Concept Extraction Based on Semantic Models Using Big Amount of Patents and Scientific Publications Data

Kaliteevskii Vasilii, Deder Arthur, Peric Nemanja, Chechurin Leonid

DOI: 10.1007/978-3-030-86614-3_11

Concept extraction based on semantic models using big amount of patents and scientific publications data

Abstract. Formalisation of heuristic methods for supporting the conceptual design stage of product and technology development has been extensively evolved in industry during the last half of the century and gradually more formally appears in academic context nowadays. Due to the considerable interest from the Industry and the Academia, heuristic approaches such as TRIZ have been strongly developed over the past decades. Thus, TRIZ evolved from a set of empirical inventive principles into a considerably formal approach including techniques for modeling technical problems with the possibility of further overcoming them using formal methods. Moreover, during the last decades, TRIZ has been extensively digitized. Several generations of software have appeared that facilitate the use of inventive methods (Goldfire, Invention Machine). From the trend of digitalisation and the success of machine driven processes, it can be assumed that the further fate of invention methods and formal algorithms for overcoming non-trivial problems lies in the plane of Machine Learning and Artificial Intelligence approaches. The position of the authors is that the idea of automating inventions looks extremely attractive, although in the coming time, digital approaches will rather complement the intelligence of engineers and scientists, rather than replace it. Taking a certain preparatory step towards AI driven inventions, we present a semantic model that can form the basis of future approaches, at the same time, having already sufficient functionality to support the heuristic stage of technology. As part of this work, over 8 millions of patents and scientific publications have been analyzed to extract semantic concepts. A model was built based on Machine Learning methods and Natural Language Processing techniques with the following discussion and application examples.

1. Introduction

As a source of inventive solutions from all the technical areas, patents form an ample origin of different heuristic concepts of devices and methods proposed by inventors to deliver certain functionality in novel ways. With an idea of systematisation of inventive patterns from such a heuristic solution database, Genrich Altshuller introduced the TRIZ methodology in the late 60s of the XX century [1, 2]. Being born on the analysis of thousands of patents TRIZ accumulated its main tools as Inventive Principles and Trends of Engineering Systems Evolution at the first stage, then Contradiction Matrix and Inventive Principles, followed by Ideality concept and substances/field analysis with Inventive Standards at a dawn of TRIZ development led by the creator of the theory [3]. At the later generations of TRIZ, such tools as OTSM-TRIZ [4] appeared and later TRIZ gradually stepped in the digitalized era [5], [6, 7]. All those tools and methodologies from the very beginning have been evolving around the central idea of formalisation of inventive process, as well as systematisation of the heuristic stage of product and technology design [8]. Keeping the fact that amount of patents grew rapidly during last half a century (and keeps exponential growth at some directions [9]), the main ambition of the present article is to bring together TRIZ idea of heuristic

methods formalisation and modern opportunities of machine learning and natural language processing (NLP) applied to automatic patent analysis in order to open up new functional possibilities for TRIZ community based on novel data analysis algorithms.

Idea of automatic extraction of the semantic features from the text with the help of text mining and natural language processing is not new [10]. Thus, there are the full set of methods and algorithms used in NLP for automatic textual analysis. These algorithms start with the basic preprocessing of a text such as stopword removal, stemming, lemmatization and tokenization, lasts with a semantic term relations identification such as POS tagging and syntax parsing and finally ends with a techniques to analyse corpuses of textual documents to identify inter-term and inter-document relations such as TF-IDF metrics, Word2Vec, Doc2Vec, LDA, hierarchical text clustering algorithms and others [11]. These natural language processing techniques are also used to analyse patent textual data for automatic search, tagging, classification and automatic patent landscaping analysis. Thus, the automatic method of construction of a knowledge organization system is presented in the paper [12] with the help of LDA topic modelling algorithm, K-Means clustering and PCA for results interpretation. The algorithm built in the paper allows to perform patent automatic classification and automatic categorical refinement of the searching results patents pool. Another interesting development is the CPPAT tool that based on the automatic analysis of uploaded patents and scientific publications can produce comparative analysis of the topic based on the uploaded set and present main extracted keywords as main terms and topics. Tool also presents relevant statistics on countries, assignees, inventors and related Universities [13]. The method of comparing claims sections of different patents and the corresponding graphical representation is developed and schematically described in [14], the algorithm allows to compare different patents claim in a convenient way, visualise patent structure and search for similar patents. The algorithm of concept-based search is also presented in the [15] in the context of complementing the CPC patent clusters with automatically classified related patent documents. The automatic patent landscaping is presented in the [16], the idea of the algorithm is that based on some seed pool of patents with the help of neural network the tool produces either narrower or broader set of relevant patents based on extracted textual features and CPC codes.

