Automatic extraction of potentially contradictory parameters from specific field patent texts

Daria Berdyugina, Denis Cavallucci

Automatic extraction of potentially contradictory parameters from specific field patent texts

Daria Berdyugina[1] and Denis Cavallucci[1]

[1] ICUBE/CSIP, INSA de Strasbourg, 24 Boulevard de la Victoire,
67084 Strasbourg, France
dberdyugina@etu.unistra.fr
denis.cavallucci@insa-strasbourg.fr

Abstract. Nowadays, Altshuller contradiction matrix is used by many TRIZ practitioners, especially by beginners, thanks to its simplicity. However, establishing the link between user's specific problems issued from their experience in their domain of knowledge makes the use of the matrix often difficult. Applying specific terms of domain to formalized language of TRIZ tools necessitate an expertise that users often don't have time to build. Our previous finding based on Natural Languages Processing (NLP) tools and techniques, made possible to process a corpus of patents from a given field and thanks to Topic Modelling technique we achieved to link the technical parameters extracted out of patents to their context representation on a vector space in the text. However, this approach is not pertinent to identify the contradictory relations between extracted parameters. For this reason, we applied antonyms identification technique in order to better process the relations of oppositions between extracted parameters. The goal of this research it to extract automatically potential contradictions and set them up in an Altshuller-like matrix. Such an approach could facilitate the application of this famous TRIZ tool for practical user's problems. Moreover, setting up the matrix for patents of the new domain of knowledge could help to construct easily the state of art for these types of domain and keep the users informed without spending a lot of time and human resources for reading and analyzing large quantities of texts appearing continuously in each domains.

Keywords: NLP, Altshuller Matrix, Text Mining.

1 Introduction

Patents as a source of inventive information attracted attention of researchers and industry for a long time. Their application represents a huge area of scientific and practical research. They cover almost every domain of knowledge in industry nowadays that is the reason why a lot of TRIZ-related approaches use patents as the basis of its functioning.

The main purpose of patent institutions is to register inventions using their technical descriptions and consequently associating current limitations in a domain and what an

applicant is claiming. The legal nature of the patent text is manifested equally in peculiar style of writing of a document, i.e. long and complex sentences so as the presence of a lot of repetitions of the same information in order to better precise its borders. In the length of a document that could sometimes exceed one hundred pages for certain inventions with an often purposefully confusing structure. Precisely for this reason, for non-experienced readers, especially for people unfamiliar with the jurisprudence, the reading and the understanding the patent context may present an obstacle to use it in their work, especially in inventive problem-solving process.

On the other hand, the patent text contains a huge amount of peer-reviewed technical information, such as certain features of newly invented object or system, the state of the art of a given field and all other detailed information about an invention. These details represent an enormous interest for engineers, scientists and industry. However, despite the fact that the patent readers are usually familiar with technical information expressed in patent text and they are using to read the documents written in technical language, the double nature of the patent document presents a real barrier. The reading and understanding of such text demand a lot of human resources and is time-consuming. Hence, thanks to the development of computer science, especially Natural Languages Processing (NLP), we could exploit the information automatically and save time and resources.

In the context of theory of the resolution of inventive-related tasks (TRIZ) [1], patent texts present an object of interest because of the fact that they contain a huge amount of inventive information. The founder of TRIZ, the Soviet Engineer G S. Altshuller, analyzed manually about 40,000 of patents. This analysis permitted him to notice that all inventions obey to the certain laws or evolution and arrive to the conclusion that an inventive process can be formalized.

The analysis of huge amount of patent texts allowed G. S. Altshuller to create a famous tool which called Contradiction Matrix (CM) [2]. Nowadays, with the development of TRIZ, more experienced TRIZ practitioners prefer to use more complex tools such as ARIZ85C [3] or Vepoles [4]. However, despite the fact that the CM was created in 1969 and despite the attempts to change this tool, the CM is still popular among TRIZ users thanks to its accessibility and simplicity.

Nevertheless, with the development of modern science and technology and with the emerging of a lot of new domains of knowledge, the main terminology and vocabulary used in CM are becoming obsolete making this tool out of date. Moreover, for the specialists of a given field, it is often difficult to link their specific vocabulary to the TRIZ terms. This fact creates an obstacle for the use of CM.

