Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 131 (2015) 635 – 643

Procedia Engineering

www.elsevier.com/locate/procedia

World Conference: TRIZ FUTURE, TF 2011-2014

Natural Language Processing (NLP) - A solution for knowledge extraction from patent unstructured data

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

Patents are valuable source of knowledge and are extremely important for assisting engineers and decisions makers through the inventive process. This paper describes a new approach of automatic extraction of IDM (Inventive Design Method) related knowledge from patent documents. IDM derives from TRIZ, the theory of Inventive problem solving, which is largely based on patent's observation to theorize the act of inventing. Our method mainly consists in using natural language techniques (NLP) to match and extract knowledge relevant to IDM Ontology. The purpose of this paper is to investigate on the contribution of NLP techniques to effective knowledge extraction from patent documents. We propose in this paper to firstly report on progress made so far in data mining before describing our approach.

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

Keywords: TRIZ, Inventive Design, Text mining, Patent mining, Knowledge discovery;

1. Introduction

Patent mining has long been considered as useful to engineers for R&D management and according to the World Intellectual Property Organization, 90% to 95 % of all inventions are found in patent documents [1]. Thus, patents constitute a first choice media, when speaking of technological information, where one can understand the evolution mechanism of an artifact by its observation throughout the years.

* Corresponding author. Tel: +33 3 88 14 47 00.
E-Mail address: wendemmi.souili@etu.unistra.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.2015.12.457

However, knowledge contained in patent documents is underexploited in innovative design processes. Furthermore, due to the continuous increase in the amount patents data currently available and apart from few contribution on patent mining, it becomes more than necessary to valorize and automate the extraction of patents knowledge to assist the process of designing new products.

Nevertheless, most of existing patent mining tools are only focusing on explicit or structured data; and in this context, Text Mining (TM) techniques turn out to be a good approach to knowledge extraction from unstructured and implicit textual data.

IDM derives from TRIZ, the Theory of Inventive Problem solving. Both were developed to assist engineers in their invention process and consist of logic, data; and research rather than intuition [2; 3] to better assist R&D activities, using patent mining. IDM was developed to enhance classical TRIZ in formalizing its methodological translation and therefore solve most of its limits. Actually, the comprehension of classical TRIZ is often felt as a complex task and it is difficult to implement an abstract computational design model upon its techniques and concepts. OTSM [4] later developed these concepts but is unfortunately unachieved so far. In this context, patent documents can be reckoned as a reliable source of human activity and IDM as well as TRIZ its underlying theory are grounded on the assumption that artifacts evolve according to a number of objective laws. Unlike other traditional problem solving methods, IDM helps to break psychological inertia which is vital in an inventive process through the identification of contradictions.

The transition from a generation of an artifact to another is symbolized in a patent in a non-explicit way through the overcoming of a contradiction with no compromise. Thus, patent documents can be considered as vital source of knowledge for IDM.

Ontology may be defined as the standard representation of a set of concepts or objects and their relation within a field or domain.

A universal ontology was created to interface IDM software for technology intelligence, in order to solve problems belonging to different fields and the make easier the finding of the most difficult contradiction. Main concepts of IDM ontology are Problems, Partial Solutions, Action Parameters and Evaluation Parameters. We later define these concepts and what we mean by "contradiction".

The target of the research underlying this paper is to reports on an on-going project to conceive a knowledge extraction system based on TM to assist inventive design R&D experts in their activities. Our approach is completely new and proposes the combination of TM techniques to automate a qualitative extraction of IDM knowledge after the analysis of the generic linguistic phenomena that occur within patent text; using syntactic, semantic and discourse information and graph visualization tool to display IDM concepts [5] with their relations.

Firstly, an overview of current contributions in TRIZ and patent mining are proposed. Section 3 is devoted to the presentation in detail of IDM ontology. We later expose our method in section 4. Before concluding in section 6 with the conclusion and perspectives part, a discussion on the proposed method is presented in section 5.

