Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 131 (2015) 1130 – 1139

Procedia Engineering

www.elsevier.com/locate/procedia

World Conference: TRIZ FUTURE, TF 2011-2014

Identifying and reformulating knowledge items to fit with the Inventive Design Method (IDM) model for a semantically-based patent mining

Achille Souili[a*], Denis Cavallucci[a], François Rousselot[b]

[a] LGeCo / INSA Strasbourg, 24 Boulevard de la Victoire, 67084 Strasbourg Cedex, France
[b]rousselotfr@gmail.com

Abstract

The analysis of initial situation is necessary when dealing with the complex engineering situation. Furthermore, innovation is the underlying foundation of technological advances and today's competitive economy. Patents contain important research and it is also possible to find within them a kind of history of the evolution of an artifact. Used effectively, patents can help to provide businesses and individuals work with innovative ideas. In this context, the engineers may very often need to analyze them in order to benefit from the knowledge contained therein to organize their inventive task. However, patents are lengthy and rich in technical terminology; analyzing patents may require a lot of human effort. Thus, automating the process is very timely. To facilitate this, several patent analysis tools have been created. There are also patent analysis tools dedicated to highlighting various values of tools, but very few of them are designed to extract information or knowledge contained in patents. This paper presents an ongoing research on knowledge extraction for the Inventive Design Method (IDM), which extends from TRIZ, the theory of inventive problem solving.

Peer-review under responsibility of the Scientific Committee of TFC 2011, TFC 2012, TFC 2013 and TFC 2014 – GIC

Keywords: Knowledge discovery; Patent mining; NLP; Text-mining, IDM; TRIZ

[*] Corresponding author. Tel.: +33.(0)3.88.14.47.55 .
*E-mail address:*achille.souili@insa-strasbourg.fr

Peer-review under responsibility of the Scientific Committee of TFC 2011, TFC 2012, TFC 2013 and TFC 2014 – GIC
doi:10.1016/j.proeng.2015.12.432

1. Introduction

The arrival of the Internet has made a huge amount of data available, particularly patent documents. According to the World's Intellectual and Property Organization, patents contain more than 80 percent of the world technical knowledge. Additionally, patent databases are filled with thousands of new contributions each year.

Nomenclature

NP	Noun Phrase
VP	Verb Phrase
Adv	Adverbial Phrase

Patents turn out to be a good contribution to research. They are valuable to designers, industries and businesses. In patents, it is possible to find a history of the evolution of an artifact. When carefully analyzed, patents can reveal important details such as past or potential evolution of the artifact. This information can later be used to make a decision about the method or tool being developed. In this context, the designer may very often need to do research in order to benefit from the knowledge contained therein to organize the inventive process. Unfortunately, the knowledge contained in patents remains underused by engineers who do not take the time to analyze them and so benefit from their content. The Inventive Design Method (see section 3) derives from TRIZ [1] the theory of inventive problem solving and was developed to assist designers in their innovation process. Unlike other traditional methods for solving inventive problems, IDM can overcome psychological inertia, an essential step in the innovation process. Its main goal is to bring the designer to cross the borders of his/her field of study to find solutions by analogy, by looking for generic knowledge contained in the patent documents. However, experts currently do this research manually. Creating and updating the graph of problems that is actually a representation of all known problems and partial solutions network is time consuming and requires a lot of effort. It becomes thereby necessary to make patent more accessible to engineers.

In this research, we seek to support effective and efficient patent mining and knowledge extraction by the use of text mining techniques. From visual and bibliometric analysis to semantic searches, many approaches exist in patent analysis. However, these approaches have not reached maturity yet. Despite their obvious efficiency, they are still domain dependent and do not always address unstructured data. The output is often a re-treatment of surface data, such as summaries, charts, maps and tables. Furthermore, great amounts of the work that aim at extracting knowledge from patent just perform quantitative techniques and try to display trends. Therefore, providing a tool that can effectively meet the designers' needs is in great demand.

This paper presents an ongoing method to automate IDM knowledge extraction. We firstly report on ongoing progress made in the field of patent analysis, after introducing TRIZ.

Second, we present the Inventive design method and its ontology. Finally a case study and results are presented and discussed to measure the performance of the model of extraction.

2. An overview of patent mining tools

Knowledge extraction is young and in recent years many approaches have been proposed to extract knowledge from patent documents and, with this extracted knowledge, automate the TRIZ process.

