ISBN: 978-981-18-5852-9
WCSE 2022 Spring Event: 2022 9th International Conference on Industrial Engineering and Applications
doi: 10.18178/wcse.2022.04.028

A Multi-granularity Knowledge Representation and Mining Method for Patent Texts

Lin Gong [1], Mingren Zhu [2], Zhenchong Mo [3], Ziyao Huang [4]

[1] School of Mechanical Engineering, Beijing Institute of Technology, Beijing, China
Yangtze Delta Region Academy of Beijing Institute of Technology, Jiaxing, China
gonglin@bit.edu.cn

[2] School of Mechanical Engineering, Beijing Institute of Technology, Beijing, China
2508828250@qq.com

[3] School of Mechanical Engineering, Beijing Institute of Technology, Beijing, China
837598830@qq.com

[4] School of Mechanical Engineering, Beijing Institute of Technology, Beijing, China
huang_ziyao@126.com

Abstract. To support patent knowledge reuse for product concept design, multi-granularity representation of patent knowledge should be taken into consideration. Traditionally, an ontology model called TCO (Techspecs Concept Ontology) is used for representing the hierarchical architecture of a product, which consists of components layer, function modules layer, and product layer. However, the interactions such as the context relevance among components are ignored in the TCO model. In this paper, a modified model called PSG-TCO is proposed, where the PSG (Patent Semantic Graph) part can be used to capture components' interactions. Based on the PSG-TCO model, an automatic knowledge extraction method is proposed to construct PSG-TCO instances from a large number of patent texts. All of the PSG-TCO instances form a multi-granularity knowledge base, which can be used for engineering knowledge retrieval, design concept discovery, and providing potential innovation stimulus.

Keywords: knowledge-based design, patent analysis, PSG-TCO model

1. Introduction

The development of product innovation design is an important aspect to improve the innovation ability comprehensively. It not only affects the enterprise competitiveness, but also plays a vital role in industrialization level of the country. In the process of product design, especially in the conceptual design stage, the divergent thinking of designers is the main influence factor to form innovation. However, such divergent thinking is usually limited due to the thinking pattern formed by the designer's professional knowledge or personal experience. Therefore, how to broaden and explore the design space is the key problem of product innovation design.

To deal with this problem, it is traditional to work cooperation among a few designers. Methods such as brainstorming [1] and 6-3-5 [1] are used to explore the design space. However, these methods often need to consume large manpower and time resources, and increase the management cost [2]. With the development of artificial intelligence technology in recent years, data and knowledge driven methods have become the mainstream technology solutions to design space expansion problems. There are three main advantages of such methods: (1) external knowledge can increase the objectivity during design process; (2) huge knowledge base can effectively cover the design space; (3) automated programs can provide accurate and rapid creative knowledge retrieval services. The three advantages inspired a lot of works about knowledge-based design methods, and the basis of these works is to construct a knowledge base with potential innovation stimulus ability.

The data sources for the knowledge base can be varied among related works, such as internet resources [3], industrial data [4], scientific papers [5], and patent database [6].The data sources for knowledge base can

be varied among related works. Compared to other data sources, the design knowledge contained in patents is more systematic and organized, especially for the invention patents. From the micro aspect, invention patents contain a wide range of design knowledge about products, functions, structures, configurations, working principles and operating mechanisms. From the macro aspect, invention patents are organized by the IPC classification system, cover all technical fields and accumulate over time. Such advantages make invention patents a good data source to build the knowledge base for supporting engineering design [7].

Representation of knowledge is the key issue in knowledge base construction and it should be decided based on the data source and application needs. Eventually, different representation will lead to different extraction, storage and retrieval method of the knowledge base. Given the data source and the application needs, the representation of knowledge should be designed carefully. For example, to classify or cluster the patents for recommendation, the representation should cover the main information of the whole patent or abstract text, which could be bag-of-words [8], tf-idf [9], topic distribution [10], and other deep language model semantic vectors [11]. Those representation methods can be used for inspiring designers in document-level. For knowledge-based design, the representation should be more detailed, easy to read and understand. Semantic networks or knowledge graphs are used increasingly in various activities of design process [12]. WordNet [13] and ConceptNet [14] are the most often used public semantic networks for supporting design activities [12]. However, these two knowledge bases are built in general field and not focused on engineering design. Semantic networks built with invention patents can be focused on engineering design, like B-Link [12] and TechNet [7]. These semantic networks are all supporting engineering design in term-level.

