Improvement of automatic extraction of inventive information with patent claims structure recognition

Daria Berduygina[1] and Denis Cavallucci[1]

[1] INSA of Strasbourg, France

daria.berdyugina@insa-strasbourg.fr
denis.cavallucci@insa-strasbourg.fr

Abstract. Our recent research finding produces methods for automatic extraction of inventive information out of patents thanks to the use NLP; notably the automatic text processing. However, these methods have drawbacks due to a high amount of noise (duplicates, errors) in the output result that prevent the further use of TRIZ methodology. In the mean-time, we observed that patent claims are the most important source for inventive information. These text paragraphs have nevertheless a dual nature (combining legal and technical vocabulary) and this nature engender part of the observed noise. We postulate that taking into consideration claims hierarchical structure and its structural information can reduce the time for extraction and refine the final output quality, which is the principal aim of the paper. In this paper, we report on the methodology we have employed based on the patent claim structure recognition as a way to address our objectives

Keywords: TRIZ, Text Mining, Natural Language Processing (NLP).

1 Introduction

Today's progress is coming fast. For this reason, engineers and scientists aim to find a creative idea that can lead to invention. To help them, the researchers have developed methods that facilitate the inventive process, such as Brainstorming [1], Delphi method [2] and Synectics [3]. The TRIZ (Theory of Solving Inventive Problem) [4] began to be developed and adopted in the 1990s with the aim of making the inventive process easier and faster. This theory has earned its place among creativity techniques as an effective method which can be applied in all areas of human activity.

However, the classic TRIZ methods are difficult to understand because of the absence of formalized ontology. One more drawback is due to the fact of difficulty to perform any computation on its abstract concepts. The IDM (Inventive Design Methodology) was created by our laboratory to extend the limitations of TRIZ mentioned before. In the IDM ontology the core elements to define a problem situation and a solution consist mainly of three concepts: problems, partial solutions and parameters. We aim to extract these three concepts to automate a problem-solving process.

Patents represent an abundant source of information related to IDM. By examining patents, we can learn about technological advances over time and, more significantly, the technological challenges and solutions that have been invented by specialists and engineers in the area. Given the importance of patents as a source of information, a number of academic research and patent exploitation activities have been carried out in recent years.

Nowadays, the number of patent applications is increasing, thus it is mandatory to use adequate methods and processing tools because it can lead to better results in any patent-related activity. NLP (Natural Languages Processing) techniques related to the distinctiveness of the patent field are encouraging enhancing the quality of patent document processing. It is known that patents have extremely long and complex sentence structures with peculiar style. This feature is due to the double nature of patent text which is at the same time a legal and technical document aiming to protect the inventor and identify the boundaries of the invention.

The use of NLP analyzer for patents (with morphological, syntactic or semantic modules) is essential goal. The overall task of patent analysis is to find repeatable inventive steps that can be applied to new problems. During the last few years, our team has constructed such tool for automatic extraction of IDM-related knowledge from English-language patents. However, our tool does not take into account the hierarchical structure of patent text notably a structure of patent claims. Therefore, this is one of the reasons that our tool produces a lot of noise at the output. The adequate automatic treatment of this part of patent document could be a rich source for IDM-related knowledge, thus, because of the difficulties to process this, there is no efficient technique to extract the precious knowledge out of claims.

The double function of patent document is represented by two central parts of patent text. Firstly, a Description defines an invention, secondly, the Claims 'define the matter for which protection is sought' [5]. The description is written in similar to scientific papers style. It may contain examples that aid engineers to understand the content. The claim is the central point of the patent disclosure. It describes the essential properties of the invention. And it is subject to legal protection. This part is usually written by special patent agents for other patent experts. For this reason, the legal language is used for.

The detailed analysis of claims structure enables us to conclude that claims refer to each other. Referencing is a main feature of a legal document which aims to elucidate all aspects of invention in face of legislation. Simple test makes obvious the presence of hierarchical structure of claims (there are independent claims explaining the general characteristics of invention and dependent claims further clarifying that has already been claimed). A dependent claim may refer to one or more previous claims.