Another pool of relevant cases is formed by the text mining techniques applied to patent data in the context of TRIZ. Thus the attempt to automatically classify patents based on the recognised Contradiction was described in [17]. Another research related to automatic extraction of patent resolved contradiction is presented in [18]. The Inventive Design Method (IDM-Matching) which is a construction to automatically build links between target problems and inventive solutions semantically extracted from patents data is presented in [19]. One more related research is automatic patents parameters extracting tool based on topic modeling techniques for identification of the contradictory parameters is presented in [20]. The work aimed at computer-aided patent oriented search with the central idea of relevance to the level of innovation is shown in [21].

The structure of the paper is as follows: first, we present the premises of the research in the Introduction, referring to the research that was made in the field of automatic patent analysis, and

also unstructured text analysis in the context of TRIZ challenges. Then, we describe the related work, a new presented model with the Concept Extraction feature and concluding notes.

2. Related Work

The similar related work based on the automatic semantic concept extraction analysis is presented in [22]. Authors built a statistical patent network model based on the 4 millions granted patents from USPTO from 1976 till 2013 years with the help of natural language processing techniques applied to title, keywords and abstracts of the patent documents. Authors additionally use the network analysis of forward and backward citations and CPC patent classifications to refine model connections. As a result authors derived one of the first published models of patents with a semantic classification.
The related work to the present paper was conducted and described in [23]. In this paper the way the dataset of patents and scientific publications is collected and preprocessed is described. Thus Authors compiled a dataset from 8 millions of documents consisting from patents and scientific publications and performed text mining techniques in order to extract features. The methodology uses Doc2Vec algorithm for the whole corpus textual documents vectorization, which are further clustered with a KMeans algorithm [24, 25]. The LDA Topics modeling is exploited over each cluster thus providing a semantical concepts that consist from patents and scientific publications and presented by a relevant bag-of-words with relevance coefficients [26-28]. The resulting model allowed to build concept evolution graphs and compare different semantic concepts.

3. Semantic preprocessing

3.1 Dataset and semantic preprocessing stages

As in [23], in present research the dataset consists of patents documents scrapped from USPTO bulk archive and scientific publications from UK Core Collection [29, 30]. The whole script is implemented via python language and deployed for graphical interface with the help of Django framework. First of all the patents and publications data parsed from corresponding XML and JSON files and put into PostgreSQL database. The steps of the preprocessing are schematically shown on Fig. 1.