There are lots of patent knowledge bases that could be used to support designers to explore the potential design space and get new ideas from existing patent knowledge at different levels, i.e., the document level and the term level. However, the architecture of patent products is ignored in most related works. The architecture of a product can be represented by TCO, an ontology model which states a product can be decomposed into three layers: components layer, functional modules layer, and product layer [15]. Components are the basic physical units of a product. A functional module consists of a few components and has its own specific function to support the whole product. A product consists of a few different functional modules. However, traditional TCO can't capture the interactions among components and functional modules because of the tree architecture. In this paper, we propose the PSG model to deal with the interactions because graphs have natural advantages in representing relationships. In addition, PSG can be combined with TCO to form a new ontology model called PSG-TCO, which is able to capture both the hierarchy architecture and the interactions of a patent product. Based on the PSG-TCO model, we propose an automatic knowledge extraction method to construct PSG-TCO instances from a large number of patent texts. All of the PSG-TCO instances form a multi-granularity knowledge base, which can be used for engineering knowledge retrieval, design concept discovery, and providing potential innovation stimulus.

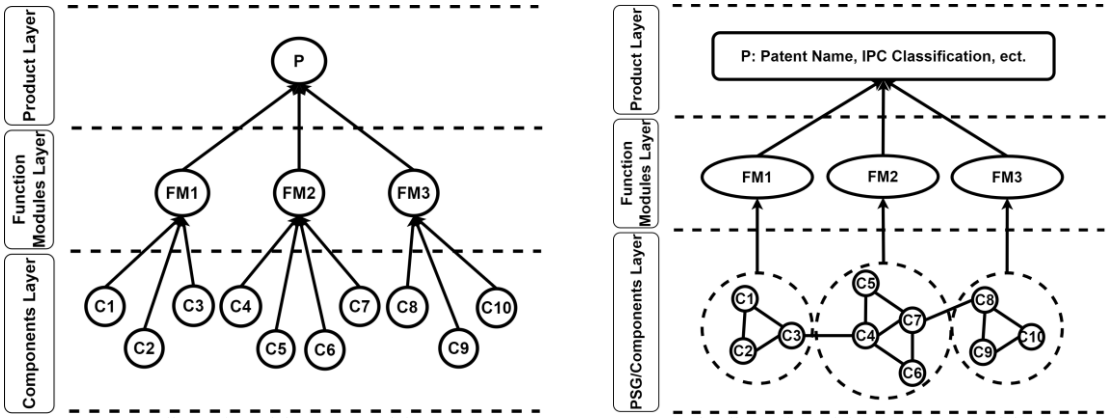
2. Knowledge Representation

We propose a multi-granularity knowledge representation and automatic mining method for patent texts in this paper. To capture patent product architecture and inner interactions, we propose the PSG model and combined it with the traditional TCO model, which form a new ontology model called PSG-TCO. Based on the PSG-TCO model, we propose an automatic knowledge extraction method to build PSG-TCO instances from a large number of patent texts. In this section, we will describe the detailed definition of the PSG-TCO model, and then give out the framework of the proposed automatic knowledge extraction method in the next section.

The traditional TCO model treats the product architecture as three layers: the components layer, the function modules layer, and the product layer. Components are basic physical units of the product. A function module consists of a few components separately, and has its own specific function to support the product. A product consists of a few different modules. The TCO model can be represented with a triple (Cs,FMs,P), where $Cs = \{C_i | i = 1,\cdots,n\}$ is a set of the product components, $FMs = \{FM_j | j = 1,\cdots,m\}$ is a set of the function modules and P stands for the product. Fig. 1(a) shows the tree architecture of the TCO model.

