Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 131 (2015) 418 – 425

Procedia Engineering

www.elsevier.com/locate/procedia

World Conference: TRIZ FUTURE, TF 2011-2014

A lexico-syntactic pattern matching method to extract IDM- TRIZ knowledge from on-line patent databases

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 increasingly of great importance to innovation performance. A patent document is an important source of technological knowledge where artifact evolves according to problems and partial solutions. This paper reports on an on-going research whose goal is to provide design engineer with an efficient patent mining tool that re-structures a patent corpus in a problem graph to better highlight state of the art of given field of knowledge. The methodology employed to achieve such a functionality is based on a NLP method that manage both IDM-TRIZ concepts and patent corpuses contents. The paper also presents a case study in steel making industry to illustrate the methodology.

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

Keywords: Knowledge engineering, Patent mining, Ontology, Inventive Design, NL;

1. Introduction

"More than 80 percent of man's technical knowledge is described in patent literature" [1]. The internet has made available large amount of patent databases supplied by thousands of contributions each year. It becomes thereby important to extract and structure information from patent documents in order to make patent knowledge accessible to engineers.

Much of the information contained in patent documents is expressed in natural language texts which need collaborative tools from fields like computational linguistics, cognition and informatics. From visual and bibliometric analysis to semantic searches, many approaches exist in patent analysis. However, despites their obvious efficiency, patent mining techniques have not reached maturity yet. They are still limited in their domain of application and do not address all the patent sections.

The development of TRIZ, the theory of problem solving [2], has led to the automation of its use and many authors since its efficiency was obvious. TRIZ and its derived method Inventive Design were created to assist

Peer-review under responsibility of the Scientific Committee of TFC 2011, TFC 2012, TFC 2013 and TFC 2014 – GIC
doi:10.1016/j.proeng.2015.12.437

engineers in their invention process. They consist of logic and data, but not intuition. They thereby turn to be relevant for better assisting R&D activities [2]; [3].

TRIZ is primarily about technical and physical problems. However, it is now applied to almost any problem and situations. The theory assumes that technical systems evolve in a similar way and any problem once reduced to a functional level can be solved by analogy using standard solutions and problem solving techniques, regardless from its domain. Yet there are some disadvantages with classical TRIZ as it has not been completely formalized. In addition, its comprehension is complex task and it is hard to create a computational model upon its techniques and concepts.

IDM was, on that score, created to clarify the concepts involved in TRIZ [2] and solve its limits with the purpose to address wider and more complex and problematic situation.

IDM and its underlying theory assume that artifacts evolve according to a number of objective laws. The transition from a generation of an artifact to another is symbolized in a patent in a non-explicit way through the solving of a contradiction without any compromise. Patent document can be viewed as a reliable sign of human innovative activity.

NLP techniques are good contribution to patent mining tools efficiency. The aim of this paper is to report on an on-going project which deals with IDM knowledge extraction from patent databases.

First, we present a brief state of art on patent analysis for TRIZ. Secondly, we introduce the context of this study. Then, we expose our method based on the use of generic linguistic markers. Finally, we propose experiments validating our approach before presenting conclusion and perspectives for future works.

Nomenclature

PE	Evaluation Parameter
PA	Action Parameter
PB	Problem
SP	Partial Solution

2. NLP tools for patent mining

NLP tools are necessary for intelligent patent mining and knowledge extraction. This section reports on major projects made so far on patent mining in general and specifically on patent mining for TRIZ. The development of Internet has made available a large amount of patent database with the challenge to manage this huge source of knowledge and provide it to engineers.