Nomenclature

EP	evaluation parameter
AP	action parameter
PB	problem
SP	partial solution

2. Theoretical background

Data mining, also known as knowledge discovery, is a recently new discipline. This section is devoted to major contribution made so far on patent mining in general and more specifically on patent mining for TRIZ. The development of Internet has made available a large amount of patent databases and the number of patent documents is daily increasing dramatically. Many knowledge discovery systems including some related to TRIZ have intended

to automate knowledge retrieval or extraction from patent documents. And despites many existing studies on patent mining; the discipline has not reached maturity yet.

They typically use hybrid methods combining statistics and linguistics. An example is Feldman & al [6] method which implements a text mining at the term level to produce a list of candidate words to be keywords after a basic linguistic preprocessing. Such an approach is promising for processing unstructured data. Other studies such as Ghoul & al [7] present a processing chain achieving an automatic semantic patents annotation through a structural ontology and domain ontology; biology as regarding their case.

With the development of TRIZ, several researches were taken to automate its processes. Among theses, S-A-O based approaches which suits better for unstructured patent sections mining are worth noting. S-A-O based approach explicitly represents relationships between the components of a patent. It is intrinsically connected to the concept of function, understood differently by different authors.

According to Cascini et al. [8; 9], functions performed by or on components are represented by Action which constitutes with the Subject, the component of a system. They particularly advocate the use of functional analysis to identify a problem and generate leading-edge solutions. Solving a problem involves breaking it into its component functions which are later broken down in sub functions until the function level for solving the problem is reached. Expressed as Subject-Action-Object triads [9], functional analysis is found to be relevant for representing knowledge related to patents key findings and inventor's domain of expertise.

Another understanding of S-A-O is given by Moehrle & al. in [10]. They propose to tune Multi–Dimensional Scaling (MDS) to S-A-O structures to map technological convergence between two companies. It is also worth noting Yoon & al. [11] proposition to automatically identify TRIZ trends by using binary relations of "verbs+nouns" or "adjectives+noun" to define specific trends and trend phases through semantic sentence similarity measuring.

3. IDM ontology

The starting point of our research is the specific need of design engineers for a tool that rapidly captures all information related to a specific topic on their projects to starts projects with exhaustive understanding of initial situations. Within the framework of IDM, designers often use patents documents to benefit from the knowledge contained therein to assist their invention process. However, the development of Internet has made available a huge amount of patents texts, thus making the research task longer more tedious. Therefore, automating the patent mining task to facilitate the work of engineers is welcome.

Despites the large amount of approaches and tools available on patent mining and knowledge extraction for TRIZ, very few matches the real needs of designers looking for systematic innovation. Most of the existing tools use domain specific ontologies and are limited to these domains.

In order to meet these needs, IDM (see in the next section) was created. It derives from TRIZ [3], has a generic ontology which includes several key concepts like partial solutions, problems, parameters and values inherited from OTSM reflexions. The next section is devoted to the presentation of IDM knowledge model.

As a reminder, ontology may be defined in the context of knowledge extraction as a specification of a conceptualization. It is a description of the concepts and relationships that can exist for an object/concept or a category of objects/concepts. IDM deals with artifacts evolution and was developed to address wider and more complex and problematic situations [5]. The method assumes like TRIZ that artifacts are the result of evolution guided by objective laws. IDM ontology is generic and is intended to be applicable to any domain.

The basic concepts of IDM Ontology are problems, partial solutions, and contradictions which contain elements, parameters and values. Problems express unsatisfactory features within a system or a method while on the contrary, partial solutions are element of change or improvement presenting a result known in the domain or proved by experience. A problem, according to IDM, has the following literal structure: *Subject+Verb+Complement*. It must express an essential problem. As for partial solution, this must comply with the following syntax: *Infinitive+Complement*; and be the most simple possible. Parts or components of a system are called elements. Parameters are complement nouns and are of two sorts. On the one hand, there is EP, the value of which can only be observed and having one direction of evolution wanted while the other unwanted. They are useful to evaluate the results of a design choice. On the other hand, we have AP that can be modified by experts and has two opposite and

logical directions of wanted evolution. As of values, the fundamental aspect of the concept of contradiction lies in the opposition between values; and in the fact that two opposing values must be expressed.