However, components in a product do not exist in isolation, instead they are related to each other. The tree architecture can not representation the interactions among the components, which makes the TCO model

inadequate. Graphs have the natural advantage of representing interactions and are used in many filed such as social science [16] and biological medicine [17]. Here we propose the PSG (Patent Semantic Graph) model to capture the product components' interactions contained in the related innovation patent text. The PSG model can be represented as a graph $G = (Cs,Rs)$, where Cs is the set of components, $Rs = \{(C_i,C_j,R_{ij})|C_i,C_j \in Cs,R_{ij} \in R\}$ is the interactions among the components and R is a set of different types of relationships defined in the following. In fact, some related works have defined a few types of relationships, such as the CFG model [18] which classify the interactions into energy flow, material flow and signal flow. However, those relationships are fuzzy and difficult to confirm by automatic methods, which can lead to much more cost in building such a knowledge base. The semantic relationships between words is easy to define and calculate by automatic methods, which is used in many semantic networks knowledge base [12]. Here we define two semantic relationships of different strength: (1) if two component words appear in the same sentence in the patent text, it is considered that there is a strong semantic relationships between the two components; (2) if two component words appear in adjacent sentences, it is considered that there is a weak semantic relationships between the two components; (3) otherwise, there is no semantic relationships between the two components.



(a) Illustration of the TCO model (b) Illustration of the PSG-TCO model

Fig. 1: The traditional TCO model (a) has a tree architecture. The modified PSG-TCO model (b) combines the PSG model and the TCO model, which can capture the interactions among components based on graph.

The PSG model can be combined with the TCO model to form a modified ontology model called PSG-TCO, which is illustrated in Fig. 1(b). Although the PSG-TCO model treats the patent product architecture into three layers as same as the traditional TCO model, there are some differences between the two. First of all, the components layer in PSG-TCO is changed into the PSG instance G instead of the components set Cs. Secondly, a function module FM_j in PSG-TCO still consists of a few components of the patent product, but they are not separate instead related to each other, which can be represented as a subgraph of the PSG instance. More specifically, a function module in PSG-TCO is defined as a subgraph consisting of closely related components of the PSG instance and has its own function. In this paper, we propose 5 types of function modules according to the common functions of invention patents:

- Dynamic module: Including drive, transmission and action implementation modules;
- Logistic module: Including material supply, storage, transportation and consumption modules;
- Information module: Including information collection, interaction, control, calculation and transfer modules;
- Energy module: Including energy other than mechanical energy supply, store, transfer, convert and consume modules;
- Support module: Including support, stability, clamping, positioning modules.

Additionally, patent attributes such as its name and IPC classification can be added to the product layer to describe the patent product. The PSG-TCO model captures the hierarchical architecture and inner interactions of patent products, which can provide a more complete structured knowledge to support designers in their design space exploration.

3. Knowledge Base Construction

Based on the PSG-TCO model, we propose an automatic knowledge mining method for a large number of patent texts. Fig. 2 shows the overview of our method.

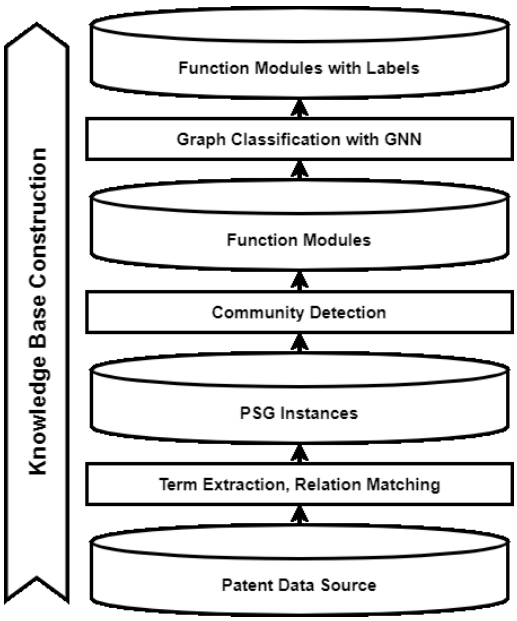


Fig. 2: Overview of the proposed automatic knowledge mining method.

In this section, we will describe the details of the proposed method. Specifically, the patent data source used in our research is downloaded from the patent data service system of Chinese National Intellectual Property Administration(http://patdata.cnipa.gov.cn/). Patents in other language like english can also be easily fitted into this method. We introduce the NLP technologies used to build PSG instances from patent texts, the community detection technologies used to extract function modules from related PSG instance, and the graph neural networks used to classify function modules.