Patent analysis approaches can be classified into three main categories including: visual and bibliometric patent analysis, data mining and Information Retrieval (IR); and semantic search. While bibliometric patent analysis, also known as bibliometric focuses on the analysis and the organization of large amount of historical data to support decision making [4] ,data mining and information retrieval aim at analyzing data from various perspectives and summarizing it into useful knowledge. Several data mining system have tried with the development of TRIZ to automate its process. Such systems which address knowledge extraction from patent usually utilize hybrid methods by associating statistics and linguistics with the purpose of using the inventive principle to solve problems in different domains. With glossaries, ontologies and thesauri, NLP techniques made a good contribution to the grown of patent analysis tools. Thus, several researches related to TRIZ were taken on this basis with techniques such as SAO-based approaches which suit better for unstructured patent section mining. SAO-based patent analysis which controls the syntactical structure of subject (Noun Phrase), Action (Verb Phrase) and object (Noun Phrase) explicitly represents relationships between the components of a patent. It is intrinsically connected to the concept of function understood differently by different authors. Savranski [5] calls it “the action changing the feature of any object” whereas for Cascini et al. [6] and [7] functions performed by or on components are represented by Action which constitutes with Subject the component of a system. [6] and [7] particularly advocate the use of functional analysis to identify a problem and generate innovative solutions. To solve a problem, this one is broken into its component functions which are later divided in sub functions until

the function level for solving the problem is reached. Expressed as Subject-Actions-Object triads [7], functional analysis proves to be relevant for the representation of knowledge related to the patents key findings and the inventor's domain of expertise.

As for Moehrle et al. [8], they proposes to adapt Multi-Dimensional Scaling (MDS) to SAO structures, in order to map technological convergence between two companies. In addition, Yoon et al. [9] propose to automatically identify TRIZ trends. According to their approach while property refers to a specific characteristic of a system and is usually described using adjectives, function indicates an action which alters the feature of an objects and is usually described with verbs. Therefore, they propose the use of binary relations of "verbs+nouns" or "adjectives+noun» to define specific trends and trend phases through semantic sentence similarity measuring. As for Dewulf, properties are the attributes of a product. "What a product is or has". "They are mainly expressed in adjectives and is related to physical parameters" [10].

3. Context of the study

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 so as their own knowledge in order to start projects with exhaustive understanding of initial situations. In the context of inventive design, designers are often asked to search in patents documents to benefit from the knowledge contained therein to assist their invention process. However, resources are important and their research task is long and tedious. Thus any method of facilitating the work is welcome, especially those based on patent mining.

For this purpose, several works were tempted to automate the analysis process from different approaches as we reported it in the state of art. However, despites the large amount of the approaches and tools available, very few matches the real needs of designers in their quest for systematic innovation. Most of them are based on specific ontologies and are not applicable universally.

To meet this need, IDM (see later) which derives from TRIZ [2] was created. The method has a generic ontology and includes several key concepts like problem, partial solution, parameters and values. More precisely, our work deals with knowledge acquisition to fill IDM ontology published by [11].

4. IDM knowledge model

Ontology is the standard representation of a field or domain of an important category of objects or concepts which exists in the field or domain, showing the relation between them. IDM is an extension of TRIZ and deals with artifact evolution. It was developed to address wider and more complex and problematic situations [12]. IDM assumes that any object created by human being (artifacts) is the result of evolution guided by objective laws. As opposed to other ontologies, IDM ontology is generic and is meant to be applicable to any domain. It is also dynamic insofar as it describes the impact of changes on each other [13].

IDM basic concepts are problems, partial solutions and contradictions which include elements, parameters and values. Problems describe unsatisfactory features within system or a method while partial solutions are element of change or improvement expressing a result know in the domain or proved by experience. A problem complies with the following literal form (<*subject+verb+complement*>) and must express the essential problem. As for the partial solution, it must be the most simple possible and have the following syntax (<*Infinitive+complement*>). Parts or components of a system are called element. According to IDM model, parameters are complement nouns and can be classified into two categories: Evaluation parameters (PE) the value of which cannot be modified but are useful to evaluate the results of a design choice and Action parameters (PA) which can be modified by engineers.

5. IDM knowledge extraction

Our method is based in the use of lexico-syntactic patterns [14] to match and extract IDM knowledge from patent database. This method comes from NLP and automata theory. The method was first used by Hearst [15] to retrieve hyponyms from texts. It was then developed by Morin [16] within the framework of the acquisition of patterns for the identification of hierarchical relationships between terms.

