handle these types of searches swiftly and efficiently.

In the field of patent search and patent analysis, the existing semantic-based search tools would be a good starting point for FORMAT. Especially the **Subject-Action-Object approach** could be very useful. It would be interesting to know the limits of such an approach and whether the search criteria could be extended to handle specific ‘properties’ or ‘qualities’ of the subject, object or action. The FORMAT methodology should also be able to recognize emerging technologies and future directions of R&D in a certain technology field, to support a company’s strategic decisions. For this purpose, the **“patent mapping” approaches**, if accordingly adapted, could be able to provide very easy-to-read, visual overviews.

Two things that need to be taken into account, regardless of the purpose of the search, are the following. First of all, it is highly important to find the most relevant keywords. The relevance of keywords sometimes depends on the object of the search (e.g. describing the technological problem). Whether you start from a set of documents or you start from an idea, the importance of **involving experts to determine the right set of keywords** is stressed in almost every paper in the field of patent searching. Second of all, depending on the search query, different **sections of the patents** should be analysed (e.g. claims describe the technological solution).

7. INNEN text mining resources and competencies

The general goal of text mining techniques is to turn text into data for analysis, via application of **Natural Language Processing** (NLP) and analytical methods.

A typical application is to scan a set of documents written in a natural language and either model the document set for predictive classification purposes or populate a search index with the extracted information and with the additional enrichment data grabbed from the cloud.
After the initial classification, done with a large enough document corpus, the system is able to classify automatically new documents and to assign them appropriate tags or calculate a set of documents, semantically related to the one inserted.

Another possibility is to discover relations between the documents and a set of patents or technical papers, coming from a publicly available repository, in order to suggest potential innovations for the involved companies in their products or business processes.

One of the engines that will be included in the text mining module, developed by INNEN, will provide an automatic multithreaded retrieval engine for European patents available from the European Patent Office (EPO) and a similar engine for freely available technical papers.

Specific analytical tasks can discover in such sets of information applicable tools to perform the same action in a different context, possibly suggesting different and more efficient ways of doing the same things, new materials that can innovate their products or potential inefficiencies or harmful operations in the current production process.
The Innovation Engineering (INNEN) approach to perform these searches sprouts from the above mentioned Subject-Action-Object parsing techniques, here referred as the OAT (Object Action Tool) paradigm

Some of the features of the INNEN text mining technologies are:

- automatic indexing and text extraction from virtually any type of binary document (MS Office suite document formats, Open Office document formats, PDF, images, etc.),
- storage and retrieval of the contents from any CMIS compliant document repository,
- automatic classification and tagging of the documents

- advanced searches, with search terms highlighting, text snippets extraction, facet based categorization and clustering of the results (facets can be dynamic, like data ranges);

The solution is in active development and is in continuous evolution; other semantic features will be added in the near future, to support:

- similarity and fuzzy searches
- POS (Part of Speech) tagging inside each sentence of the text extracted from the documents
- automatic identification of the core "concepts" the document speaks about

The solution is based on a customization of the enterprise search server Apache Solr, with multi-core and multi-domain capabilities, and on Apache UIMA, a project originally developed by IBM, but subsequently donated to the Apache community for the analysis of unstructured content such as text, audio and video.

INNEN is developing its technical resources in the field of semantic search engines and analysis and will also benefit from the results of two important research projects:

- the **INSEARCH semantic search engine** (ended in December and currently in review process)
- the **MARKET-IT and PROGRESS-IT projects** (in progress)

The INSEARCH project is an information system, based on advanced search, retrieval and analysis techniques, able to support European SMEs in the technology innovation process and it has been developed by Innovation Engineering in collaboration with the Universities of Rome Tor Vergata and Kufstein (Austria). Its main scope is to collect and organize relevant knowledge for innovating internal products and processes and supporting human decisions and actions. It includes sophisticated tools for content filtering and ranking of the collected results and a semantic data management system able to send proactive recommendations to the users, through an intelligent monitoring system of external resources and Web sites, relevant for the specific user business case.
It is equipped with a powerful and reusable Web spider, able to extract, index and store data from Web pages, European patent databases and technical papers repositories.

The MARKET-IT and PROGRESS-IT projects will contribute the lexical and NLP (Natural Language Processing) engines for Italian and English languages, as well as the semantic similarity engine for finding relations between words and documents (based on a complex and innovative vector space distance calculation between words and concepts in a specific knowledge domain).