Henceforth, thanks to the modern NLP and computing technique, the automatic extraction of TRIZ-related information is becoming more possible. With the exploitation of linguistic textual markers, we achieved to extract the main subjects discussed in patents thanks to the Topic Modelling approach [5]. Based on distributional hypothesis [6], which claims that linguistic items with similar distributions have similar meanings and that the semantic meaning of a word is characterized by its context [7] (the theoretical basis and origins of this hypothesis are discussed in [8]), we may analyse statistically a huge amount of textual data and exploit not only linguistic features but also

statistical representation of tokens[1] [9, p. 111]. This hypothesis allows to make a conclusion that in domain-specific text, the most repeated word are the terms of this domain. Consequentially, we exploit these approaches in order to automatically extract the inventive information.

Despite the strength of TRIZ, the absence of formalized ontology that disables the possibility of performing the computation on abstract parameters, our laboratory elaborated the Inventive Design Method (IDM) in order to extend this limitation of the ground theory [10]. Based on TRIZ, IDM permits to easily perform the problem-solving process. According to IDM, three main concepts represent the solid base for this process: parameters, partial solutions and problems [11].

The goal of our research consists of automatic extraction of potentially contradictory parameters from the domain-specific corpus thanks to the NLP techniques. Thanks our recent research, we elaborated the tool that permits us to extract three main concepts automatically out of patent texts. In the present article, we discuss the method that allows to represent the context space of extracted parameters and compute the score of its similarity in order to extract the contradictory relations.

In the chapter 2, we describe the state of art, including IDM, antonyms and used NLP techniques. In the chapter 3, we describe our applied methodology. The chapter 4 is dedicated to the result and its evaluation. In the final chapter 5, we present a conclusion.

2 State of Art

In order to better precise the methodology of the present research, in this chapter we discuss the IDM concepts and NLP techniques used for achieving our goal.

2.1 Inventive Design Method

In the present research, we aim to extract IDM-related information out of the domain-specific corpus. The object of our particular interest is parameters because they represent the elements of contradiction. For the clarification purpose, we discuss above the main concepts: problems, partial solution, parameters and contradiction.

In the present research we aim to mine the contradictory relations between parameters. However, for the understanding of the complete process, we need to describe all essential concepts of IDM.

According to IDM, the problem-solving process comprises four steps [11]:

1. Extraction of inventive information, notably the problems and the partial solutions;
2. Formulation of contradictions;
3. Solving of key contradictions;
4. Choice of the most pertinent solution.

[1] Lexical or category unit

A problem represents the situation 'where an obstacle prevents progress, an advance or the achievement of what has to be done' [12]. A partial solution 'expresses a result that is known in the domain and verified by experience' [12].

The problems and partial solutions need to be extracted in order to perform the first step of problem-solving process. The second step comprises the contradiction formulation that may be done based on problems or parameters. For this step, it is important to give a definition of these concepts.

According to our ontology, we distinguish action parameters (AP) and evaluation parameter (EP). 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)' [13]. The EP '[…] can evolve under the influence of one or more action parameters' and makes possible to 'evaluate the positive aspect of a choice made by the designer' [13].

Thus, in order to facilitate the use of TRIZ in industrial innovations, IDM gives the definition of contradiction notion. According to IDM, a contradiction is '[…] characterized by a set of three parameters and where one of the parameters can take two possible opposite values Va and $\overline{Va}$.' [13].

For the clarification purpose, we provide the graphical representation of contradiction notion below [13].

$$\mathrm{AP}\frac{Va}{\overline{Va}}\begin{pmatrix} EP_1 & EP_2 \\ -1 & 1 \\ 1 & -1 \end{pmatrix} \qquad (1)$$

In the present research, we are interested in EP extraction and in contradiction relation detection between extracted parameters. Henceforth, we use our tool that permits to extract the parameters automatically. Then, we detect the antonym relation between them using NLP techniques discussed below.

2.2 Antonyms Classification

According to The Oxford Dictionary of English Grammar, an antonym is defined as 'a word in opposite meaning to another' [14, p. 29]. But as soon as an antonym represents more complex linguistic phenomena, we need to describe it in more detailed way.

An antonym is more than just linguistic term, this is also related with psychological aspect because, according to its definition, antonym appears only in the pair of words and could not have the opposite meaning without another word. Hence, the opposition could not exist without human knowledge about the object of opposition. Moreover, a word may have more than one opposite word.