For clarification purposes, an example would be:

1. 'In a fluid transport hose comprising […]; a passage composed of […].
2. The transport hose according to claim 1, wherein said resilient body is helically formed and said channel is formed between adjacent turns of said helically formed resilient body.
3. […]
4. […]

5. The transport hose according to claim 3, wherein said flexible bag is provided under its deflated and folded condition with a coloring agent sandwiched between two outside folded surfaces of the folded flexible bag.' [6]

As it has been told, the set of claims form a hierarchical structure. This structure could be represented as directed graph:

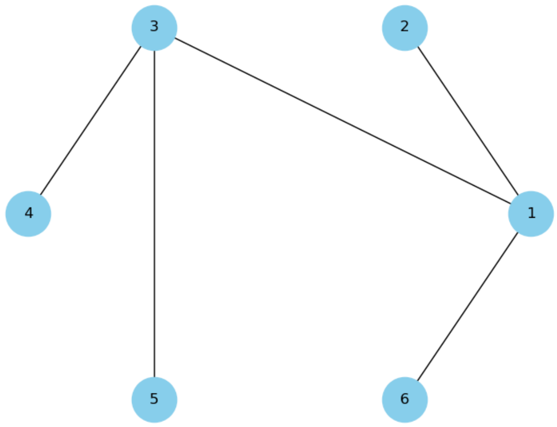

Fig. 1. Structure of claims of patent US4259553A [6] made by our claims analyzing code).

As shown on the **Fig. 1.** Structure of claims of patent US4259553A [6] made by our claims analyzing code), the directed graph represents a hierarchical structure from set of 18 patent claims with 11 dependent claims. We can observe that claims no 1–9 form a group, claims no 10–12 and 13–18 form another group. This is a simple relatively simple structure but we can also find a patent document which has more complex structure because of the number of claims that could be above more than 30.

Dealing with this kind of structure could improve the quality of extraction of our tool in terms of noise reduction. Thus, limitation of information retrieval algorithms to one group of claims could drastically reduce the noise during the process of extraction.

In this article, we propose to look through an overview of IDM-methods and its tool for automatic knowledge extraction from patent documents and literature review about patent claims structure recognition and other methods to process the claim text (State of art). Thereafter, we shortly describe the tool for the automatic extraction of the IDM-concepts from patent texts, which was recently constructed by our laboratory (Extraction tool). Then, we describe our methodology concerning the improvement of IDM-related information extraction (Methodology) and we present results of our experimental work (Evaluation and implementation).

2 State of art

In this section, we describe the IDM-methods and its tool for automatic knowledge extraction from patent documents and literature review about patent claims structure recognition and other methods to process the claim text

2.1 The Inventive Design Method

For our goal of extraction of IDM-related information, we have to define the main notions and the basic statements of these Methods.

The theory developed by Genrich Altshuller is the basis for a significant part of the work carried out by the CSIP team: the TRIZ. This theory is based on four fundamental elements [7]. The Inventive Design Method (IDM) based on TRIZ extend the limitations of the grounding theory.

The IDM describe the four necessary steps for problem-solving process. The first step consists on extraction of the information and on its organization into a graphical form comprising ‘problems’ and ‘partial solutions’. The second step involves using the first to formulate a list of contradictions according to the specific model. The third step includes the individual solving of each key contradiction. Finally, the fourth step is to select using statistics and engineers evaluation the most suitable Solution Concept [8].

For extending the limits of TRIZ and for making this theory useful for industrial innovation, the IDM proposes a practical definition of the contradiction notion [9]. According to this definition, the contradiction is ‘[…] characterized by a set of three parameters and where one of the parameters can take two possible opposite values Va and $\overline{Va}$.” [9]

Thus, it is important to distinguish the *action parameter* (AP) and the *evaluation parameter* (EP). The first one, the AP, “[…] is characterized by the fact that it has a positive effect on another parameter when its value tends to Va and that it has a negative effect on another parameter when its value tends to $\overline{Va}$ (that is, in the opposite direction)” [9].