8. Bibliography

8.1. Websites:

http://www.epo.org
http://www.uspto.gov
http://www.wipo.int/
http://www.jpo.go.jp/
http://www.pantrosip.com/
Apache Solr open source enterprise search server, http://lucene.apache.org/solr/
Apache UIMA, engine for analysis of unstructured data, http://uima.apache.org/
The Linking Open Data cloud diagram, http://richard.cyganiak.de/2007/10/lod/
The Semantic Web, Linked and Open Data,
http://wiki.cetis.ac.uk/images/1/1a/The_Semantic_Web.pdf

8.2. Scientific literature:

[1] Järvenpää, H., In Technology Forecasting Using Bibliometrics. What Information Source is Relevant When? Exploring Different Source Types - PICMET 2009 Proceedings
[2] Choi, C. and Park, Y., Monitoring the organic structure of technology based on the patent development paths, Technological Forecasting & Social Change 76 (2009) 754-768
[3] Yoon, B. and Lee, S., Applicability of Patent Information in Technological Forecasting: A sector-specific Approach, Journal of Intellectual Property Rights 17 (2012) 37-45
[4] Nijhof, E., Subject analysis and search strategies – Has the searcher become the bottleneck in the search process?, World Patent Information 29 (2007) 20-25
[5] Damm, A., Technology and competitor mapping designed to support strategic business Decisions, World Patent Information 34 (2012) 124-127
[6] Porter, A. *et al.*, Mining PICMET: 1997-2003 Papers Help You Track Management of Technology Developments
[7] Pilkington, A., Technology Commercialisation: Patent Portfolio Alignment and the Fuel Cell
[8] Li, Y. *et al.*, Extracting the significant-rare keywords for patent analysis, Expert Systems with Applications 36 (2009) 5200-5204
[9] Azzopardi, J. *et al.*, A Survey on Patent Users Search Behavior, Search Functionality and System Requirements, IRF Report 2010-00001
[10] Daim, T. *et al.*, Forecasting emerging technologies: Use of bibliometrics and patent analysis, Technological Forecasting & Social Change 73 (2006) 981–1012
[11] Bonino, D. *et al.,* Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics
[12] Porter, A.L. and Detampel, M.J., Technology Opportunity Analysis, Technological Forecasting and Social Change 49 (1995) 237-255
[13] Xie, Z. and Miyazaki, K., Evaluating the effectiveness of keyword search strategy for patent identification, World Patent Information xxx (2012) 1-11
[14] Fujita, S., Technology survey and invalidity search: A comparative study of different tasks for Japanese patent document retrieval, Information Processing and Managemen 43 (2007) 1154-1172
[15] Greve, H.R., Exploration and exploitation in product innovation, Industrial and Corporate Change 16(5) (2007) 945-975
[16] Laursen, K., Keep searching and you'll find: what we know about variety creation through firms' search activities for innovation, Industrial and Corporate Change 21(5) (2012) 1181-1220
[17] Makri, M. and Lane, P.J., A search theoretic model of productivity, science and innovation, R&D Management 37 (2007) 303-317
[18] Gupta, V.K., Multi-disciplinary Studies on IPR in R&D: A Review, Journal of Intellectual Property Rights 9 (2004) 33-42
[19] Phadnis, R. and Hirwani, R.R., Patent Analysis and a Tool for Research Planning: Case Study of Phytochemicals in Tea, Journal of Intellectual Property Rights 10 (2005) 221-223
[20] Praveen Raj, R.S., Inventive Step or Non-Obviousness of an Invention, Journal of Intellectual Property Rights 10 (2005) 232-234
[21] Kardam, K.S., Patenting in the Emerging Fields of Technology, Journal of Intellectual Property Rights 12 (2007) 15-29
[22] Baruah, A., Patent Specification: Engineering the Technical Output of Novel Invention, Journal of Intellectual Property Rights 14 (2009) 423-431
[23] Ghafele, R. and Gibert, B., Driving Innovation Through Patent Application Review: The Power of Crowdsourcing Prior Art Search, Journal of Intellectual Property Rights 16 (2011) 303-308
[24] Yoon, B., Applicability of Patent Information in Technological Forecasting: A Sector-Specific Approach, Journal of Intellectual Property Rights 17 (2012) 37-45
[25] Jun, S., Patent Management for Technology Forecasting: A Case Study of the Bio-Industry, Journal of Intellectual Property Rights 17 (2012) 539-546
[26] Magdy, W. and Jones, G.J.F., A Study on Query Expansion Methods for Patent Retrieval, Proceedings of PaIR 2011
[27] Parapatics, P. and Dittenbach, M., Patent Claim Decomposition for Improved Information Extraction, Proceedings of PaIR 2009
[28] Atkinson, K.H., Towards a More Rational Patent Search Paradigm, Proceedings of PaIR 2008
[29] Tiwana, S. and Horowitz, E., FindCite – Automatically Finding Prior Art Patents, Proceedings of PaIR 2009
[30] Ernst, H., Patent information for strategic technology management, World Patent Information 25 (2003) 233–242