5.1. The Semantic resources

The first step of our research consisted in the creation of semantic resources. Indeed, the issue of linguistic resources is of prime importance in the field of NLP [17]. The patent domain is very spread out as it includes many technical fields, ranging from Chemistry to Engineering. Given the syntactic complexity of patent texts and for the purpose of a fine tuning, the creation of a broad coverage lexicon was at stake. As such, a heterogeneous corpus constituted of 100 patents was constituted with the aim to observe how IDM concepts i.e. problems and partial solutions are expressed in patents and determine the regularities between the information relevant to IDM and the morpho-syntactic structure of the patent documents. Candidate markers was then collected and then tested on two corpora of 87 patents from the steel-making industry and 7 patents from vehicle transportation. The amount of patents used to make experiments may seem small at first but as Sinclair [18] says, it is best to work on small corpora for more efficiency. The results of this analysis show that the linguistic markers can be classified into two main categories which are: super-markers and polyvalent markers.

Super-markers are terms the meaning of which can be determined without knowing the context. For example, "*improve*" expresses an improvement and "*deteriorate*" conveys a problem. As for polyvalent markers, their meaning can only be determined within the context. For example, depending on the context " *increase* " may mean an improvement or a deterioration (see the example below). A sample of linguistic markers are proposed in table 1.

Example: (Decrease is employed to mean problem): *"Non-metallic inclusions decrease the workability of metals and lead to surface flaws in the rolled product"*

(Decrease is used to mean amelioration): *"Molting of the ladle nozzle due to flow velocity can be significantly decreased and breaking of the nozzle will never been caused".*

Table 1. Classified list of some linguistic markers

Grammatical category	Super-markers		Polyvalent markers
	Problem	Partial solution	
Verbs	blemish, break, bug, complicate, crack, damage deflect, etc.	ameliorate, detect enhance, ameliorate, detect, etc.	allow, change, create, decrease, differentiate, etc.
Nouns	failure, flaw, imperfection, instability, limitation, etc	amelioration, enhancement, improvement, detection, etc.	increase, generation, intensification, production, rise, etc.

5.2. IDM knowledge identification

The identification phase started with the collection of sentences in which we find propositions expressing problem and/or partial solution. At this point of the study, we notice that patent writers use a variety of syntactic structures and sometimes rather prefer general language instead of specific to express what they claim. In light of this, we started building automata to tag patent text to extract relevant knowledge with the assumption that a paragraph conveys an idea. Automata were then build to match the relevant segment at the level of the paragraph. The complete identification process can be found in our previous papers [19]; [20]. Figure below (Fig.1) shows an

example of automaton built to retrieve partial solution from the claims section. It would be a very complicated to explain in a linear way this automaton. Broadly speaking, it recognizes partial solutions existing in sentences containing "claims, revendications" at the beginning.

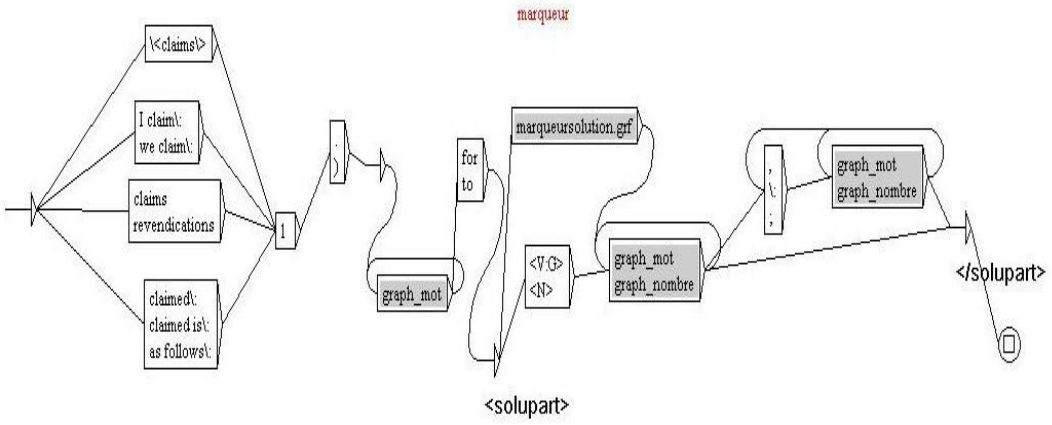

Fig.1. an example of an annotation pattern

5.3. The problem graph generation

Problem graph [21]; [19] represents problems and partial solution networks. It is constituted by all the know problem and partial solutions related to each other with their implication links:

- A problem may lead to another problem
- A problem may have a partial solution
- A partial solution can create a new problem
- A partial solution may lead to another partial solution

In order to automate the problem graph generation from patent database, we implemented a java interface to interrogate said patent stores to extract knowledge from them. For the time being, the tool uses USPTO[1] database but we are planning to extend it to other databases like Espacenet[2].

The tool provide an interface where the user can run boolean queries using operators like AND/OR. Users can then select the desired patents and start the problem graph generation (Fig. 3 and Fig 4). Let us precise that before the graph generation, user extracted results are presented to the user for validation. Figure 2 summarizes the problem graph generation. The extraction used for the time being is the corpus processor Unitex[3] used in the program as a library. Unitex is a corpus processing system, based on automata-oriented technology.

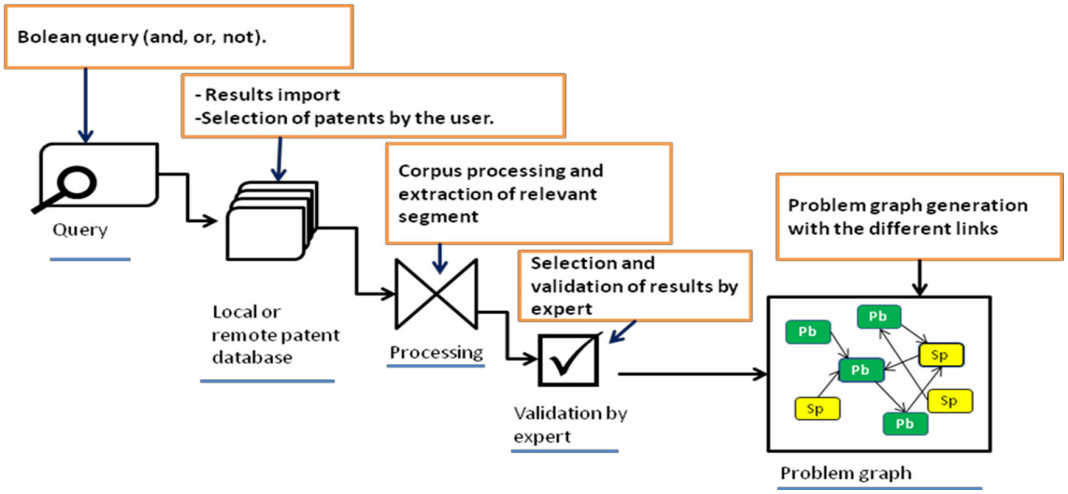