A huge amount of patent documents is now available on the internet. This situation has caused the development of sophisticated patent mining tools. Many various tools were created for analyzing patents. These tools can be used to perform a wide range of tasks including conducting strategic technology planning, detecting patent infringement, analyzing and forecasting future technological trends, determining the most promising patents and patents quality, and identifying patent vacuums and technological hotspots [2]. However, these tools are either limited in their domain of application or do not tackle the contradiction concept well. The patent mining field is very young and this section reports on major progress in the field of patent mining and more specifically in patent mining for TRIZ and IDM. Tools dedicated to patent mining are very few. Some are related to TRIZ and have the intent to automate knowledge extraction from patent documents. Such tools are typically using hybrid approaches by associating statistics and linguistics. Feldman et al [3] for example, present a document explorer which implements text- mining at the term level. A list of candidate words is produced to be keywords after a basic linguistic preprocessing. Furthermore, Ghoula et al. [4] expose a processing chain achieving an automatic semantic patent annotation through a structural and domain ontology. Such approaches are promising for patent document unstructured parts processing.

However, the above mentioned approaches are limited in the way that they do not take into account artifact improvements. Hence, they do not give access to the invention process. The efficiency of TRIZ is obvious and with its development many authors have tried to automate its application in order to use the inventive principles to solve problems in a variety of domains. One of these widespread approaches is S-A-O (Subject-Action-Object) which is intrinsically linked to the concept of function, the understanding of which differs from one author to another.

For example, Savranski [5] defines function as the "action of changing the feature of any product" while for Cascini et al., action and subject may refer to the components of a system where the action refers to the functions performed by and on components. More precisely, they propose a method to automatically identify the contradictions subjacent to a given technical situation or system using patent mining [6]. According to them, functional analysis can be used to identify problems or generate innovative solutions. In functional analysis, a problem is broken down into its component functions that are further divided into sub-functions and sub- sub functions, until the function level of solving the problem is reached. Functional analysis is relevant for the representation of knowledge related to patents key finding and the inventor's domain of expertise.

3. The target description

This paper aims at describing an ongoing research to conceive a patent mining tool relevant to automatically extract IDM related knowledge since engineers manually do this so far. This research explores the use of text mining technics to improve knowledge extraction. More specifically, it explores the use of generic linguistic markers to match and retrieve concepts likely to be of interest to IDM.

3.1. TRIZ and IDM knowledge model

IDM, as its mother theory, TRIZ, is based on logic, data and research rather than intuition. It is primarily about technical and physical problems. TRIZ [1; 7] was developed and enunciated in 1946 by the Russian engineer and scientist Genrich Altshuller, when he first discovered that objective laws govern the development of technical artifacts. Stated after a study of nearly two million patents, the theory was assuming the existence of universal principles of invention. This study led him to make the first three following findings [8]

- First, some problems and solutions reappear frequently in industry as well as in science. A predictive solution to these problems can be found by categorizing the contradictions existing in each problem. Contradictions are the cornerstones of TRIZ. Contradiction may be defined as a conflict existing between characteristics within a system. More precisely, it occurs when the improvement of one parameter or characteristic of a system or process negatively affects the same or other parameters of the system
- Second, Patterns of technical evolutions may also be reappear frequently across many different industries.
- Third, in general, creative innovation representing these technical evolutions emerges outside the field where they were developed. The theory of inventive problem solving is different from other traditional approaches of problem solving. The latter rarely offer satisfactory solutions to the outcome of the implementation of a technique or a method; and often disappoints the designer. These approaches are:
- The trial and error approach that accepts compromise between system elements while seeking a solution randomly;
- The "brainstorming" approach which is closely linked to individual skills;
- The experimental design approach, which is complex and only allow finding solutions in a known direction.

Unlike the above approaches, TRIZ excludes compromise when solving a problem. It focuses instead on an ideal solution that is the key to innovation [9]. It has now become universal and apart from being primarily used for technical and physical problem solving, TRIZ is currently applied to solve non-technical domain problems and situations.

3.2. IDM and its knowledge model

- *The Inventive Design Method*

IDM was developed to solve the drawbacks encountered with conventional TRIZ [10]. Actually, TRIZ is not formalized and also has the disadvantage of not being able to be instantiated. Therefore, modeling of TRIZ has been initiated within the LGeCo, Laboratory of Design Engineering, with the aim of providing assistance to inventive design experts.