The semantic research [15] distinguishes some basic characteristic of opposites:

- Binarity manifested in the occurrence of opposites as a lexical pair:
- Inherentness expressed in the relationship may be presumed implicitly;
- Patency presents the quality of how obvious a pair is.

According to the nature of opposite relationship, there are three groups of antonyms: gradable, complementary and relational antonyms [16].

Gradable antonyms are known as the most represented class of antonyms. They express the pair of words with opposite meaning where these two meanings lie in a continuous spectrum [17]. For example, the weight could be *heavy* or *light*, hence these two words appear in the opposite ends of the spectrum, so there is a gradient of opposition, that is why this type of antonym is called 'gradable'. The other examples of such pairs are: big/small, old/young, dark/light, etc.

Complementary antonyms, also called binary or contradictory antonyms [18], represent a pair of words where two meanings does not lie in a continuous spectrum. For example, the pair of words vacant and occupied does not have a continuous spectrum between them, however, they are opposite in meaning and that is why they are complementary antonyms. The other examples of complementary antonyms: entrance/exit, exhale/inhale, mortal/immortal.

Relational antonyms could be defined as a pair of words that refer to the opposition from the opposite point of view [19]. For example, semantically, there is no opposition between *pupil* and *teacher* but we may oppose them in certain contexts. This fact allows us to call this type of antonym as relational since they exist only in pairs depending on the context. The other examples: parent/child, come/go, husband/wife, etc.

In the point of view of TRIZ contradiction notion, the EPs between which we aim to identify opposite relation, according to the classification cited above, represent the relational antonyms since there is no opposition in language between surface and pressure but they form a contradiction. That is a reason why in order to identify the opposite relation, it is necessary to set a context representation of every extracted parameter.

The techniques of opposite relation identification in the context are described in the next section.

2.3 Topic Modelling approach and antonyms identification

Patent mapping technic exists for a long time and is widely used for graphical representation of patent content. This is an important task because of the large number of patents is publishing daily henceforth it is difficult to track all of them manually. The graphical representation of patent content presents an accessible and comprehensible way to display all main features of patent content and then to choose a field to focus on.

The examples of use of patent mapping could be found in [20]–[22]. However, the most common way to establish a patent map is based on the structured data such as dates, citations or assignees. Hence, all this information may be analysed using traditional bibliometric techniques [23]. Thereafter, the text-mining techniques are equally suitable not only for terms-extraction task, but also for extraction of key information [24]. The technique of automatic text summarization allows to extract an essential information out of patent text and present it into the form of short text that is easily understandable than an input text [25].

However, in the context of our goal, we aim to extract the relation between textual elements out of unstructured data. Moreover, we are focusing on one-domain text collection that is the reason why we cannot predict the vocabulary used in text and we need to turn for the computation techniques, notably the unsupervised learning techniques.

The main advantage of such method consists in the fact that it generates an output without any information about environment. The formal structure of such algorithms allows to find the pertinent patterns. Conversely, the supervised or reinforced learning techniques demand the annotated input data to get an example of that should be given at the output.

For statistical corpus analysis, the one of the most suitable techniques based on unsupervised learning logic is Latent Dirichlet Allocation (LDA) [26]. Briefly, this technique could be described as follows: 'The LDA model assumes that the words of each document arise from a mixture of topics, each of which is a distribution over the vocabulary' [5]. This tool is based on the distributional hypothesis described above (1). I.e., in sample text talking about, for example, the cats and dogs, the words like 'milk', 'fish' and 'meow' would appear together with cats and the words 'flesh', 'bark' and 'bone' would appear near dogs. That is a topic representation of a text and LDA allow us to establish tis representation in order not only to get the set of domain terms, but also to achieve to form a context space. However, since that Topic modeling technique represents the bag-of words approach, all syntaxic relations between words are lost after the processing. This fact could impact the precision of contradiction identification.

Moreover, according to existing methodology, topic models ' … can extract surprisingly interpretable and useful structure without any explicit "understanding" of the language by computers' [5]. For a detailed explanation on the algorithm refer to [27] and for an evaluation analyzing scientific publications refer to [28].

A lot of research is focused nowadays on the task of antonym identification. This semantic relation of opposition represents a powerful index for many language-based approaches of information extraction. The most common application for antonym identification is opinion mining [29], [30]. The Deep Learning techniques could be used to identify the antonyms [31]. However, the simple language contradiction identification is not the object of our interest. We are focusing of the extraction of opposite relations in the point of view of TRIZ.