The two other parameters in a contradiction definition are called an EP which “[…] can evolve under the influence of one or more action parameters” and which make possible to “evaluate the positive aspect of a choice made by the designer” [9].

For the clarification purposes, we add the possible formulation of the model of contradiction according to IDM postulates (**Fig. 2**).

$$\begin{array}{cccc} & & EP_1 & EP_2 \\ AP & Va & -1 & 1 \\ & \overline{Va} & 1 & -1 \end{array}$$

Fig. 2. Possible representation model of contractions (physical and technical) [9]

The understanding the way of contradiction formulation helps in the process of information retrieval and information extraction. The information that we aim to extract from patent text comprises the four elements: problems, partial solutions, APs and EPs, moreover, if it is possible, their Va and $\overline{Va}$ values.

With the help of research on IDM, we have a basic definition for the notion of *problems* as well as of *partial solutions* (how it should be represented syntactically and graphically and which information should it content).

The following schemas show the graphical representation of the problem (**Fig. 3.** Graphical representation of a problem [10]) and of partial solution (**Fig. 4.** Graphical representation of a partial solution [10]).

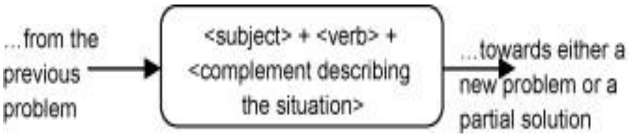
...from the previous problem
<subject> + <verb> + <complement describing the situation>
...towards either a new problem or a partial solution

Fig. 3. Graphical representation of a problem [10]

A problem (**Fig. 3.** Graphical representation of a problem [10]) "describes a situation where an obstacle prevents progress, an advance or the achievement of what has to be done" [10].

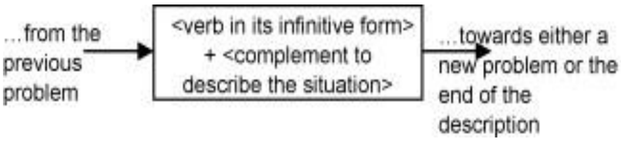
...from the previous problem
<verb in its infinitive form> + <complement to describe the situation>
...towards either a new problem or the end of the description

Fig. 4. Graphical representation of a partial solution [10]

A partial solution (**Fig. 4.** Graphical representation of a partial solution [10]) "expresses a result that is known in the domain and verified by experience" [10].

2.2 Extraction from Patent Texts

Patent texts are an important source of IDM-related information. However, this type of text presents a challenge for NLP applications because of its double nature (in the same time legal and technical) [11]. The technical knowledge issue from patent documents is rare and it is difficult to find into it innovative information of the same quality of, like for instance in scientific papers [12].

A basic inventive principle method of TRIZ and IDM relates on the fact an inventive solution could be found in another domain thanks to analogy. i.e., to find a solution for a problem it is necessary to search for analogical solved problem belonging to other domains. These analogies could be found in patent texts because this type of text represents an available inventive solution. However, searching for required patents and reading a mass of texts even by professions is a time-consuming process. For time-saving purposes, our team constructed a tool [13], that extracts out patent database the IDM-related information selected by users in English language (see the section 3). This tool can also construct problems, solutions and parameters graph which is helpful to understand the user's problem through contradictory representations and to find an appropriate solution from the same or even from another distant domain [7].