In addition to TRIZ, IDM also derives from OTSM-TRIZ [9]. And like OTSM-TRIZ, IDM uses the contradictions to solve problems and shares its fundamental assumptions. These assumptions, also known as postulates, are the following:

- Any problem solvable by TRIZ must be formulated as a contradiction
- Technical systems evolve according to objective laws. The best solution is the one that complies with these laws.

The best solution is the one that involves the least possible new resources.

Bultey et al [11] worked within this framework on an ontological model based on substance-field concepts analysis to stimulate problem solving through the use of description logics. An ontology of IDM was also built by Zanni et al 2008 [12]. The next section presents the IDM ontology and its different concepts.

- *IDM ontology*

An ontology may be defined as the standard representation of a domain or field of the important categories of objects or concepts, which exist in the field or domain, showing the relations between them. In other terms, ontology uses concepts and relations to organize domain knowledge and even to support knowledge extraction. IDM ontology, unlike other ontologies, is generic and wants to be applicable in all areas without any restriction [13]. Therefore, it differs from other ontologies applied for patents that are very specific and static. The main concepts of IDM were presented previously in [14]. There are the problems, the partial solutions and contradictions that include elements parameters and values.

Problems describe unsatisfactory features of a system or a method while partial solutions bring progress or improvement to the method or the artifact. A problem must represent the main problem. As for a partial solution, it must be the simplest possible.

Parts or components of a system are called elements. They have parameters. Two types of parameters can be distinguished. On the one hand, we have Parameters of Action on which one can act. On the other hand, there are Evaluation Parameters which cannot be changed, but remain useful to measure the results of a design choice. As for values, they are used to qualify parameters.

IDM proves to be an effective method that contributes to reduce the time spent in R&D during a design process.

4. The proposed methodology

The previous section reported on current progress made in patent mining and more specifically on patent mining for TRIZ. This part presents firstly presents main approaches in knowledge extraction before exposing the methodology we use to extract knowledge relevant to IDM.

Two main approaches can be distinguished in the field of information extraction. We have, on the one hand, the data oriented approach; and on the other hand the knowledge oriented approach.

Data oriented approach consists in a statistical processing. In this approach, analysis results are represented as clouds or charts. This allows for quick interpretation of results. Data oriented approach includes:

- Factor analysis method i.e. canonical approach, discriminant analysis, etc.

- Automated classification method, i.e. bottom-up and top- down methods, etc.

A patent document contains several items for analysis. It is made up of several parts. Some are structured like patent numbers; filing date, etc. and others are unstructured, such as claims, abstracts, claims and descriptions of the invention. The unstructured patent section contains narrative text i.e. the patent title, abstract, claims, and description. As for the structured patent section, it comprises information, such as the inventor of the patent, assignee, date of publication and citation information. From all these parts, it appears that it is the cover sheet, that is predominantly processed by a quantitative method. As an example, the bibliometric method is known to be used for patent trend detection.

However, as Bereau & Dou [15] mentioned it, there is a dissociation between structured parts analysis and unstructured parts one. Data oriented approaches fit well to structured data. Unstructured data require other methods of processing such as knowledge oriented methods.

A data oriented approach rests upon text-mining processing method and is based on linguistic analysis. This approach generally consists of text pre-processing such as lemmatization, tagging, and segmentation, named entities or concept recognition, and uses statistics sometimes.

The approach used in our approach is knowledge oriented since the knowledge to be extracted is located in the patent document unstructured parts. More specifically, we propose to use a set of generic linguistic markers. Marti Hearst [16] demonstrated that, in unstructured texts, it was possible to look for specific lexical relations, which are frequently expressed throughout the text. After developing a list of terms, which reflects the desired relationship, patterns of expressions are grouped in representative patterns. Another study of Simon Teufel [17] tries to identify meta-speech markers to reveal the semantic and the logic organization of text.

4.1. The implementation

The underlying idea of our method is the use of generic linguistic markers as clue words to identify and extract knowledge useful for IDM. Unlike existing approaches in Patent mining, our approach is domain-independent. It performs the discovery of IDM knowledge using the patterns created from the observation of how IDM concepts are expressed syntactically in patent documents.