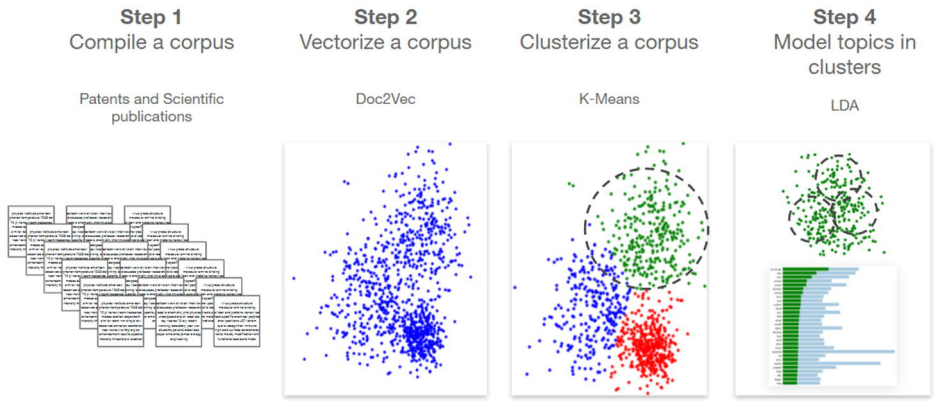

Fig. 1. The preprocessing stages performed over a textual documents corpus (patents from USPTO and publications from UK CORE collections). Step №1. Raw patents and publications are scrapped and put in the database. Step №2. Semantic space of documents is built via textual document vectorisation process (Doc2Vec). Step №3. Semantic space clustering via K-Means algorithm is performed (the amount of clusters on the picture is schematic). Step №4. The Topic Modelling algorithm (LDA) is exploited over each cluster to extract relevant semantic concepts represented by bag-of-words and related patents and scientific publications.

First of all, textual documents are validated if they've not been damaged and empty (some documents may have empty abstract or damaged xml structure - such documents are not put for further analysis). Then documents are checked for a language with the help of NLTK library, only those documents that are in english are used in preprocessing. The main first computational step is a vectorisation of the textual documents corpus. Thus, all the documents are vectorised to a 100-dimensional numerical vector based on their semantic affinity with the help of Doc2Vec algorithm. After each document is vectorised, the next step is to calculate clusters to extract groups with semantically close to each other documents. When those groups are built, the LDA topic modelling algorithm is calculated over each cluster and semantic topics are extracted. The LDA algorithm principle is depicted on Fig. 2.

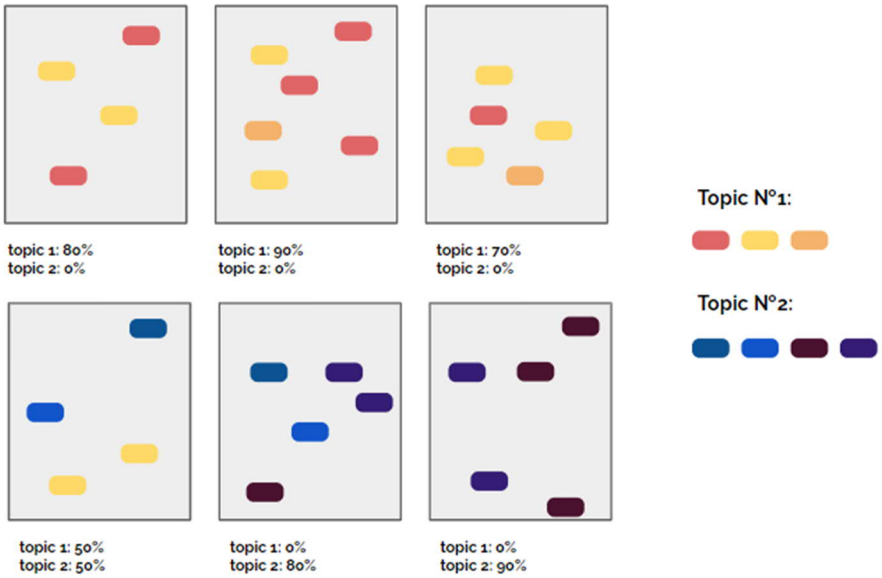

Fig. 2. On this Figure the scheme of the LDA algorithm is presented. On the example, two topics (on the right) are extracted from the six textual documents (on the left) containing the set of words (different words are shown in different colours). Based on words frequency and relevancy within the document the topic relevance to the document is calculated.

3.2 Concept extraction

The central idea of the framework is the semantic concept extraction. The whole preprocessing analysis is aimed to group semantically close textual documents of patents and scientific publications together that represent some common physical conception. Thus, at the last step of the preprocessing the bag-of-words for each semantical concept has been extracted. Thus, being represented by the most relevant keywords each concept might be interpreted and relevant (semantically close) concepts are recommended by a machine. A cloud of recommendations is depicted on Fig.3. To make the semantic model of concepts more representative for engineers and researchers the physical properties and effects have been added to the model.