However, this tool has the drawbacks, notably, the noisy extraction. To evaluate the state of work of the tool before starting, the authors made an analysis of the quality of extraction. We took 20 patents randomly in the domain of Machine Translation, then, we made our algorithm work. Thanks to this analysis, we could note that the doubles presented in the output of extraction are abundant. This fact is shown on the **Fig. 5.** Quality of extraction**.**

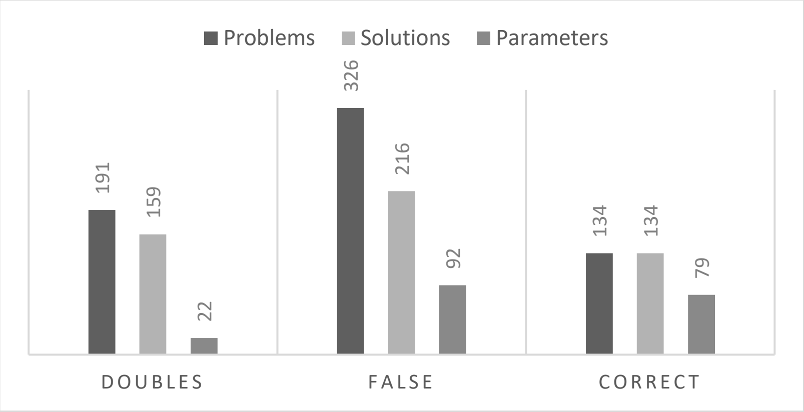

Fig. 5. Quality of extraction

As shown above, the doubles are extracted for each concept. The presence of redundant information deteriorates drastically the quality of extraction, notably the statistical scores (Table 1)

Table 1 Statistical calculation of the extraction quality for each IDM concept

	Problems	Solutions	Parameters
Precision	0,2913043478 (29%)	0,2913043478 (29%)	0,461988304 (46%)
Recall	0,881578947 (88%)	0,817073171 (81,7%)	0,975308642 (97,5%)
F-score	0,437908497 (43,7%)	0,521400778 (52%)	0,626984127 (62,6%)

Despite the fact that the recall is relatively high (yet in this analysis it is not possible to calculate a good number, which is why we consider that the tool does not miss anything), the precision is still poor. The F-score (see Table 1) seems fine for this evaluation point, yet it is only an approximation. In reality, the results can be even worse.

2.3 Patent Claims

Nowadays, there are a number of tools that can be used to recognize at least partially the patent claim hierarchical structure. These tools are created by companies or institutions for their own purposes of analysis. For example, Espacenet [14] is the European Patent Office's search engine which permits to build a tree-like representation of claim structure in their viewer. Dependent claim detection is offered by the French company Intellixir [15]. The TotalPatent [16] (the LexisNexis product) also constructs a hierarchical structure visualization of patent claims.

Moreover, the recent research in the patent information retrieval domain is focused on claims. The Information Retrieval Facility [17] is the series of conferences started in 2006, which carried off patent documents. In addition, there are several projects, financed by the European Union, that conduct research about patent searching and analyzing (PATEXPERT [18] and iPatDocs [19]).

The linguistic approaches prevail in the works of Sheremetyeva S. [20] and Shinmori et al. [21]. They aim to break the complex sentences into sub-sentences for making easy to read and understand it. This topic continues to interest the research, for instance Parapatics P. et al. [22]. The efforts to make a structural parsing are related in the works of Verberne et al. [23], D'hondt et al. [24], and Yang and Soo [25]. However, these structural parsers are focused on searching of grammatical relations in claims.

The theme of patent claims dependency dominates at the work of Hackl-Sommer R. et al. [26]. They make two strong hypotheses that led us to advance that "the occurrence of references in patent claims is a direct indicator to identify and separate independent from dependent claims" [26]. i.e., the presence in the text of a patent claim with such phrase like "according to claim 1' lead us to conclude that this claim is dependent. Inversely, the absence of this type of phrase is the index of independent claims. The second hypothesis is 'the formulaic language of patent claims allows for pattern-based analysis of the claims to identify references' [26]. i.e., the legal language used in patent claims facilitates the claims structure analysis and extraction with purposes to minimize the redundant information extraction.

3 Extraction tool

In this section, we shortly describe the tool for the automatic extraction of the IDM-concepts from patent texts, which was recently constructed, by our laboratory.

3.1 Tool description

Before we start to describe our methodology, we shortly present the toolkit for automatic extraction of IDM-related concepts.