Hence, the computation of semantic similarity is one of the most used techniques for calculation how close two words or two texts are. There are a lot of approaches of similarity computation based on knowledge-based approach [32]. In the point of view of text similarity, the technique of lexical matching is applied in [33].

The similarity metrics are another way to compute the similarity between words [34]. They are based on computation methods and in order to perform any calculation, the corpus should be represented as vector space.

The suitable distance metric for our approach is Hellinger distance since it proposes to compute the similarity in vector space. In probability and statistic, this metric is used to calculate the similarity between two probability distributions. This distance is defined by Hellinger integral, described in [35]. For more information, refer to [36].

The following equation is used in Topic Modeling algorithm to calculation the Hellinger distance. The P and Q represent two probability measure in continuous and dP and dQ is a brief form of writing of Radon-Nikodym derivatives of P and Q [36].

$$H^2(P,Q) = \frac{1}{2}\int(\sqrt{dP} - \sqrt{dQ})^2 \qquad (2)$$

3 Methodology

In this chapter, we describe the method for contradiction identification based on Topic Modelling technic and the distance metric.

3.1 Corpus presentation and extraction tool

As it is described above (1), the goal of the present research consists of extraction of contradictory relations out of the domain-specific patent corpus. This extraction is possible thanks to the distributional hypothesis and NLP techniques such as Topic Modelling and similarity distance computation.

First of all, for clarification purposes, we discuss the tool for parameters extraction. This tool elaborated recently in our laboratory is based on linguistic and statistical approach which is suitable for information extraction out of unstructured data [10]. Our tool is described briefly in [37]. The tool permits to extract three main concepts of IDM out of patent text. In the present research we are interested in parameters extraction. In order to perform this extraction, the tool use the dictionary of markers that, according to the previous research [38], are used to identify TRIZ parameters. The dictionary includes the list of the terms used to express the parameter notion in the patents. This list is obtained by statistical analysis of patent texts previously in [39].

In the framework of the present paper, we perform all workflow on domain-restricted corpus comprising four patents from door latch mechanism field. All these patent texts are accessible via Google Patents. The corpus consists of 379,898 words and the texts are written in English language.

In the illustrative purpose, we provide the scheme describing the applied workflow below (see **Fig. 1**).

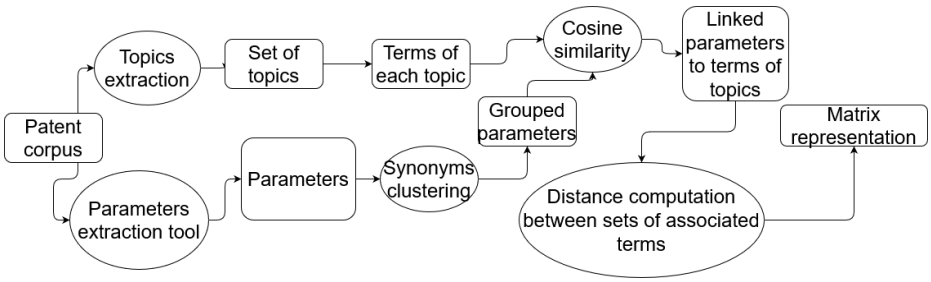

Fig. 1. Workflow representation

3.2 Corpus preprocessing

For the present research, we need to apply classical NLP reprocessing pipeline in order to get a suitable input for the LDA model, i.e. the first step consists of transformation of the text into the vector representation (Word Embeddings [40]).

However, before performing the transformation, it is necessary to clean the corpus. Firstly, we transform the text of the input in lowercase. Then, thanks to the special tool

for statistical corpus analysis (Antconc [41]), we extract the most common used words into the patent texts and add them to the classical English-language stop words list. This step is required to eliminate from the result the unnecessary words.

The second step consists of removing the punctuation and to concatenate the polylexical terms and collocations together. Thus, we search the words with the highest probability to appear together in text. We are interested in bi- and trigrams identification. After identifying such collocations, we replace the space for underscore. By performing this step, we make our algorithm recognize not only simple terms but equally the multiword expressions.

The final step includes the lemmatization and the part of speech identification. This step allows to exclude from the corpus space the verbs and adverbs that do not present in the string corresponding to parameters. By applying LDA, as it is discussed in 2.3, we lose all syntaxic relations between extracted terms. The identification of which action (verbs) and the description of action (adverbs) are removed from the vector space in order to better contextualize the parameters (which represent the noun phrases).