4. Our method for extracting IDM knowledge

We propose to use lexico-syntactic patterns [12] to match and extract IDM knowledge from patent databases. Our method come from NLP and is based on finite state automata theory. Firstly used by Hearst [13] to retrieve hyponyms from texts, it was then developed by Morin [14; 15; 16] within the framework of the acquisition of patterns for the identification of hierarchical relationships between terms.

The method consists in three steps including the semantic resources collection, IDM knowledge identification and the problem graph generation. They were already presented in detail in previous papers [17; 18].

Briefly speaking, a set of 100 patent documents extracted from USPTO was used in the development of the system; 87 patents were reserved for the evaluation. More precisely, the first set of 100 patents were manually selected. These patents were from various domain with the concern to build as much as possible a complete model of matching and extraction. A patent is a set of structured information in several fields. Due to lack of space, we cannot reproduce an extract of the corpus in the present paper. However, for those looking for more information on the patent structure, the website of WIPO[†] could be of great interest. Relevant information in patent databases may be sometime drowning in a complex syntactic structure. Sentences in patent documents are sometimes long and may contain more than 500 words. Not all the information contained in Patent documents are useful for IDM experts. Some information may also be duplicated. Therefore, it is important to create pre-treatment modules to extract relevant sections and structure them when necessary. These processes are built in java and includes four main steps. When a user inputs a query, relevant patent documents are retrieved from USPTO, Espacenet or from a local patent database to form the extraction corpus. Then comes the extraction process with these following four main steps:

- Selection of relevant text areas: The areas selected for the mining process are the patent number, the first claim, prior art and the description part. We also selected the international code that is useful for determining the technology covered by the patent.
- Segmentation in paragraphs: we chose to segment the patent document at the paragraph level. The motivation of such a choice is that we consider that a group of sentence about one main idea, the topic.
- The matching and tagging of knowledge relevant to IDM: this step consists in matching and tagging relevant knowledge with XML tags to later parse and extract them. For this task, we use finite state automata. Finite state automata provide efficient and convenient tools for the representation of linguistic phenomena. Within the framework of NLP, the use of finite state automate has already proved to be successful in various areas [18] As for IDM concepts, our previous have presents more details on finite state automata.
- The knowledge extraction to fill IDM ontology and the problem graph generation: The final step before the problem graph generation is the knowledge extraction. Once IDM concepts are identified, candidate concepts are selected and presented to the user. The latter has the possibility to edit boxes and complete the relationships.