The toolkit [13] to be improved uses linguistic and statistical methods to extract concepts related to IDM. It is based on knowledge-oriented approach (in contrast with data-oriented approach: tokenization, lemmatization, segmentation, naming entities recognizing concepts; used generally for structured data) [27]. However, the patent text represents the unstructured data, that is the reason why the knowledge-oriented approach was used.

This approach consists of an automatic extraction of the relevant linguistic patterns for each concept (problem, partial solution and parameters). Firstly, two corpora of patent texts were built (the first corpus was used to complete the list of linguistic markers and the second one for the result evaluation). The classical NLP approaches such as corpus pre-processing, stop word elimination, linguistic marker weighting, part of speech tagging and lemmatization were applied for the training corpus [28].

The linguistic markers are extracted from the patent corpora with help of the TF-IDF methods (term frequency — inverse documents frequency) [29] and the identification of a contiguous sequence of n items methods, also called n-gram identification. The last one is based on the extraction of all the word sequences from 1 to 10 tokens and on the calculation of the most frequents.

This approach conducted to analyze all the n-grams to choose the most relevant linguistic markers and to study it in the context. For example, the problems are preceded by markers such as 'it is known that…' or 'resulting in…'. And the partial solution is preceded by the phrases like 'the present invention relates to…' or '…an object of this invention is to…' [13].

After construction of the list of linguistic markers for each IDM concept and its classification, the API was built to operate this extracted data using the Python language. At the input, a user gives a patent text, then the algorithm perform the extraction based on the lists of linguistic markers.

4 Methodology

This section contains a methodology concerning the improvement of IDM-related information extraction.

4.1 Corpus analysis

In order to extract from the patent, the information about its structure, we need to find the linguistic clues that permit to establish the dependency between patent claims. The formulaic language used in patent claims construction enables to say that it exists certain amount of determined dependency constructions. Therefore, we need to obtain the list of dependency constructions.

For this purpose, we chose the patent from different technical domain of knowledge in text format from our database. The style of writing patent document is formalistic but the lexical and syntactic construction can be dissimilar that is the reason that we analyze as much domain as possible. We chose 20 patents randomly from chemistry, engineering science and linguistics.

Thereafter, we use the AntConc [30] which is an open-source corpus analysis toolkit for concordancing and text analysis. This software permits to find all sequences of searching features in corpus by entering a query word.

Due to formalistic style and language used for writing patent claims, the dependency constructions repeat in each document. The lexical and syntactic structure of phrases is independent of the domain of knowledge, i.e., same constructions are used in each domain.

For example, engineering domain [31]:

1. […]
2. The seal device as set forth in claim 1, wherein said contact surfaces have a ring configuration.

3. The seal device as set forth in claim 2, wherein said sensing member also has a ring configuration. […]

In linguistic domain [32]:

1. […]
2. The method of claim 1 wherein providing the translation output comprises […]
3. The method of claim 2 and further comprising: calculating a confidence measure for each translation output.
4. The method of claim 3 wherein calculating comprises: calculating the confidence measure […]

By the assumption that all dependent claims contain the word ‘claim’ and the number of claim/claims on which they refer to, we conduct the research using this key word. Through this analysis, we arrive to find 34 typical claims dependency clues like, for example, ‘according to claim Num., wherein,’ ‘in accordance with claim Num., wherein.’

During the analysis, we divide the claim dependency structures by the following groups:

1. foregoing term, for example, ‘referenced above,’ ‘one of the,’ ‘above-mentioned’;
2. interval, for example, ‘Numb. to Numb,’ ‘between Numb. and Numb.’;
3. filler adverbs: for example, ‘before,’ ‘previous’;
4. enumeration, for example,” according to claims 1, 3 to 5 and 10–20.’

Moreover, we can find the numerous combinations of these type of dependency clues like a foregoing term + interval, for example, ’one or more of claims 1 to 5’. To obtain the best results, we should take into account all types of dependency clues, even the combinations.