Hence, the starting point of our method is the building of the training patent corpus. The process was explained previously in [18 ; 19].As a reminder, the corpus contains 100 patents from various domains ranging from Engineering to Chemistry. In order to deal with issues regarding representation and IDM knowledge model so to produce an effective knowledge extraction process, our working model has been divided into two steps. The first step is the preprocessing aimed at producing both training information for further evaluation and the initial population of extraction automata. The second step constitutes the knowledge extraction itself. More precisely, this aims at applying our automata on test corpora and evaluating them to complete or extend the initial population with unseen extraction patterns.

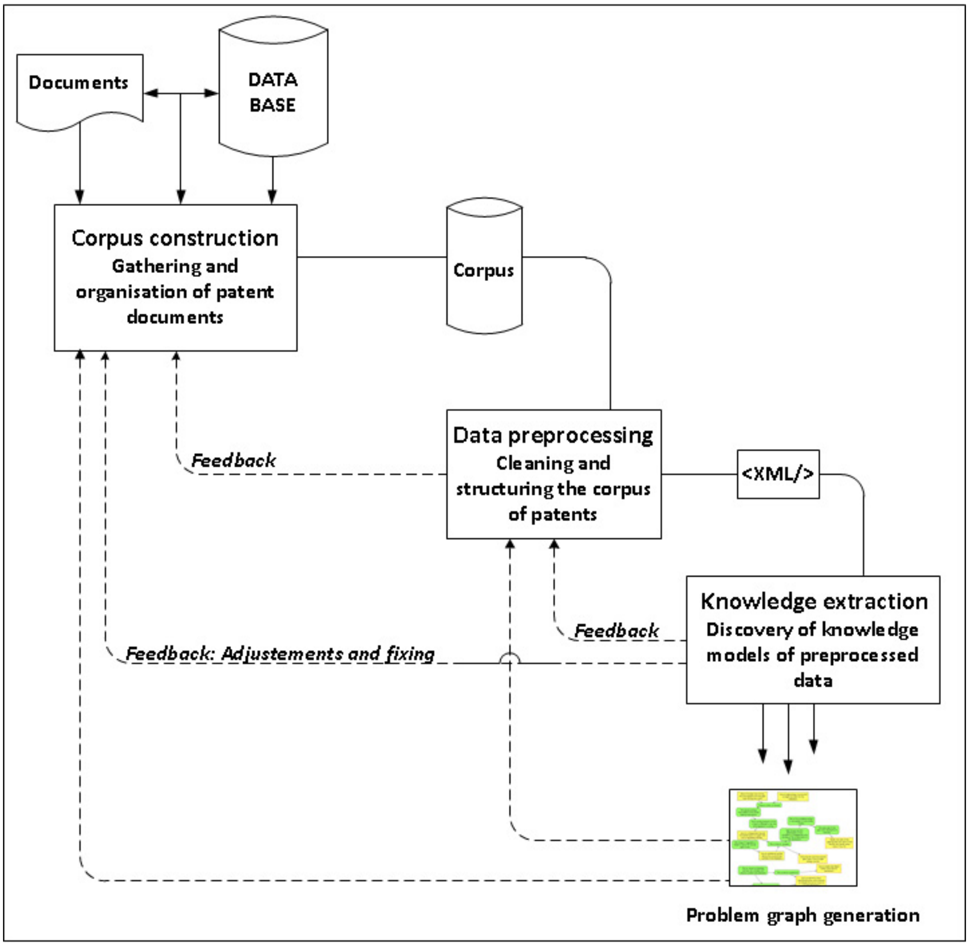
Documents
DATA BASE
Corpus construction
Gathering and organisation of patent documents
Corpus
Data preprocessing
Cleaning and structuring the corpus of patents
<XML/>
Feedback
Knowledge extraction
Discovery of knowledge models of preprocessed data
Feedback
Feedback: Adjustements and fixing
Problem graph generation

Fig. 1 The training process

4.2. Text preprocessing and training step

The preprocessing phase includes two main goals: to extract relevant information from patent documents and to use that information to build extraction automata.

With respect to text preprocessing, an essential principal in our method is to be able to make a good use of the patent document structure for the knowledge extraction. It is known that the patent document is very specific and has inherent complexities [20]. So we have restricted our scope to some extend to study a few parts of the patent documents i.e. the cover sheet (mainly the abstract), the first claim and the descriptions part.