3.3 Topic Modelling and similarity Computation

With the development of the domain of statistical corpus analysis, there are a lot of tools and frameworks allowing to mine the topics. For example, BigARMT[2], Stanford Topic Modelling Toolkit[3] and topic-model [4] R package.

In the context of our project, we use Gensim Framework [42] because this framework permits to perform the basic NLP tasks and is suitable for Python programming language.

As an input, the LDA model takes the vector representation of the corpus. We transform our cleaned corpus into the vectors thanks to doc2vec[5] technique. After all manipulations, we achieve to establish the topic representation of the corpus. The most suitable number of topics is chosen after the calculation of coherence score for the models having different number of topics (from 2 to 40).

The next step consists on the linking the extracted parameters with terms of topics. Parameters are extracted thanks to the tool described in 3.1. The tool focuses on the extraction of general TRIZ parameters and the step of contextualizing allows to exclude the noise and to add more domain-related information in the result. Every topic set consists on 10 of the most representative terms. The number of parameters depends on the extraction tool. We compute the similarity between all extracted parameters and topic terms in order to identify all contexts of every word, even hidden.

The final step consists of calculation of the Hellinger distance between of the sets of associated topic terms. The Hellinger distance metric gives an output in the range [0.1] for two probability distributions, with values closer to 0 meaning they are more similar.

---

2 Available at https://github.com/bigartm/bigartm

3 Available at https://downloads.cs.stanford.edu/nlp/software/tmt/tmt-0.4/

4 Available for https://cran.r-project.org/web/packages/topicmodels/index.html

5 https://radimrehurek.com/gensim/models/doc2vec.html

Thus, we could estimate how far the two parameters are in the point of view of contextual semantic space in order to identify the opposite relations between them.

4 Case study, results and evaluation

In order to validate our approach, thanks to one of our industrial partner ArcelorMittal, we've been using a series of 14 patents from the domain of door latch mechanism for automotive industry. The parameters extraction tool extracted from the corpus 237 parameters. Thanks to our method, we reduced this number to 15. This quantity of parameters is suitable for creation of small domain-restricted CM based on antonym recognition.

Hence, we could form a CM comprising 105 potential contradictions and then calculate the distance between those candidates. For finding the most interesting parts, we calculate the average score and we highlight the parts that exceed this score.

We use human extraction of contradiction as an example. The main difficulty is the interpretation of the result because the algorithm extracts the words from the corpus without any modification and that is the reason why the comparison is hard to perform.

The human extraction comprises 18 contradictions and thanks to our method we highlight 56 candidates that have the Hellinger distance score higher than average value.

Precision	0.16
Recall	0.5
F-score	0.24

Table 1. Statistic of extraction

The **Table 1** shows the statistical result of the extraction. The F-score is relatively low but in terms of recall, the result is good.

The fable result encourages us to search for another indexes for contradiction identification in order to not only filter the parameters, but to find the most pertinent approach for our goal.

However, our method has been coded in an API allowing to explore domain-specific collection of patent texts in order to establish the matrix representation of opposite parameters. The work is still in progress.

5 Conclusion

In this article, we describe the methodology to opposite relation identification between parameters in a domain-specific corpus of patent texts. The present technique, thanks to the identification of additional semantic information, reduces the quantity of extracted parameters and prioritize only the domain-related parameters. The described method allowed us to create an API that may identify the potential contradictorily relation automatically between filtered parameters.

However, as a future work, we aim to validate the quality of extraction and if it is necessary, to identify linguistic and discursive markers permitting to better distinguish the domain-specific parameters and contradictory relations since our extraction tool focuses on the extraction of general parameters. That is the reason why the precision is not adequate even after filtering and contextualization steps.

For instance, we are working on the dressing of the 'borders' of CM, but the choice of the content of cells remain uncertain at this time. Hence, we need to create a methodology to link the elements of the matrix to the useful TRIZ-related information such as, for example, inventive principles or partial solutions. In a longer perspective, since not all necessary TRIZ-related information resides in patent texts, we also intend to exploit other sources of inventive information, such as scientific papers to complete empty spaces with other sources extraction.