In closing our corpus analysis, we need to mention the different ways of claim numbers referencing. For this goal, the authors of patent claims use Arabic numerals as well as Roman numerals. These two types are relatively easy to process. However, we need to take into account the existence of spelled-out numerals used frequently in patent claims constructions in order to establish the hierarchical structure.

4.2 Workflow

After constructing the list of dependency structure, we could start of claims hierarchy identification and extraction. The workflow is relatively straightforward. The following steps describe all processes.

1. Segmentation. We need to detect the beginning and the end of each patent claim as well as find the section with claims in patent text.
2. Number recognition. Each claim is numbered consecutively and this number needs to be identified for each claim.
3. Classification. Each claim needs to be categorized as dependent or independent.
4. Selection. In case of dependent claims, the parent claim has to be extracted.

We will discuss each step of our workflow in subsequent sections.

Segmentation and number recognition

For identification of the beginning and the end of each claims as well as the claims section in patent text, we need to find a reliable method.

The claim section is always located at the end of patent text. The beginning of this section is identified in the same way: it usually started by “Claims”: sequence. Thus, the automatic retrieval of this section is a relatively easy task.

The individually claim segmentation is more complex. However, we introduce a number of rules permitting to identify the beginning of each one. The beginning might be represented by a new line, by a number preceded by the term “Claim” like ’Claim 1’, by a number followed by a character (blank, dot, closing parentheses or hyphens) like ’1’, ’1’, “1)” and “1 — ”. These rules have been included in our algorithm of claims segmentation.

Seldom, the beginning of a new claim is not separated by a new line, thus the claims in text appear in the block. Alternatively, we can find a blank or any other character between mentioned clues, for example, “Claim 1 3. A method…” We added in our algorithm the rules to deal with this type of structure.

Classification and selection

The next two steps in our workflow aim to classify the claims according to our definition of dependent and independent claim and to select which claim we need to extract.

To achieve this goal, we use the algorithm that allows finding the dependency structures described in **Corpus analysis** section (**4.1.**). Then, we extract the number of the parent claims. In case, when this information is not founded for a claim, we consider this claim as an independent.

5 Evaluation and implementation

The results of our experimental work is presented in this section.

5.1 Method evaluation

In order to evaluate our methodology of claim segmentation and hierarchical structure extraction, we processed random patent texts from different domains issued from our database. In particular, we are interested in patent containing an important number of claims (more than 20).

At the output of our algorithm, we can see a list of dependent and independent claims with their number as well as a directed graph represented the hierarchical structure of claims.

As the language used for claim construction is formal with limited number of formal constructions, the class of an analyzed claim is evident. For example, the claim, “2. The sealing ring of *claim 1* wherein said electrode is embedded within said body and is spaced from said exterior surface.” [33] depends on claim 1.

In our analysis we processed 13 documents. The results of output are:

- accuracy of claim segmentation = 92.3% (12 of 13 documents)
- accuracy of dependency recognition = 100% (12 of 12 documents).

Obviously, the algorithm does not process the claims that were wrongly segmented or not recognized. The failure is due to the format of the original document, after changes of code, we arrived to segment this 13th document.

5.2 Implementation

After claims hierarchical structure recognition, we make our IDM-concept analyzing tool does not take into account child claims. Before, we processed our dataset without any changes. The result is following:

- before child claims extraction:
  - 42 concepts have found, their 13 problems and 29 partial solutions,
  - processed in 81.31 seconds;
- after child claim extraction:
  - 32 concepts have found, their 13 problems, 19 partial solutions;
  - processed in 75.31 seconds.

As shown above, our method allows reducing the time of the procession of our tool as well as the quantity of partial solutions extraction. It is obvious because in claims are listed ready solutions, not problems.

The 8 of 10 dropped patterns considered as partial solutions are doubles and the other 2 was errors. If we will focus on these 19 partial solutions extracted, the 16 are correct, 3 are incorrect. The doubles are represented by 3 phrases in the output anymore.

This change allows to conclude that the work of our IDM-concept extraction tool has been improved:

- improvement of processing time = 7.38%
- noise reduction = 76.92%.