From this study a set of linguistic markers where found interesting to find IDM knowledge [19]. Patents are semi-structured documents. Each patent makes the state of the art of the artifact by firstly underlying known technical problems and before displaying the provided solutions. Therefore, it is relatively easy to discover some regularities and translate them into extraction rules.

The first clues which appear in the first sentence of the abstract section are phrases such as "*a method for*", "*a process for*", "*an apparatus for*", "*for an invention*", "*the purpose of this invention is to*." These expressions are often used to describe the invention or the method patented. Note that these clues are repeated in the claims and the descriptions sections. They are mostly followed by verbs expressing improvement and are productive markers.

Consider, for instance, that we are given the following extracts where bold sequences of words represent the markers triggering the extraction patterns:

> The invention relates **to a process for**
> GOAL[separating non-metallic inclusions from hot liquid metal],
> **in particular**
> OBJECT[from aluminum killed soft steel, in continuous casting plants],
> **wherein**
> METHOD[the metal supplied into a tundish is guided at least once in upward direction to the surface of the metal sump in said tundish under formation of at least one stream at a speed greater than that of the metal in the tundish prior to its being deflected],
> **whereby**
> RESULT[a wave is formed at the surface of the metal covered by slag].

> When (condition)
> ACTION [partial_solution {the content is more than about 4%}]
> ,
> CONSEQUENCE [problem {segregation may seriously occur}]
> **and, as a result,**
> THEN_CONSEQUENCE [problem {toughness and pipe-expansion properties are degraded}].
> **Hence,**
> SOLUTION [partial_solution {the content of Mn is set in the range of about 0.5% to about 4%}].

In the first frame, the phrase "*the invention relates to a process for*" introduces the field of the patent as well as the artifact patented. Furthermore, expressions such as "wherein" and "whereby" which are respectively a conjunction and a pronoun can be noticed. These are formal forms which have respectively "*in which*" and "*by which way, or by which method*" as equivalent. While "*wherein*" is used to give precision on the results or the effects provide by the invention, "*whereby*" and "*thereby*" bring more precision on the consequences of these results.

In the second frame, a typical sequence relevant for IDM knowledge extraction is shown. As we can notice, sentence connectors, such as those highlighted in bold in the previous frames, may be used to identify implication links for the different concepts. These are:

Condition:
So long as, as long as, if, provided that, etc.…
Opposition:
however, conversely, by contrast, in contrast, whereas, contrary to this, etc..
Purpose:
in order to, so as to, so that, with the purpose of, etc.
Result:
as a result, therefore, thus, in consequence, etc.

5. Results and discussion

To evaluate the accuracy of our model, comparison was made between automatic and manual extraction of IDM concepts. The results obtained showed some imperfections. For example the presence of duplicates i-e similar sentences were noticed. Furthermore, some extracted concepts were not complete or did not comply with IDM ontology requirements.

Relevant results sometimes need to be reformulated to meet the criteria for IDM. The problems syntax is "*subject* + *verb* + *objects*". As for partial solutions, they must obey the following "*infinitive* + *complement*" syntax. Considering the following example:"*Deteriorating the casting nozzle*", taken from the sentence "*The casting process is interrupted, thereby deteriorating the casting yield*", even if it is relevant from the point of view of IDM has not the right format. Correct reformulation would be "*The casting yield is deteriorated*".

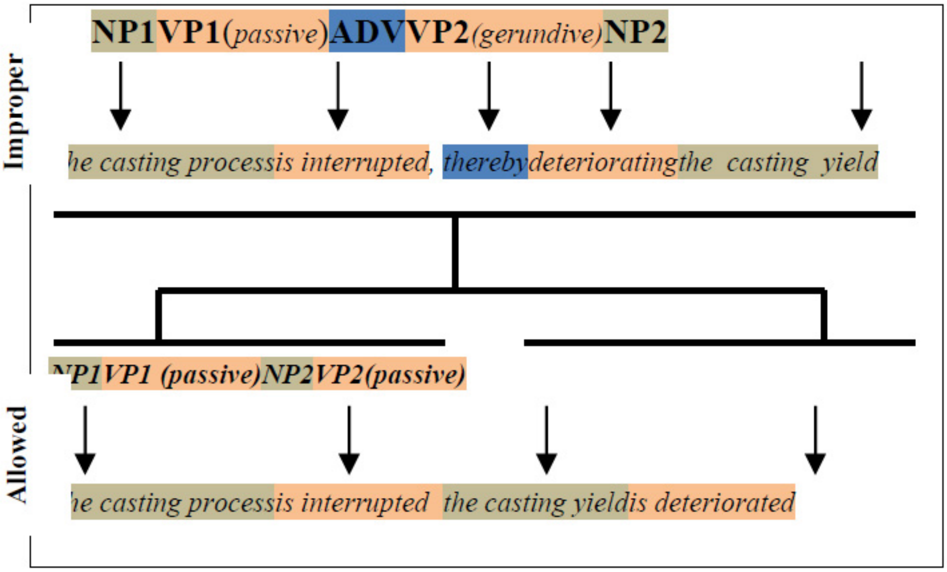
Improper
NP1VP1(passive)ADVVP2(gerundive)NP2
he casting processis interrupted, therebydeterioratingthe casting yield
Allowed
NP1VP1 (passive)NP2VP2(passive)
he casting processis interrupted the casting yieldis deteriorated