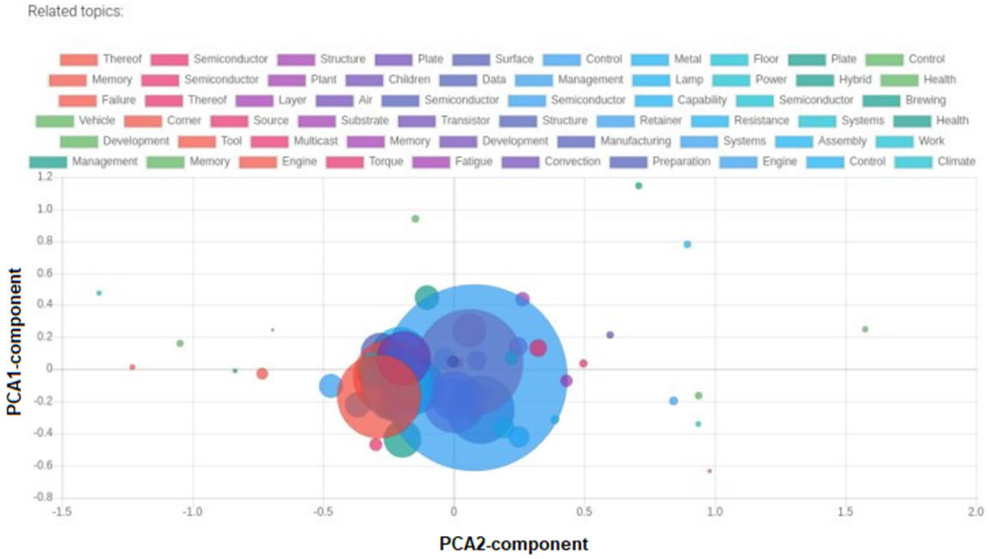

Fig. 3. Different concepts that are semantically close to the current selected concept. The recommendation system is working comparing concepts by 100-dimensional numerical vectors and depicts the closest ones based on PCA analysis.

The idea of physical effects extraction is not new in TRIZ. Thus, there are physical effects databases have been integrated to the majority of digital TRIZ supporting softwares [6], or the one available online without licence required [31]. The research describing the database of the physical effects is also presented in [32]. In our model we integrated the database of physical effects combined in [33]. Besides the effects (207 in total), the physical, electrical, thermal, magnetic, optic and mechanical properties have been extracted (30 in total). The importance of each property term is calculated based on the amount of occurrences in the full text of patent or scientific publication. With the help of this functionality the system allows to perform the physical effect oriented conceptual search and identify relevant concepts automatically. Several machine produced examples of identified concepts are depicted on Table 1.

Table 1. Examples of the semantic concepts identified.

Concept №1. Transistor			**Concept №2. Solar Cell**		
Terms	**Properties**	**Physical Effects**	**Terms**	**Properties**	**Physical Effects**
Current	Conductive	Electrical field	Nanowire	Optical	Electromagnetic induction
Field	Electromagnetic	External field	Solar	Piezoelectric	Ultraviolet
Gate	Resistive	Electrical charge	Cell	Flexible	Absorption
Drain	Dielectrical	Field-Effect	Silicon	Conductive	Polarized light
Insulator	Electronic	Tunnel effect	Semiconductor	Solid	Light carriers

3. Discussion

The main result of the present research is a model that, based on a large amount of input data (more than 8 million text documents, patents and publications), builds objective semantic models based on text mining techniques and natural language processing algorithms. A feature of this model is the extracted semantic concepts, taking into account various physical properties (mechanical, electrical, etc.) and physical effects. Once built, such models make it suitable to navigate among the semantic concepts, find relevant ones and obtain related patents and publications, and find other concepts by the degree of proximity as well. In addition, taking into account the fact that there are textual documents with a known publication date is tied to each concept, these concepts can be used to identify their evolution trends, which combines the digital and automatic version of the S-curve tool from TRIZ. Another feature is that such a model makes it possible to conduct patent-oriented search more effectively, since semantic models include related physical effects. As demonstrated by the example, the extracted concepts are meaningful and easy to interpret even for not expert users of the system.