Fig. 2. Methodology of generation

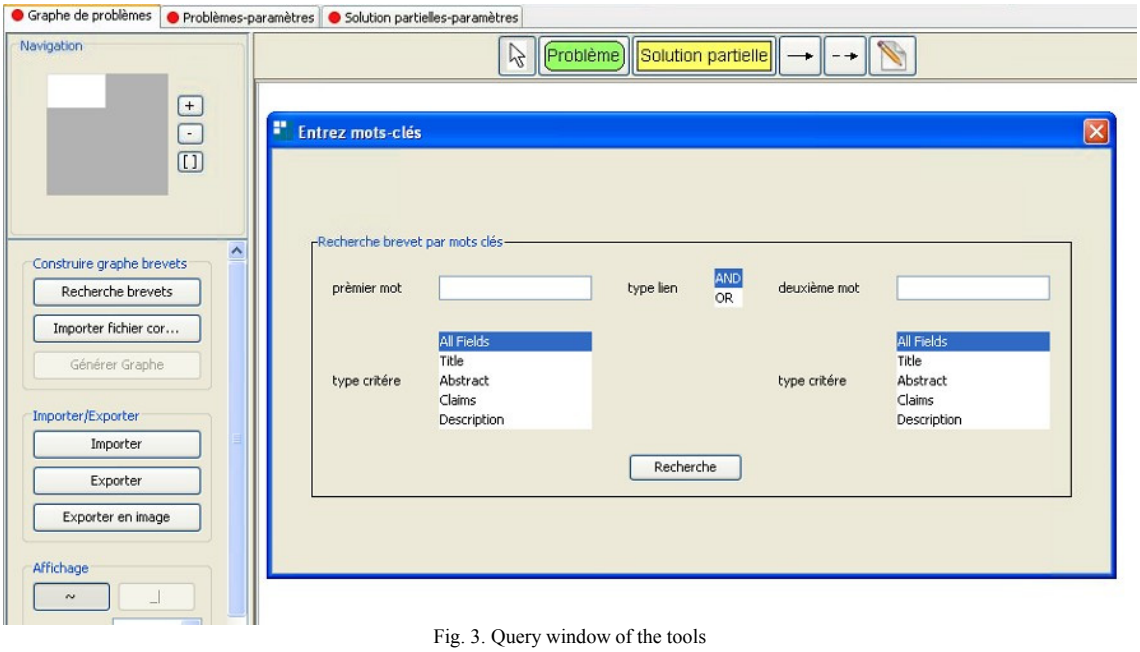

Fig. 3. Query window of the tools

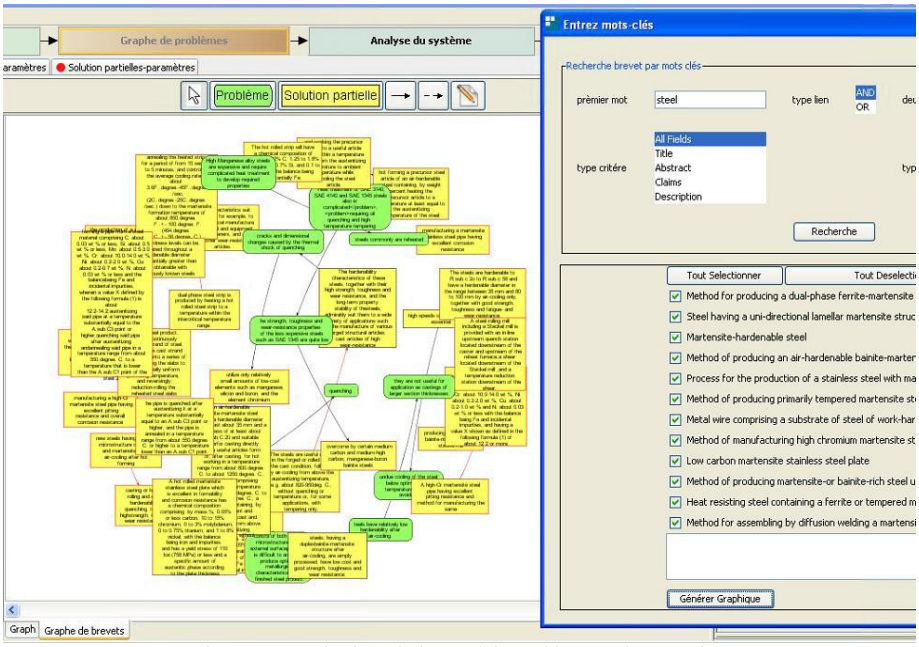

Fig. 4. Patent selection window and the problem graph generation

Figure 3 shows the automatic problem graph. It was generated from 15 patents related to steel and martensite. Partial solutions are contained in yellow boxes and problems in green boxes.

6. Results and discussion

The goal of this research is to make expert task simpler when analyzing patent documents. The collection of IDM concepts currently consumes an exuberant amount of time and money. We are in the beginning of our research and the results are not satisfactory yet. Partial solutions population is not proportional to that of problems when compared to human built graphs. We also notice orphan boxes i.e. they are not linked to any other box. We also see that segments extracted does not always respected IDM knowledge syntax. They are sometimes very long. Finally, many redundancies, both on problems and partial solutions are observed.