[†] www.wipo.int

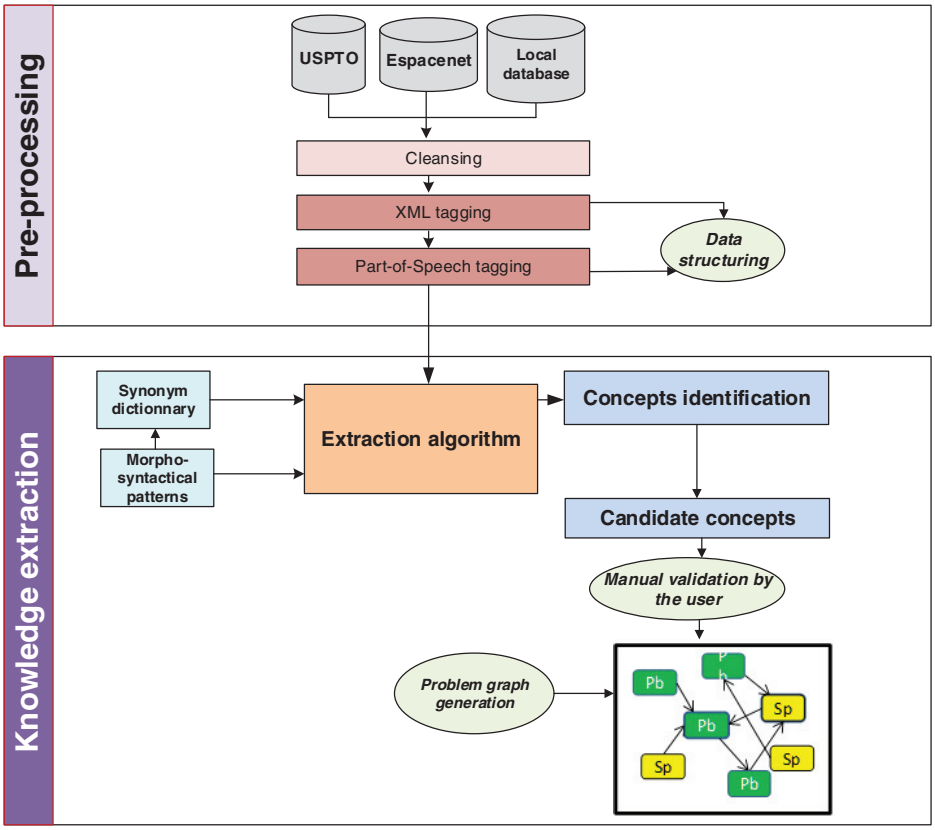

Fig. 1: Problem graph generation

The method was integrated in STEPS a research tool developed in our laboratory, LGeCO (Design Engineering laboratory). The user has the possibility to either create the problem graph manually or automatically from a corpus of patent retrieved from the patent databases. Fig. 1 shows the query window where the user can search keyword contained in the different parts of patent document i.e. title, abstract, description, claims; and use Boolean operators “and/or” to refine his research. Let us precise that for the time being, we do not have our own search engine. We currently use patent databases search engine remotely to perform our queries. The result is shown as a checkbox list (Fig. 3) and the user can select the select the patents to be used in the extraction corpus. Fig. 4 presents a problem graph generated from a corpus related to “mascara” and “enzymes” (fig.5). Relevant knowledge are annotated with respective XML tags.