Of course, the semantic approach has its limitations related to the fact that a property mentioned in a patent or publication can be mentioned as not achieved within the framework of the automatically analysed study, but on the one hand, such information is still interesting for a potential decision-maker, since the property is still related to the topic, and on the other hand, such flaws can be handled in future steps of the model development through more advanced natural language processing algorithm that allows to identify relations of the mentioned terms. Another fair limitation of the

present model is the lack of quantitative measures of the extracted properties. Being integrated numerical characteristics of the concepts would allow for highly relevant recommendations due to possible quantitative search for parameters, however, currently this refinement is planned for the next step. The authors do not consider the use of only english versions of patents as a significant limitation, since most patent families have copies in the USPTO (which are published in english).

4. Conclusion

TRIZ was born as a theory being based on a manual analysis and systematization of patent documents over 50 years ago. Since then, effective digital tools have appeared and computational possibilities for automatic analysis of textual data have grown and many research studies have been published that try to extract heuristic information from patents using text mining algorithms and natural language processing [34, 35]. Thus, in our study, semantic models were built based on the patents data (3,514,730 patents applications from USPTO) and scientific publications (4,875,744 publications from UK CORE).

A key feature of semantic models is the independence and objectivity of the data. Despite the fact that the proposed model does not allow establishing functional relationships or cause-effect chain relationships, the presented model is ready for identifying a concept, extracting information about evolution trends of the concept development, and is able to find semantically close and distant concepts, as well as compare different concepts one with each other and find relevant research documents (scientific publications and patents) that semantically (by the matching relevancy of corresponding terms) are between the selected concepts.

The presented model assists in the conceptual design stage and supports the heuristic state of technology development. Since the model not only counts the most frequent and relevant words, but also extracts related physical properties and effects, such a model can provide more information to the researcher or engineer about the concept and makes the search and recommendations significantly more relevant.

As a further direction of research, it is planned to add extracting data from images attached to patent applications, as well as establish functional and cause-effect chains relationship between the elements of the extracted concepts to increase the capabilities of computer-aided design.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie grant agreement № 722176.

References

1. Salamatov, Y. and Souchkov, V., 1999. TRIZ: the right solution at the right time: a guide to innovative problem solving (p. 256). Hattem: Insytec.