[31] Jeon, J. *et al.*, How to use patent information to search potential technology partners in open innovation, Journal of Intellectual Property Rights 16 (2011) 385-393

[32] Klampanos, I.A., Azzam, H. and Roelleke, T., A case for probabilistic logic for scalable patent retrieval, Proceedings of PaIR 2009

[33] Tseng, Y.-H., Wu, Y.-J., A study of search tactics for patentability search – A case study on patent engineers, Proceedings of PaIR 2008

[34] Tiwana, S. and Horowitz, Z., Extracting problem solved concepts from patent documents, Proceedings of PaIR 2009

[35] Koch, S. *et al.*, Iterative Integration of Visual Insights during Patent Search and Analysis, IEEE Symposium on Visual Analytics Science and Technology, 2009, 203-210

[36] Chen, L., Tokuda, N., Hisahiro, A., A Patent Document Retrieval System Addressing Both Semantic and Syntactic Properties

[37] Ryley, J.F., Saffer, J. and Gibbs, A., Advanced document retrieval techniques for patent research, World Patent Information 30 (2008) 238-243

[38] Bonino, D. *et al.,* Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics, World Patent Information 32 (2010) 30-38

[39] OuYang, K. and Weng, C.S., A New Comprehensive Patent Analysis Approach for New Product Design in Mechanical Engineering, Technological Forecasting & Social Change 78 (2011) 1183-1199

[40] De Vries *et al.*, Search by Strategy, ESAIR 2010, 27-28

[41] Takaki *et al.*, Associative Document Retrieval by Query Subtopic Analysis and its Applications to Invalidity Patent Search, CIKM 2004, 399-405

[42] Moehrle *et al.,* Patinformatic as a business process: A guideling through patent research tasks and tools, World Patent Information 32 (2010) 291-299

[43] Lee *et al.*, An approach to discovering new technoliogy opportunities: Keyword-based patent map approach, Technovation 29 (2009) 481-497

[44] Trappey *et al.*, Development of a patent document classification and search platform using a back-propagation network, Expert Systems with Applications 31 (2006) 755-765

[45] Wanner *et al.*, Towards content-oriented patent document processing, World Patent Information 30 (2008) 21-33

[46] Kim *et al.,* Visualization of patent analysis for emerging technology, Expert Systems with Applications 34 (2008) 1804-1812

[47] Choi, C and Park, Y., Monitoring the organic structure of technology based on the patentdevelopment paths, Technological Forecasting and Social Change 76 (2009) 754-768

[48] Tong, L.H. *et al.*, Automatic classification of patent documents for TRIZ users, World Patent Information 28 (2006) 6-13

[49] Cong, H. and Tong, L.H., Grouping of TRIZ Inventive Principles to facilitate automatic patent classification, Expert Systems with Applications 34 (2008) 788-795

[50] Son *et al.*, Development of a GTM-based patent map for identifying patent vacuums, Expert Systems with Applications, 39 (2012) 2489-2500

[51] Park *et al.*, Identification of Promising patents for technology transfer using TRIZ evolution trends, Expert Systems with Applications 40 (2013) 736–743

[52] Jeong, C. and Kim, K., Technology Relationship Analysis Using Problem and Solution Similarities, Proceedings of the 2012 IEEE ICMIT, 516-521

[53] Yoon, J.and Kim, K., Detecting signals of new technological opportunities using semantic patent analysis and outlier detection, Scientometrics 90 (2012) 445–461

[54] Gerken, J.M. and Moehrle, M.G., A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis, Scientometrics 91 (2012) 645-670

[55] Cascini G., Fantechi A., Spinicci E., Natural Language Processing of Patents and Technical Documentation, Lecture Notes in Computer Science, Springer-Verlag Heidelberg, ISBN: 3-540-23060-2, DOI: 10.1007/b100557, December 2004, pp. 508-519.

[56] Cascini G., Russo D., Computer-Aided analysis of patents and search for TRIZ contradictions, International Journal of Product Development, Special Issue: Creativity and Innovation Employing TRIZ, Vol. 4, Nos. 1/2, 2007, pp. 52-67.