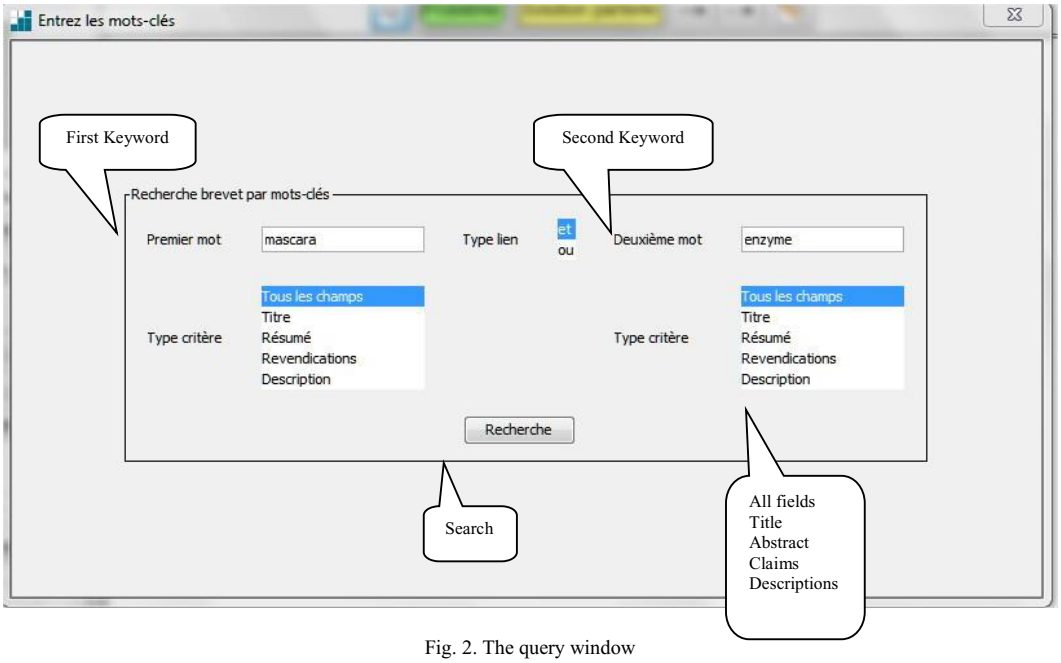

Fig. 2. The query window

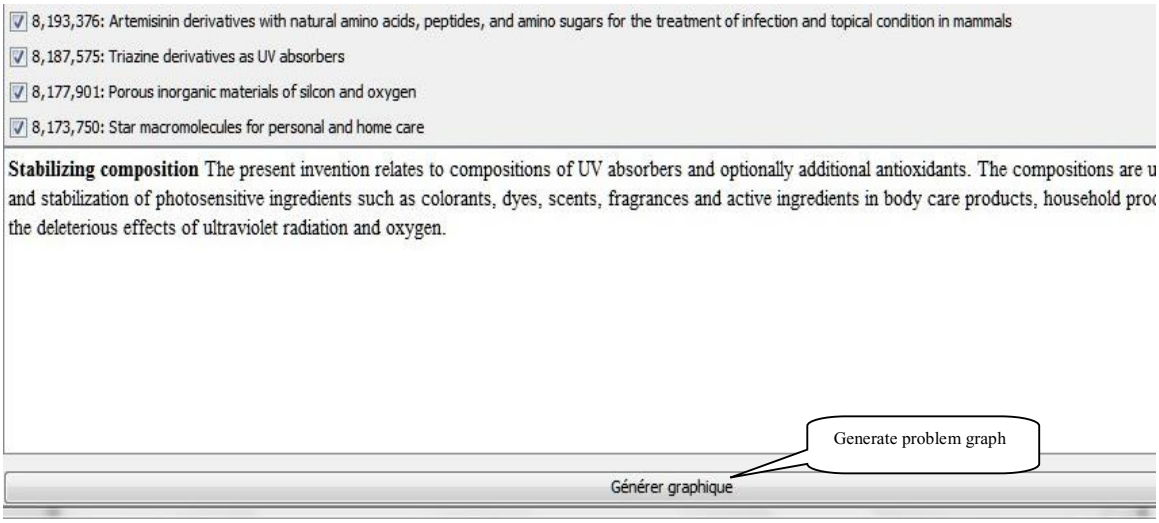

Fig. 3. Result shown as a checkbox list with an overview of abstracts

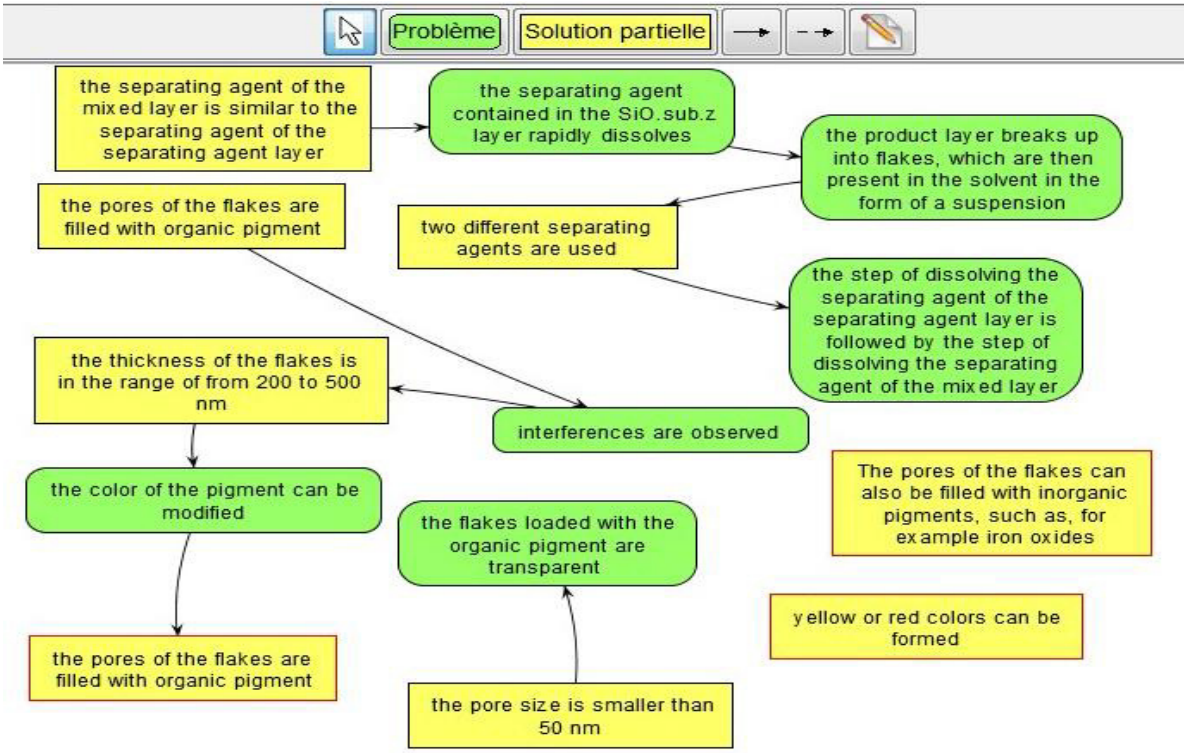

Fig. 4.Problem graph

```
<corpus>
<patent num="8,177,901">
<title>Porous inorganic materials of silcon and oxygen</title>
<abstract>...</abstract>
<claims>...</claims>
<description>
...The temperature of the solvent should be so selected that its vapor pressure is in the indicated pressure range.
<seq>With mechanical assistance, the separating agent layer and if <solupart>the separating agent of the mixed layer is similar to
the separating agent of the separating agent layer</solupart>,<problem> the separating agent contained in the SiO.sub.z layer
rapidly dissolves</problem> and <problem>the product layer breaks up into flakes, which are then present in the solvent in the form
of a suspension</problem>. If <solupart>two different separating agents are used</solupart>, <problem>the step of dissolving the
separating agent of the separating agent layer is followed by the step of dissolving the separating agent of the mixed layer
</problem>.</seq>...</description>
...
</corpus>
```

Fig. 5. Sample of an extraction corpus.

5. Evaluation

For the purpose of measuring the performance of our method, we applied it on two different corpora. The first corpus (Corpus A) was constructed with the keywords “steel and martensite” and the second corpus (Corpus B) with “mascara” and “Enzyme”. Both results were then evaluated to calculate their respective recall and precision rates. Recall is the ability of an algorithm to present all relevant concepts. As for precision, it is the ability of an algorithm to present only relevant concepts.

Table 1. Recall and precision for corpus A

	Recall	Precision
Partial solution	44,22%	73,53%
Problem	46,10%	86,10%

Table 2. Recall and precision for corpus B

	Recall	Precision
Partial solution	34,14%	62,17%
Problem	39,40%	68,13%

As you can notice, performance measures vary from a corpus to another. However, our main concern is the precision rate we are trying to maximize. Therefore, unlike our previous studies we have this time focused on precision by adopting strict algorithms, even if this means reducing the recall rate. Furthermore, increasing the precision involves multiplying evaluations on different domains to assess the generic-ity of our model of extraction.