The presence of doubles is because these partial solutions are extracted from other parts of patent texts.

5.3 Discussion

The method of automatic extraction of IDM-related information, described in this article, proved our hypothesis that dual nature of patent texts makes all the document structure more complex. This complexity poses many problems for automatic concept extraction as well as for text understanding.

Extracted output containing noise represents a difficulty with analyzing the results of extraction because our global goal is to help engineers to find an appropriate solution using as much as possible sources of information. The precision of information extraction is important because in real algorithm application situation, the doubles and errors are barriers and they need to be eliminated from output as much as possible.

Despite the fact that we improved the quality of extraction, we also need to refine our approach because we processed a small amount of texts to complete the list of dependency structures.

As a future work, we need to refine the quality of the output. Firstly, we suggest dealing with hierarchical structure of patent text, notably for reducing the noise in the output of problems, which are located mostly in the Abstract and Description section.

Secondly, the resolution of co-references such as an anaphora, cataphora or split antecedents that are represented in the patent texts can also reduce the noise and make the output phrases more coherent and clear.

Thirdly, the most efficient way to improve the quality of extraction it is an implementation of a method of user's validation of the output. For example, once the user reports an extracted sentence as a noise, the algorithm record it and learn not to extract the similar sentences.

6 Conclusion

The contribution of our method according to the section 4 is important. However, the analysis of working has made in small patent corpus, i.e. we need to repeat testes with bigger corpus to find more limitations and fix it before implementation in the toolkit.

Moreover, we suggest that all patent document, not only claim section has a hierarchical structure. This treatment of this hypothesis and an it adequate implementation could reduce drastically the noise even remove it in certain cases.

References

1. J. P. Parker and L. G. Begnaud, *Developing Creative Leadership*. Libraries Unlimited, 2004.
2. N. C. Dalkey and O. Helmer-Hirschberg, "An Experimental Application of the Delphi Method to the Use of Experts,' 1962. [Online]. Available: https://www.rand.org/pubs/research_memoranda/RM727z1.html. [Accessed: 09-Apr-2019].
3. G. M. Prince, *The Practice of Creativity: A Manual for Dynamic Group Problem Solving*. Collier Books, 1972.
4. Г. Альтшуллер, *Найти идею: Введение в ТРИЗ — теорию решения изобретательских задач*. Альпина Паблишер, 2008.
5. European Patent Office, 'Guidelines for Examination in the European Patent Office.' 2018.
6. M. Tanaka and H. Saito, 'Transport hose with leak detecting structure,' US4259553A, 31-Mar-1981.
7. D. Cavallucci, Ed., *TRIZ — The Theory of Inventive Problem Solving: Current Research and Trends in French Academic Institutions*. Springer International Publishing, 2017.
8. D. Cavallucci, "From TRIZ to Inventive Design Method (IDM) : towards a formalization of Inventive Practices in R & D Departments,' 2012.
9. F. Rousselot, C. Zanni-Merk, and D. Cavallucci, "Towards a Formal Definition of Contradiction in Inventive Design", *Comput Ind*, vol. 63, no. 3, pp. 231–242, Apr. 2012.
10. D. Cavallucci, F. Rousselot, and C. Zanni, "Initial situation analysis through problem graph", *CIRP J. Manuf. Sci. Technol.*, vol. 2, no. 4, pp. 310–317, Jan. 2010.
11. Brigitte Guyot and Sylvie Normand, 'Le document brevet, un passage entre plusieurs mondes.', *Document et organisation,* Paris, 2004.