To obtain the main components in the related patent product, we use regular expressions to match the terms contained in the illustration part as showed in Fig. 3. The main components listed in the illustration part of the patent are generally composed of component indexes, component terms, and intervals between them:

- The component indexes may precede or follow the component terms, which usually starts with an Arabic digit and consists of Arabic or Roman digits or lowercase English letters.
- The component terms are mainly composed of Chinese characters and uppercase English letters.
- The interval between the component index and the component term is not always available, but if there is an interval, it is usually represented by a space, a stop, a comma, a colon, a dash, or a period.

According to the order of component indexes and component terms and whether there are intervals or not, we set four kinds of regular expressions to extract component terms from the illustration part of patent texts, which are (1) [index][term], (2) [term][index], (3) [index][interval][term] and (4) [term][interval][index].

附图中的标记为：机架1（Body frame 1）、进料装置2（Loading equipment 2）、进料倾斜支架（Loading bracket 21）、下落通道22（Whereabouts channel 22）、进料过渡支架22（Loading transition bracket 22）、进料通道221（Loading channel 221）、进料气缸23（Loading cylinder 23）、进料板231（Loading board 231）、进料推杆24（Loading putter 24）、夹紧定位装置3（Clamping localizer 3）、夹紧模块31（Clamping module 31）、夹紧配合架311（Clamping frame 311）、传送带通孔3111（Conveyor belt hole 3111）、夹紧电机312（Clamping motor 312）、转轮3121（Runner 3121）、夹紧旋转套313（Clamping rotating sleeve 313）、第一三爪卡盘314（First 3-jaw chuck 314）、第一配合齿轮3141（First cooperating gear 3141）、第二三爪卡盘315（Second 3-jaw chuck 315）……

Fig. 3: An example of the illustration part.

To built the PSG instances, we then extract semantic relationships among those components. The rest parts such as abstracts, claims, invention contents, and implementations should be taken into consideration because of their rich interaction information. In our research, we only focus on the strong semantic relationships, which means if and only if two component terms appear in the same sentence of the above patent parts, we consider there is a link between the two. This task can be accomplished by simple matching and then the PSG instances are built.

As mentioned in section II, a PSG instance can be decomposed into several function modules, all of which are made up of closely related components. These function modules are represented as subgraphs, or specifically, communities of PSG instances, which can be extracted by community detection methods. Community detection methods proposed for different applications can be divided into 3 categories: modularity based methods [19], map equation based methods [20] and stochastic block based methods [21]. In our research, we prefer modularity based methods especially the Louvain algorithm [22], because it is more suitable for small graphs as PSG instances than map equation based methods, and faster than stochastic block based methods.

Extracted function modules are classified in a supervised manner using graph neural networks (GNNs). We experimented with both GCN [23] and GAT [24] for the function modules classification, where the adjacency matrix of the function module and the semantic vectors of its components are used as GNNs' inputs. It should be noted that the semantic vectors are given by the related node embeddings [25] of the components network, which is constructed using edit distance [26]: if the edit distance between two component terms is less than 1, then there is an edge between them.

We marked 300 function modules manually, 200 of which were divided into training sets and 100 into test sets. Table I shows the F1 scores in the test sets. It can be seen that GCN and GAT have different performance in different types of function modules. For example, the F1 score of GCN reaches 0.883 in the category of support modules, while only 0.582 in the category of energy modules. This may be caused by the large difference in the number distribution of different types of function modules. GAT has a good performance in all of the types, so it is used to classify the other function modules.

Table 1: F1 scores of function modules classification

	Dynamic	Logistic	Information	Energy	Support
GCN	0.715	0.655	0.799	0.582	0.883
GAT	0.689	0.667	0.741	0.642	0.754

An example of PSG instance and it's function modules are showed in Fig.4. It is a slotting machine and it consists of two support modules, two logistic modules and a dynamic module. This architecture conforms to the characteristics of slotting machine, a large automatic machine which needs stable support and automatic logistics.

After the above processing, we finally obtained a total of 39,463 PSG-TCO instances, 143,819 function modules, and 1,105,431 components. On average, each PSG-TCO instance has 3.6 function modules and 28 components. Each functional module has an average of 7 components.