2. Altshuller, G., & Altov, H. (1996). And suddenly the inventor appeared: TRIZ, the theory of inventive problem solving. Technical Innovation Center, Inc..
3. Litvin, S., Petrov, V. and Rubin M. (2007) TRIZ Body of Knowledge. The TRIZ Developers summit 2007. *Available online:* https://triz-summit.ru/en/203941/
4. Cavallucci, D., & Khomenko, N. (2007). From TRIZ to OTSM-TRIZ: addressing complexity challenges in inventive design. International Journal of Product Development, 4(1-2), 4-21.
5. Cascini, G. (2004). State-of-the-art and trends of computer-aided innovation tools. In Building the information society (pp. 461-470). Springer, Boston, MA.
6. http://invention-machine.com/custsupport/to_install.cfm (last visited 4/2021)
7. https://ihsmarkit.com/products/enterprise-knowledge.html (last visited 4/2021)
8. Savransky, S. D. (2000). Engineering of creativity: Introduction to TRIZ methodology of inventive problem solving. CRC press.
9. Artificial Intelligence (2019). WIPO Technology Trends 2019. *Available online:* https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf
10. Loper, E., & Bird, S. (2002). NLTK: the natural language toolkit. arXiv preprint cs/0205028.
11. Joseph, S. R., Hlomani, H., Letsholo, K., Kaniwa, F., & Sedimo, K. (2016). Natural language processing: A review. Natural Language Processing: A Review, 6, 207-210.
12. Hu, Z., Fang, S., & Liang, T. (2014). Empirical study of constructing a knowledge organization system of patent documents using topic modeling. Scientometrics, 100(3), 787-799.
13. Ranaei, S., Knutas, A., Salminen, J., & Hajikhani, A. (2016, June). Cloud-based Patent and Paper Analysis Tool for Comparative Analysis of Research. In CompSysTech (pp. 315-322).
14. Okamoto, M., Shan, Z., & Orihara, R. (2017, August). Applying information extraction for patent structure analysis. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 989-992).
15. Montecchi, T., Russo, D., & Liu, Y. (2013). Searching in Cooperative Patent Classification: Comparison between keyword and concept-based search. Advanced Engineering Informatics, 27(3), 335-345.
16. Abood, A., & Feltenberger, D. (2018). Automated patent landscaping. Artificial Intelligence and Law, 26(2), 103-125.
17. Liang, Y., Tan, R., & Ma, J. (2008, September). Patent analysis with text mining for TRIZ. In 2008 4th IEEE International Conference on Management of Innovation and Technology (pp. 1147-1151). IEEE.
18. Cascini, G., & Russo, D. (2007). Computer-aided analysis of patents and search for TRIZ contradictions. International Journal of Product Development, 4(1-2), 52-67.
19. Ni, X., Samet, A., & Cavallucci, D. (2020, October). Build Links Between Problems and Solutions in the Patent. In International TRIZ Future Conference (pp. 64-76). Springer, Cham.
20. Berdyugina, D., & Cavallucci, D. (2020, October). Setting Up Context-Sensitive Real-Time Contradiction Matrix of a Given Field Using Unstructured Texts of Patent Contents and

Natural Language Processing. In International TRIZ Future Conference (pp. 30-39). Springer, Cham.
21. Regazzoni, D., & Nani, R. (2008). TRIZ-Based patent investigation by evaluating inventiveness. In Computer-Aided Innovation (CAI) (pp. 247-258). Springer, Boston, MA.
22. Bergeaud, A., Potiron, Y., & Raimbault, J. (2017). Classifying patents based on their semantic content. PloS one, 12(4), e0176310.
23. Kaliteevskii, V., Deder, A., Peric, N., & Chechurin, L. (2020, October). Conceptual Semantic Analysis of Patents and Scientific Publications Based on TRIZ Tools. In International TRIZ Future Conference (pp. 54-63). Springer, Cham.
24. Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. Pattern recognition, 36(2), 451-461.
25. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
26. Huang, C. H., Yin, J., & Hou, F. (2011). A text similarity measurement combining word semantic information with TF-IDF method. Jisuanji Xuebao(Chinese Journal of Computers), 34(5), 856-864.
27. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
28. Řehůřek, R. and Sojka, P., 2011. Gensim—statistical semantics in python. statistical semantics; gensim; Python; LDA; SVD.
29. https://www.uspto.gov/ (last visited 5/2020)
30. https://core.ac.uk/ (last visited 5/2020)
31. Oxford Creativity. Physical effects and functions database. *Available online:* http://wbam2244.dns-systems.net/EDB/index.php (last visited 5/2020)
32. Fomenkov, S. A., Kolesnikov, S. G., Korobkin, D. M., Kamaev, V. A., & Orlova, Y. A. (2014). The information filling of the database by physical effects. Journal of Engineering and Applied Sciences, 9(10-12), 422-426.
33. Physical Effects database. *Available online:* http://bionicinspiration.org/physical-effects/ (last visited 5/2020)
34. Efimov-Soini, N.K. and Chechurin, L.S., 2016. Method of ranking in the function model. Procedia CIRP, 39, pp.22-26.
35. Renev, I., Chechurin, L. and Perlova, E., 2017. Early design stage automation in architecture-engineering-construction (AEC) projects. In Proceedings of the 35th eCAADe conference (pp. 373-382).