12. Bonino D., Ciaramella A., and Corno F., “Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics—ScienceDirect”. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0172219009000465. [Accessed: 10-Apr-2019].
13. A. W. M. SOUILI, ‘Contribution à la Méthode de conception inventive par l’extraction automatique de connaissances des textes de brevets d’invention’, Université de Strasbourg, École Doctorale Mathématiques, Sciences de l’Information et de l’Ingénieur Laboratoire de Génie de la Conception (LGéCo) – INSA de Strasbourg, 2015.
14. “Espacenet Patent search”, *worldwide.espacenet*. [Online]. Available: https://worldwide.espacenet.com/. [Accessed: 10-Apr-2019].
15. Questel, ‘Orbit Intellixir’, *Questel*, 2019. [Online]. Available: https://www.questel.com/software/orbit-intellixir/. [Accessed: 11-Apr-2019].
16. “Patent Research & Analysis Software | LexisNexis TotalPatent One^{TM}”, *LexisNexis® IP*. .
17. “Information Retrieval Facility”. [Online]. Available: http://www.ir-facility.org/. [Accessed: 22-Mar-2019].
18. ‘Advanced patent document processing techniques | Projects | FP6 | CORDIS | European Commission’. [Online]. Available: https://cordis.europa.eu/project/rcn/79394/factsheet/en. [Accessed: 11-Apr-2019].
19. BRUGMANN SOFTWARE GMBH, *iPatDoc*. 2013.
20. S. Sheremetyeva, “Natural language analysis of patent claims”, in *Proceedings of the ACL-2003 workshop on Patent corpus processing*—, Not Known, 2003, vol. 20, pp. 66–73.
21. Shinmori A, and Okumura M, “Aligning patent claims with detailed descriptions for readability”, *Proc. Fourth NTCIR Workshop Res. Inf. Retr. Autom. Text Summ. Quest. Answering Natl. Inst. Inform. Jpn.*, vol. 12, no. 3, pp. 111–128, Jul. 2005.
22. Parapatics P. and Dittenbach M, “Patent Claim Decomposition for Improved Information Extraction”, *ResearchGate*, 2011. [Online]. Available: https://www.researchgate.net/publication/226411853_Patent_Claim_Decomposition_for_Improved_Information_Extraction. [Accessed: 11-Apr-2019].
23. Verberne S., D’hondt E., and Oostdijk N., “Quantifying the challenges in parsing patent claims”, *ResearchGate*, 2010. [Online]. Available: https://www.researchgate.net/publication/228739952_Quantifying_the_challenges_in_parsing_patent_claims. [Accessed: 11-Apr-2019].
24. E. D’hondt, S. Verberne, W. Alink, and R. Cornacchia, “Combining document representations for prior-art retrieval”, p. 9.
25. S.-Y. Yang and V.-W. Soo, “Extract conceptual graphs from plain texts in patent claims”, *Eng. Appl. Artif. Intell.*, vol. 25, no. 4, pp. 874–887, Jun. 2012.
26. R. Hackl-Sommer and M. Schwantner, “Patent Claim Structure Recognition”, *Archives of Data Science, Series A (Online First)*, 2017. [Online]. Available: https://publikationen.bibliothek.kit.edu/1000069936. [Accessed: 11-Apr-2019].
27. A. Souili and D. Cavallucci, “Automated Extraction of Knowledge Useful to Populate Inventive Design Ontology from Patents”, in *TRIZ — The Theory of Inventive Problem Solving*, D. Cavallucci, Ed. Cham: Springer International Publishing, 2017, pp. 43–62.
28. A. Souili, D. Cavallucci, and F. Rousselot, “A lexico-syntactic Pattern Matching Method to Extract Idm—Triz Knowledge from On-line Patent Databases”, *Procedia Eng.*, vol. 131, 2015.
29. G. Salton and C. S. Yang, “On the Specification of Term Values in Automatic Indexing”, Jun. 1973.
30. Anthony, L, *AntConc*. Tokyo, Japan: Waseda University, 2019.

31. B. E. Bennett, “Seals with integrated leak progression detection capability”, US7316154B1, 08-Jan-2008.
32. M. Zhou, J.-X. Huang, C. N. (Tom) Huang, and W. Wang, “Example based machine translation system”, US7353165B2, 01-Apr-2008.
33. M. K. Sunkara, “Sealing ring with electrochemical sensing electrode”, US5865971A, 02-Feb-1999.