Fig. 2. Example of Meaning - Text reformulation

Thus, using the text reformulation model based on the Meaning- Text theory (NASR 1996) it is possible to obtain the fixing of the extractions obtained. This theory proposes indeed use paraphrasing to achieve reformulation. It relies on the identification in the grammar and lexicon of permitted and prohibited structures. Prohibited structures will then be replaced in the syntactic representation by authorized through the system paraphrases structures. Modeling the phrase "*the casting process is interrupted, **thereby** deteriorating the casting yield* " as shown in fig.2.

To conclude this section, it would be interesting to include performance measures of our model. Results obtained are displayed in the figure below.

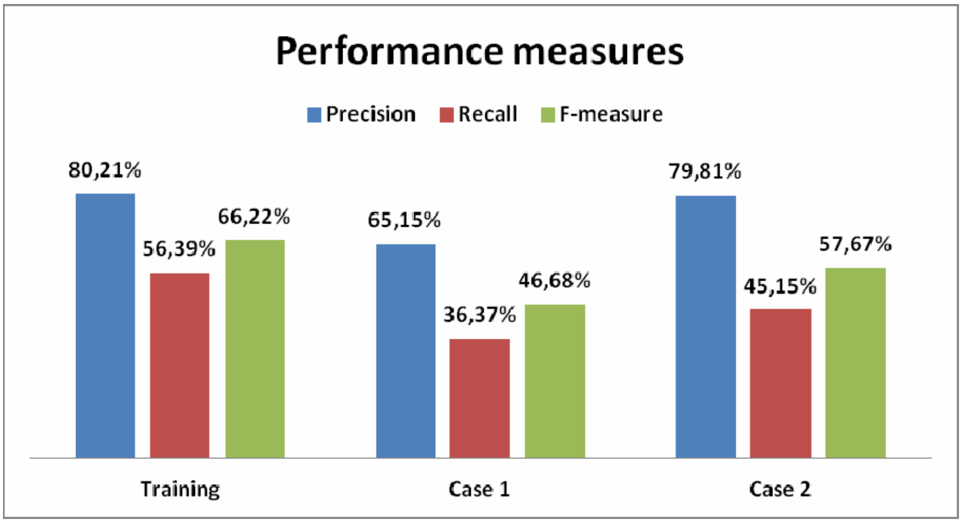

Fig. 3. Performance measures

What we can notice is the recall rate that remains low. We also notice that the quality of the corpus of extraction may highly impair the extraction quality. Thus, we are working on improving request form to make sure the patents retrieved from the Internet fits with the user needs.

In addition, we precise that the reformulation method used here only works, at the moment, for simple intendent c lauses. However, we are working to spread it to other type of clauses to greatest extent.

6. Conclusion and perspectives

In this paper, we first presented the progress made so far on patent knowledge extraction, specifically contribution made so far on patent mining related to TRIZ. Then, we pro- posed a method to automate the extraction of IDM concepts and the reformulation of items that do not comply with IDM knowledge concepts. The premise of this method is to retrieve IDM concepts using generic linguistic markers. The originality of the method is that it will provide a solution to assist IDM experts in their task. Furthermore, this method uses text-mining techniques. The progress made so far, such as the items reformulations, improves the problem graph automatic generation. Results obtained, so far, are encouraging. However, the evaluation of the result obtained reveals that although the method is encouraging, it is not fully satisfactory yet. Some limits to repel always impacts the quality of the results. We still have duplicates due to the fact that some items are syntactically different but semantically similar or close. For example, "*the tank pressure falls below the regulation pressure*" and "*the pressure in the tank decreases.*"Thus, it would be interesting to process these duplicates, improve the algorithms to improve IDM core concept representation.