References

[1] G. Altshuller, G. Al′tov, and H. Altov, *And Suddenly the Inventor Appeared: TRIZ, the Theory of Inventive Problem Solving*. Technical Innovation Center, Inc., 1996.

[2] G. Altshuller, *40 Principles: TRIZ Keys to Innovation*. Technical Innovation Center, Inc., 2002.

[3] "ARIZ : The Algorithm for Inventive Problem Solving," *The Triz Journal*, Apr. 08, 1998. https://triz-journal.com/ariz-algorithm-inventive-problem-solving/ (accessed Apr. 08, 2021).

[4] В. Петров, *Структурный анализ систем. Вепольный анализ. ТРИЗ*. Litres, 2019.

[5] D. M. Blei and J. D. Lafferty, "A correlated topic model of Science," *Ann. Appl. Stat.*, vol. 1, no. 1, pp. 17–35, Jun. 2007, doi: 10.1214/07-AOAS114.

[6] Z. S. Harris, "Distributional Structure," *WORD*, vol. 10, no. 2–3, pp. 146–162, Aug. 1954, doi: 10.1080/00437956.1954.11659520.

[7] J. R. Firth and F. R. Palmer, *Selected papers of J R Firth 1952-1959*. London: Longmans, 1956.

[8] M. Sahlgren, "The distributional hypothesis," pp. 33–53, 2008.

[9] A. V. Aho, M. S. Lam, R. Sethi, and J. D. Ullman, *Compilers: principles, techniques, and tools.* Harlow, Essex: Pearson, 2014. Accessed: Apr. 23, 2021. [Online]. Available: http://www.vlebooks.com/vleweb/product/openreader?id=LeedsUni&isbn=9781292037233&uid=none

[10] 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. doi: 10.1007/978-3-319-56593-4_2.

[11] D. Cavallucci and N. Khomenko, "From TRIZ to OTSM-TRIZ: addressing complexity challenges in inventive design," *International Journal of Product Development*, vol. 4, no. 1–2, pp. 4–21, Dec. 2006, doi: 10.1504/IJPD.2007.011530.

[12] D. Cavallucci, F. Rousselot, and C. Zanni, "Initial situation analysis through problem graph," *CIRP Journal of Manufacturing Science and Technology*, vol. 2, no. 4, pp. 310–317, Jan. 2010, doi: 10.1016/j.cirpj.2010.07.004.

[13] F. Rousselot, C. Zanni-Merk, and D. Cavallucci, “Towards a Formal Definition of Contradiction in Inventive Design,” *Computers in Industry*, vol. 63, pp. 231–242, Feb. 2012, doi: 10.1016/j.compind.2012.01.001.

[14] S. Chalker and E. S. C. Weiner, *The Oxford Dictionary of English Grammar*. BCA, 1998.

[15] A. Cruse, *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford University Press UK, 2011.

[16] D. A. Cruse, “Three classes of antonym in English,” *Lingua*, vol. 38, no. 3, pp. 281–292, Jan. 1976, doi: 10.1016/0024-3841(76)90015-2.

[17] “English Gradable Antonyms: Implicit Comparison and Explicit Comparison--《Journal of Zhejiang University(Humanities and Social Sciences)》2004年04期.” https://en.cnki.com.cn/Article_en/CJFDTotal-ZJDX200404016.htm (accessed Apr. 23, 2021).

[18] B. Aarts, S. Chalker, and E. Weiner, *The Oxford Dictionary of English Grammar*, Second Edition. Oxford, New York: Oxford University Press, 2014.

[19] S. Jones, M. L. Murphy, C. Paradis, and C. Willners, *Antonyms in English: Construals, Constructions and Canonicity*. Cambridge University Press, 2012.

[20] B. Yoon and Y. Park, “A text-mining-based patent network: Analytical tool for high-technology trend,” *The Journal of High Technology Management Research*, vol. 15, no. 1, pp. 37–50, Feb. 2004, doi: 10.1016/j.hitech.2003.09.003.

[21] S. Lee, B. Yoon, and Y. Park, “An approach to discovering new technology opportunities: Keyword-based patent map approach,” *Technovation*, vol. 29, no. 6, pp. 481–497, Jun. 2009, doi: 10.1016/j.technovation.2008.10.006.

[22] Y. G. Kim, J. H. Suh, and S. C. Park, “Visualization of patent analysis for emerging technology,” *Expert Systems with Applications*, vol. 34, no. 3, pp. 1804–1812, Apr. 2008, doi: 10.1016/j.eswa.2007.01.033.