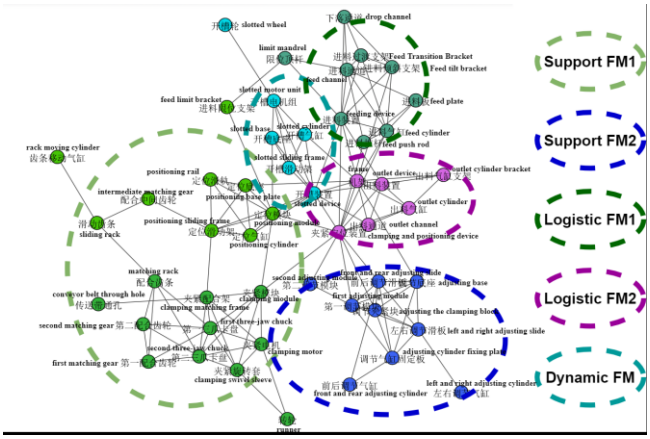


Fig. 4: Example of a slotting machine PSG instance and it's function modules.

4. Knowledge Application

The knowledge base of PSG-TCO instances can provide a multi-granularity innovation stimulation, especially for inexperienced novice designers. PSG-TCO instances provide architecture information about existing designed products for designers to retrieve and reuse knowledge at product level, module level and component level. If we want to design a new product, we can retrieve the existing design in PSG-TCO instances searching for innovation chances. Taking the trash can for example, Fig.5 shows three PSG-TCO instances of different trash cans.