However, traditional performance made on results extracted from 10 patents[4] document shows a

high precision rate even though the recall rate is still very low but we are working to improve performance of the method.

Table 2. Recall and Precision

	Partial solution	Problem
Recall	43,56 %	39, 29 %
Precision	69, 40%	75,20%

7. Conclusion and perspectives

To conclude with our paper, NLP techniques represent powerful tools for patent mining. Despites existing tools in patent analysis, most of them are using specific ontologies. Our approach which advocates the use of generic linguistic indicators to extract IDM knowledge seems to be encouraging considering the results obtained. Speed of capturing and representing knowledge is incomparable (from one minute when automated to several

days of several experts for human built). The finality of our research, when turned into a software prototype, may provide design engineers with a universal tool contributing to the efficiency of their invention process. However, results are not satisfactory yet, therefore we are working to improve performances score and extend it to other patent database.

In the future, we are also planning to reduce redundancies and eliminate orphan boxes. We also need to build a more exhaustive list of markers and multiply automata and algorithms to match the most segments possible. Another perspective which must be worth mentioning is the reformulation of results that do not comply with IDM syntax for the purpose of cleaner results.

References

[1] Yeap T, Loo G.H, Pang S. Computational Patent Mapping: Intelligent Agents for Nanotechnology. In: IEEE proceedings of International Conference on MEMS, NANO and Smart systems, 2003; 274-178

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

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

[4] Norton, M., Introductory concepts in information science. 2000: Information Today, Inc.

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

[6] Cascini G, Fantechi A, Spinicci E. Natural language processing of patents and technical documentation. In: Lecture Notes in Computer Science, 3163; 2004, p.508–520.

[7] Cascini G, Russo D. Computer-aided analysis of patents and search for TRIZ contra-dictions", Int. J. Product Development. 2007 .

[8] Moehrle M, Geritz A. Developing acquisition strategies based on patent maps. In: Proceedings of the 13th International Conference on Management of Technology; Washing-ton, USA: R&D Management; 2004

[9] Yoon J, Kim K. An automated method for identifying TRIZ evolution trends from pa-tents. Experts Systems with Applications ; 2011

[10] Dewulf S.Directed variation: Variation of properties for new or improved function product DNA, a base for 'connect and develop'. In: Proceeding of the TRIZ Future Confer-ence 2006, Kortrijk, Belgium, 9-11 October 2006.

[11] Zanni-Merk C, Cavallucci D, Rousselot F. Using patents to populate inventive design ontology. In: Proceedings of the TRIZ Future conference. 2010; 52-

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

[13] Rousselot, F., Zanni, C., & Cavallucci, D. 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. 2007

[14] Jilani, I., Grabar, N., & Jaulent, M.-C. Fitting the finite-state automata platform for mining gene functions from biological scientific literature. Paper presented at the Semantic Mining in Biomedicine, Jena (Germany). 2006

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

[16] Poibeau T., Dutoit D .Automatic extraction of paraphrastic phrases from small size corpora. In Linguisticae Investigationes. John Benjamins. Amsterdam. 2009.

[17] Abeillé A. & Blache P. Grammaires et analyseurs syntaxiques. In Pierrel J.-M. Ingénierie des langues, Hermès; 2000 pp. 51-76

[18] SINCLAIR, J.M. Preface. In GHADESSY, M., HENRY, A., ROSEBERRY, R.L. Small Corpus studies and ELT: Theory and Practice. John Benjamins, Amsterdam.2000

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

[20] Souili A, Cavallucci D. Toward an automatic extraction of IDM concepts from patents, In: CIRP Design Conference 2012, Bangalore - India

[21] Cavallucci, D., Rousselot, F. & Zanni-Merk, C. Ontology for TRIZ. Proceedings of ETRIA « TRIZ Future 2009 », Timisoara, Roumanie, 4-6 novembre, Elsevier Ltd, pp. 251-260.