References

[1] Altshuller, G.S (1998) 40 Principles: TRIZ keys to technical innovation. (Lev Shulyak et Steven Rodman, Trans.). Worcester, MA: Technical Innovation Center, INC. 1998, 141p. (1st Ed, 1998), ISBN-10: 0964074036.

[2] Assad Abbas, Samee. U. Khan(2014) A literature review on the state-of- the-art in patent analysis. World Patent Information.

[3] Feldman R, Fresko M, Hirsh H, Aumann Y, Liphstat O, Schler Y, Rajman M (1998) Knowledge Management : A Text Mining Approach". In: Proceedings .of the 2nd International Conference on Practical. Aspects of Knowledge Management, Basel

[4] Ghoula N, Khelif K, Dieng-Kuntz R (2007) Supporting Patent Mining by using Ontology-based Semantic Annotations. In: Proceedings. of IEEE/WIC/ACM International Conf. on Web Intelligence, Silicon Valley, USA.

[5] Savransky S (2000) Engineering of creativity: Introduction to Triz methodology of inventive problem solving". Boca Raton. USA.

[6] Cascini G, Russo D (2007) Computer-aided analysis of patents and search for TRIZ contradictions", Int. J. Product Development

[7] Altshuller, G.S. (2004), Et soudain apparut l'inventeur : Les idées de TRIZ. Paris, Ed. Seredinski 166p. (The art of Inventing – And Suddenly the Inventor Appeared, Moscow: Detskays Literatura, 1ère édition: 1984), ISBN-10: 2952139415..

[8] Blossier, J.E. (2002), Guide d'initiation à TRIZ. PSA Peugeot Citroën, 85p. http://crrm.u-3mrs.fr/sfba/ile-rousse/1997/article1.pdf

[9] Cavallucci D, Khomenko N (2007) From TRIZ to OTSM-TRIZ, Addressing complexity challenges in Inventive design. In: International Journal of Product Development.

[10] Dubois, S. et al. (2004), Modélisation des concepts de formulation des problèmes de la TRIZ. Actes des 15èmes journées francophones d'Ingénierie des Connaissances, IC'2004, Lyon. 2p.

[11] Bultey, A., Bertrand, De Beuvron, F.et Rousselot, F. (2007), A substance-field ontology to support the TRIZ thinking approach. IJCAT, vol.30, no.1, pp.113-124.

[12] Zanni, C., Rousselot, F., Cavallucci, D. (2008), KAID: a tool for conducting the use of inventive conception in leading complex studies. Actes de SKIMA 2008, Katmandu, Nepal.

[13]Rousselot, F., Zanni, C., et Cavallucci, D. (2007), Une ontologie pour l'acquisition et l'exploitation des connaissances pour la conception inventive. Revue des Nouvelles technologies de l'information, numéro spécial sur la modélisation des connaissances.

[14] Cavallucci, D., Rousselot, F., Zanni-Merk, C., (2008), Representing and selecting problems in Contradiction Networks. 2ème session de l'IFIP sur l'innovation assistée par ordinateur.

[15] Bereau, P. et Dou, H. (1997), La classification neuronale pour la détection de nouvelles tendances de recherche et le développement de nouveaux produits. CIFRE/CRRM, Université d'Aix-Marseille, 17p

[16] Hearst, M. (1992), Automatic Acquisition of Hyponyms from Large Text Corpora. Actes de la 14ème conférence internationale sur la linguistique informatique (COLING), Nantes, pp.539-545.

[17] Teufel, S., Siddharthan,A., Tidhar, D., (2006),Automatic classification of citation function,Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing,,,103-110,2006

[18] Gabor, K., Rousselot F., De Bertrand de Beuvron F., (2009) Extraction de connaissances orientées évolution dans les textes techniques, Proceedings TOTh conférence.

[19] Souili A, Cavallucci D, Rousselot, F, Zanni, C (2011) Starting from patent to find inputs to the Problem Graph model of IDM-TRIZ. In: TRIZ Future Conference 2011, Dublin – Ireland

[20] Guyot, B. et Normand, S.. (2004) Le document brevet, un passage entre plusieurs mondes, Paris, 22p.