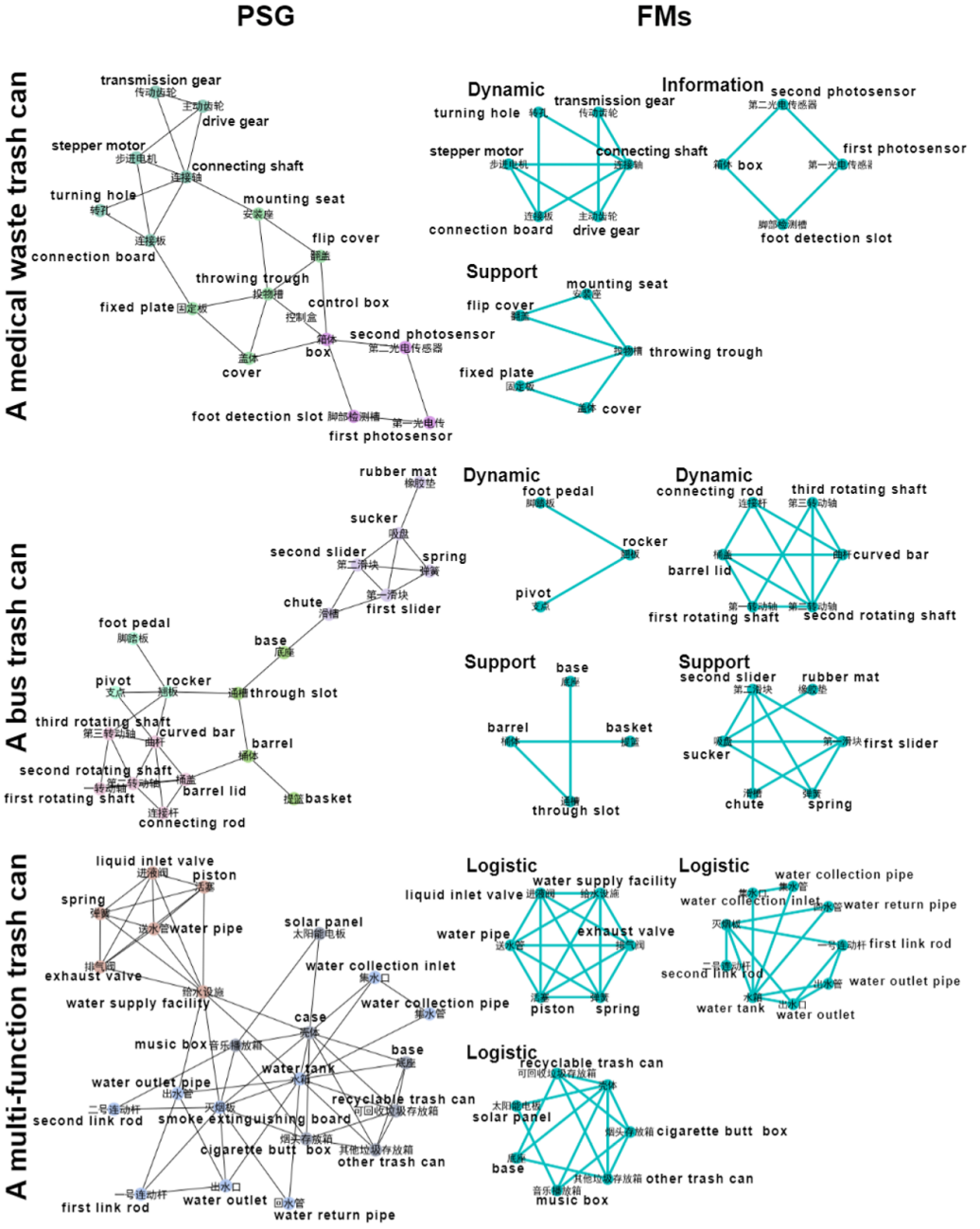


Fig. 5: PSG-TCO instances of different trash cans.

The medical waste trash can contains an information module, indicating that it adopts automated technology to assist waste delivery, which may be for safety protection. Specifically, the trash can uses two photoelectric sensors in the information module to detect the feet of waste delivery personnel. Once it detects that someone needs to drop waste, the stepper motor will start and open the lid with a fine gear drive. The but trash can is mainly made of traditional power and structure module, and adopts the traditional pedal-type

clamshell method, which belongs to pure mechanical structure. The multi-function trash can emphasizes the logistics module, which automatically extinguishes smoke by introducing water and provides the function of a music player using solar power.

The results returned by traditional inspection systems such as CNKI are often presented in the form of abstract or attached drawings. The pure text abstract makes designers have a hard reading burden, while the attached drawings do not adequately represent the structure architecture and function characteristics of patent products. In contrast, the patent knowledge represented by the PSG-TCO model is clear, which is more convenient for designers to understand the acquired knowledge and make horizontal comparison to find innovation chances. On this basis, a more precise combinatorial knowledge retrieval system can be designed to assist designers in knowledge localization because the PSGTCO model provides multi-granularity information about the PSG and the function modules, because the PSG-TCO model provides multi-granularity information about the PSG and the function modules.

5. Conclusions and Prospects

In this paper, we proposed a multi-granularity knowledge representation method for patent texts. The PSG-TCO model makes up for the shortcomings of existing methods in representing the hierarchical architecture and component interactions of patent products. On the basis of the PSG-TCO model, we proposed an automatic construction method for the corresponding knowledge base to deal with the massive and increasing patent data. Firstly, the main components of patent products are extracted from the patent illustration by regular expression, and then the semantic relations of them are matched from other parts of the patent text to construct PSG instances. Then, the function modules are extracted from PSG instances using community detection technologies. At last, GNNs are used to classify extracted function modules to supplement category information for supporting quick understanding and accurate retrieval.

We built a knowledge base from Chinese patent data and showed the help that PSG-TCO instances can provide based on a case study. The case study shows that PSG-TCO instances can effectively represent the hierarchical architecture and component interactions of patent products. In addition, the visualization of graph models is easier and faster for designers to understand existing designs and provides effective comparisons between them, which makes designers more convenient to find potential innovation chances.

In the future studies, we will build corresponding intelligent systems to support design processes and provide designers with more convenient and effective interactive experience.

6. Acknowledgements

The research was supported by a National Key Research and Development Project (No. 2018YFB1700802) and a National Ministry Basic Research Project (No. JCKY2016203A017) and a National Natural Science Foundation of China (No. 51405018).

7. References

[1] D. Ullman, *The Mechanical Design Process*, 01 2018.

[2] J. Luo, S. Sarica, and K. L. Wood, “Guiding data-driven design ideation by knowledge distance,” *Knowledge-Based Systems*, vol. 218, p. 106873, 2021.