[23] D. Archibugi and M. Planta, “Measuring technological change through patents and innovation surveys,” *Technovation*, vol. 16, no. 9, pp. 451–519, Sep. 1996, doi: 10.1016/0166-4972(96)00031-4.

[24] Y.-H. Tseng, Y.-M. Wang, Y.-I. Lin, C.-J. Lin, and D.-W. Juang, “Patent surrogate extraction and evaluation in the context of patent mapping,” *Journal of Information Science*, vol. 33, no. 6, pp. 718–736, Dec. 2007, doi: 10.1177/0165551507077406.

[25] B. Yoon and R. Phaal, “Structuring technological information for technology roadmapping: data mining approach,” *Technology Analysis & Strategic Management*, vol. 25, no. 9, pp. 1119–1137, Oct. 2013, doi: 10.1080/09537325.2013.832744.

[26] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[27] A. N. Srivastava and M. Sahami, *Text Mining: Classification, Clustering, and Applications*. CRC Press, 2009.

[28] Yau, A. Porter, N. Newman, and A. Suominen, “Clustering scientific documents with topic modeling,” *Scientometrics*, vol. 100, pp. 767–786, Sep. 2014, doi: 10.1007/s11192-014-1321-8.

[29] G. Fei, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, “A Dictionary-Based Approach to Identifying Aspects Implied by Adjectives for Opinion Mining,” p. 10.

[30] D. Lee, O.-R. Jeong, and S. Lee, “Opinion mining of customer feedback data on the web,” in *Proceedings of the 2nd international conference on Ubiquitous information management and communication*, New York, NY, USA, Jan. 2008, pp. 230–235. doi: 10.1145/1352793.1352842.

[31] S. Rajana, C. Callison-Burch, M. Apidianaki, and V. Shwartz, “Learning Antonyms with Paraphrases and a Morphology-Aware Neural Network,” in *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, Vancouver, Canada, 2017, pp. 12–21. doi: 10.18653/v1/S17-1002.

[32] R. Mihalcea, “Corpus-based and Knowledge-based Measures of Text Semantic Similarity,” p. 6.

[33] C. Corley and R. Mihalcea, “Measuring the semantic similarity of texts,” in *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment - EMSEE ’05*, Ann Arbor, Michigan, 2005, pp. 13–18. doi: 10.3115/1631862.1631865.

[34] G. Pirró and N. Seco, “Design, Implementation and Evaluation of a New Semantic Similarity Metric Combining Features and Intrinsic Information Content,” in *On the Move to Meaningful Internet Systems: OTM 2008*, Berlin, Heidelberg, 2008, pp. 1271–1288. doi: 10.1007/978-3-540-88873-4_25.

[35] E. Hellinger, “Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen.,” *Journal für die reine und angewandte Mathematik*, vol. 1909, no. 136, pp. 210–271, Jul. 1909, doi: 10.1515/crll.1909.136.210.

[36] “Hellinger distance - Encyclopedia of Mathematics.” https://encyclopediaofmath.org/index.php?title=Hellinger_distance (accessed Apr. 23, 2021).

[37] D. Berdyugina and D. Cavallucci, “Setting Up Context-Sensitive Real-Time Contradiction Matrix of a Given Field Using Unstructured Texts of Patent Contents and Natural Language Processing,” 2020, pp. 30–39. doi: 10.1007/978-3-030-61295-5_3.

[38] A. Souili, D. Cavallucci, and F. Rousselot, “A lexico-syntactic pattern matching method to extract IDM- TRIZ knowledge from on-line patent databases,” *Procedia Engineering*, vol. 131, pp. 418–425, 2015, doi: 10.1016/j.proeng.2015.12.437.

[39] 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. Accessed: May 31, 2020. [Online]. Available: https://scanr.enseignementsup-recherche.gouv.fr/publication/these2015STRAD026

[40] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” *arXiv:1310.4546 [cs, stat]*, Oct. 2013, Accessed: May 31, 2020. [Online]. Available: http://arxiv.org/abs/1310.4546

[41] L. Anthony, *AntConc*. Tokyo, Japan: Waseda University., 2019. Accessed: May 31, 2020. [Online]. Available: https://www.laurenceanthony.net/software/antconc/

[42] R. Řehůřek and P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta, 2010, pp. 45–50.