To finish with, our system is not limited in the number of patents to process. However, choosing a large number of patents may impact the system performance.

6. Conclusion and perspectives

In this paper, we first reported the existing literature in patent mining, specifically the contributions made so far on TRIZ. Then, we proposed our method to automate IDM concepts extraction from patent document. Performance measures that only focused partial solution and problem reveal encouraging good scores for precision but very low recall scores for recall. Even though we are only focusing on improving the precision scores, the scores show that our method has not reached maturity yet. Besides the multiplication of automata, effort should also focus on extending lists of markers and lists sentences connectors. We are not pretending to create a tools that will replace completely the expert (there will always be noise and silence which may impact results quality). However, our idea is to provide him an assistance to make easier the mining task and display as the problem graph the essential of the patent documents as regarding problem, partial solutions, parameters, elements and values.

Likewise, it would be interesting to measure the impact of the automatic or semi-automatic extraction of IDM knowledge on R&D teams on real cases.

References

[1] Nave Z. (2010), Patent Information Gathering and Intellectual Property, in: Globes. http://www.cgi.co.il/

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

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

[4] Khomenko, N and Ashtiani, M (2007) Classical TRIZ and OTSM as a scientific theoretical background for non-typical problem solving instruments. In: TRIZ Future Conference 2007, Frankfurt, Germany

[5] Zanni-Merk C, Rousselot F, Cavallucci D (2011) Use of formal ontologies as a foundation for inventive design studies. In: Computers in Industry 01/2011; 62:323-336. pp.323-336

[6] 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
[7] 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.
[8] Cascini G, Fantechi A, Spinicci E (2004) Natural language processing of patents and technical documentation. In: Lecture Notes in Computer Science, 3163; 2004, p.508–520.
[9] Cascini G, Russo D (2007) Computer-aided analysis of patents and search for TRIZ contra-dictions", Int. J. Product Development
[10] Moehrle M, Geritz A (2004) Developing acquisition strategies based on patent maps. In: Proceedings of the 13th Internationa1 Conference on Management of Technology; Washing-ton, USA: R&D Management
[11] Yoon J, Kim K (2011) An automated method for identifying TRIZ evolution trends from pa-tents. Experts Systems with Applications
[12] Jilani, I., Grabar, N., & Jaulent, M.-C. (2006). Fitting the finite-state automata platform for mining gene functions from biological scientific literature. Paper presented at the Semantic Mining in Biomedicine, Jena (Germany)
[13] Hearst, M. (1992), Automatic Acquisition of Hyponyms from Large Text Corpora. In : Proceedings of the 14th conference on computational linguistics (COLING), Nantes, pp.539-545.
[14] Morin, E. (1999). Acquisition de patrons lexicosyntaxiques caractéristiques d'une relation sémantique. Natural Language Processing (NLP), 40(1), 143-166.
[15] Morin, E.(1998). "Prométhée: un outil d'aide à l'acquisition de relations entre termes", In Proceedings of Natural Language PRocessing conference (TALN'1998), Paris, pp.172-181.
[16] Poibeau, T, Villavicencio A., Korhonen A., Alishahi A. (2012). Computational Modeling as a Methodology for Studying Human Language Learning . In Cognitive Aspects of Computational Language Acquisition. Springer (in press)
[17] Souili A, Cavallucci D, (2012) Toward an automatic extraction of IDM concepts from patents. In: CIRP Design Conference 2012, Bangalore – India
[18] Souili A, Cavallucci D,Rousselot F, (2012) A lexico-syntactic pattern matching method to extract IDM-TRIZ knowledge from on-line patent databases. In: TRIZ Future 2012, Lisbon – Portugal
[19] Paumier, S. (2008), Unitex 2.0 User manual. Université Paris-Est Marne-la-Vallée. http://igm.univ-mlv.fr/~unitex/