[3] X. Zhang, X. Liu, X. Li, and D. Pan, “Mmkg: An approach to generate metallic materials knowledge graph based on dbpedia and wikipedia,” *Computer Physics Communications*, vol. 211, pp. 98–112, 2017, high Performance Computing for Advanced Modeling and Simulation of Materials.

[4] J. Pokojski, K. Szustakiewicz, Łukasz Woznicki, K. Oleksi´ nski, and´ J. Pruszynski, “Industrial application of knowledge-based engineering´ in commercial cad / cae systems,” *Journal of Industrial Information Integration*, p. 100255, 2021.

[5] Y. Zhu, Q. Lin, H. Lu, K. Shi, P. Qiu, and Z. Niu, “Recommending scientific paper via heterogeneous knowledge embedding based attentive recurrent neural networks,” *Knowledge-Based Systems*, vol. 215, p. 106744, 2021.

[6] L. Liu, Y. Li, Y. Xiong, and D. Cavallucci, “A new function-based patent knowledge retrieval tool for conceptual

design of innovative products," *Computers in Industry*, vol. 115, p. 103154, 2020.

[7] S. Sarica, J. Luo, and K. L. Wood, "Technet: Technology semantic network based on patent data," *Expert Systems with Applications*, vol. 142, p. 112995, 2020.

[8] S. Ishihara, "Score-based likelihood ratios for linguistic text evidence with a bag-of-words model," *Forensic Science International*, vol. 327, p. 110980, 2021.

[9] P. Qin, W. Xu, and J. Guo, "A novel negative sampling based on tfidf for learning word representation," *Neurocomputing*, vol. 177, pp. 257–265, 2016.

[10] J. Yun and Y. Geum, "Automated classification of patents: A topic modeling approach," *Computers & Industrial Engineering*, vol. 147, p. 106636, 2020.

[11] J.-S. Lee and J. Hsiang, "Patent classification by fine-tuning bert language model," *World Patent Information*, vol. 61, p. 101965, 2020.

[12] J. Han, S. Sarica, F. Shi, and J. Luo, "Semantic Networks for Engineering Design: State of the Art and Future Directions," *Journal of Mechanical Design*, vol. 144, no. 2, 09 2021, 020802.

[13] G. A. Miller, "Wordnet: A lexical database for english," in *Proceedings of the Workshop on Human Language Technology*, ser. HLT '93. USA: Association for Computational Linguistics, 1993, p. 409.

[14] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," *CoRR*, vol. abs/1612.03975, 2016.

[15] S. Moon, S. Kumara, and T. Simpson, "Knowledge representation for product design using techspecs concept ontology," 01 2005, pp. 241– 246.

[16] M. A. Kolak, Y.-T. Chen, Q. Lin, and J. Schneider, "Social-spatial network structures and community ties of egocentric sex and confidant networks: A chicago case study," *Social Science & Medicine*, p. 114462, 2021.

[17] T. Szocinski, D. D. Nguyen, and G.-W. Wei, "Awegnn: Autoparametrized weighted element-specific graph neural networks for molecules," *Computers in Biology and Medicine*, vol. 134, p. 104460, 2021.

[18] T. Kurtoglu, M. Campbell, C. Bryant, R. Stone, and D. Mcadams, "Deriving a component basis for computational functional synthesis," *Proceedings ICED 05, the 15th International Conference on Engineering Design*, 01 2005.

[19] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[20] M. Rosvall, D. Axelsson, and C. T. Bergstrom, "The map equation," *The European Physical Journal Special Topics*, vol. 178, no. 1, p. 13–23, Nov 2009.

[21] B. Karrer and M. E. J. Newman, "Stochastic blockmodels and community structure in networks," *Physical Review E*, vol. 83, no. 1, Jan 2011.

[22] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, Oct 2008.

[23] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2017.

[24] P. Velickovi˘ c, G. Cucurull, A. Casanova, A. Romero, P. Li´ o, and` Y. Bengio, "Graph attention networks," 2018.

[25] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 701–710.

[26] R. A. Wagner and M. J. Fischer, "The string-to-string correction problem," *J. ACM*, vol. 21, no. 1, p. 168–173, Jan. 1974.