ENGINEERING KNOWLEDGE GRAPH FOR KEYWORD DISCOVERY IN PATENT SEARCH

Sarica, Serhad (1); Song, Binyang (1); Low, En (2); Luo, Jianxi (1)
1: Singapore University of Technology and Design, Engineering Product Development Pillar; 2: Singapore University of Technology and Design

ABSTRACT

Patent retrieval and analytics have become common tasks in engineering design and innovation. Keyword-based search is the most common method and the core of integrative methods for patent retrieval. Often searchers intuitively choose keywords according to their knowledge on the search interest which may limit the coverage of the retrieval. Although one can identify additional keywords via reading patent texts from prior searches to refine the query terms heuristically, the process is tedious, time-consuming, and prone to human errors. In this paper, we propose a method to automate and augment the heuristic and iterative keyword discovery process. Specifically, we train a semantic engineering knowledge graph on the full patent database using natural language processing and semantic analysis, and use it as the basis to retrieve and rank the keywords contained in the retrieved patents. On this basis, searchers do not need to read patent texts but just select among the recommended keywords to expand their queries. The proposed method improves the completeness of the search keyword set and reduces human effort for the same task.

1 INTRODUCTION

Patent retrieval has become a necessary and important task in engineering design. Engineers often search for patents related to a design topic (e.g., autonomous vehicle, additive manufacturing) in order to learn the design precedents for inspiration (Fu et al., 2013; Srinivasan et al., 2018). For instance, various methods and tools have been developed to retrieve patents and design information from patent documents to aid in the use of TRIZ (Altshuller and Shapiro, 1956) for innovative problem solving (Cascini and Russo, 2006), design-by-analogy (Fu et al., 2014), function analysis for product platform and product family planning (Song et al., 2018). Patent search and analytics also allow for inventors to evaluate the novelty of their new designs (He and Luo, 2017). In addition, patent attorneys conduct patent search to validate the legal enforceability of the claims of a newly applied patent (Fujii, 2007). Patent data have also been exploited for technology roadmapping and innovation intelligence (Jeong et al., 2015). In brief, patent retrieval and analytics support engineering design, intellectual property management, knowledge management, and innovation planning.

The richness of patent databases offers both opportunities and challenges to patent retrieval. In the United States Patent & Trademark Office's (USPTO) patent full text and image database alone, there are more than 6 million patent records from 1976 to date. It is a non-trivial task to retrieve the set of patents that are accurately relevant to a specific search interest, such as 3D printing and rolling robots, for the interests of engineers, designers, and innovators. The existing retrieval methods rely on keyword search in patent texts (Fujii, 2007; Koch et al., 2011; Nakamura et al., 2015) and mining additional structured information in patent documents, such as inventors, assignees, classification and citations (Wang, 2011; Benson and Magee, 2013; Song and Luo, 2017). Keyword search in patent texts is the key in general patent retrieval procedures, methods, and tools, but the keywords are often intuitively chosen by searchers subject to their knowledge and vocabulary on the technology of search interest (Fujii, 2007; Wang, 2011; Montecchi, Russo and Liu, 2013).

It is likely and natural for humans to overlook keywords relevant to the focal technology, which results in incompleteness. Human searchers can read patent texts from prior searches to discover additional keywords that are not obvious to them initially and expand their search queries heuristically and iteratively. However, such a human reading process is tedious, time-consuming and prone to human cognitive errors. In this paper, we use natural language processing (NLP) and word embeddings-based

semantic analysis to substitute human reading of patent texts, and partially automate and augment the heuristic and iterative keyword discovery process. The NLP-based method is expected to improve the completeness and accuracy of the search query and thus patent retrieval results, while significantly reducing human efforts and the time required for the same task. Such advantages are greater when the patent search is for larger systems and technology fields that hold many relevant patents.

2 RELATED WORK

Keyword search is the most commonly used method for patent retrieval. Typically, searchers decide keywords according to their intuition and expertise, and then combine them in search queries using Boolean operators (Alberts et al., 2011). To facilitate such a process, Wang (2011) proposed to screen the text of patents granted to the leading patentees and inventors in a focal domain for identifying relevant keywords; a computer-aided tool, PatViz, was created to enable iterative refinement of the complex search queries visually and interactively (Koch et al., 2011). Other researchers retrieved relevant patents for validity search using keywords and utilized citation link information to assess the relevance of the obtained patents (Fujii, 2007). Similar to a keyword-based search, Nakamura et al. (2015) utilized an assignee-based search to screen the patents conferred to the leading companies within a focal industry to particularly retrieve the industry-related patents.

Keyword search has also been integrated with other patent retrieval methods to improve the relevance and the completeness of the search results. In particular, a group of researchers have utilized the classification information for this purpose. A patent is categorized to one or multiple patent classes according to an established patent classification system, such as international patent classification (IPC) or US patent classification (USPC). Among them, Wang (2011) identified a few IPC and USPC classes highly related to the focal technology via an initial search and then search relevant patents within these patent classes using keywords; Alberts et al. (2011) suggested to limit the range into certain IPC and USPC classes for a keyword search specific to a focal technology to increase the search effectiveness; Benson and Magee (2013) proposed the classification overlap method (COM) that starts with a keyword search to identify the top IPC and USPC classes containing the most number of the returned patents and then take patents in the intersection of the IPC and USPC classes as the search result. In general, COM is suitable for patent retrieval for general technological domains, such as nanotechnologies, fuel cells, and electric motors. In addition, Song and Luo (2017) proposed a method to integrate the keyword search in patent texts and the searches using citation relationship and co-inventor relationship between patents, and use an iterative process to read the previously retrieved patents to expand the keyword queries, in order to arrive at a more comprehensive and accurate search.

Another strand of approaches exploits semantic analysis to search for patents in the vast databases. D'hondt (2009) developed a syntactic approach of interactive patent retrieval by normalizing syntactic and morphological variations in patent texts. Takaki et al. (2004) and Xue and Croft (2009) extracted keywords from sample patent texts, such as claims or full texts, using semantic analysis to build search queries to search related patents for validating the legal enforceability of patent claims. Mukherjea et al. (2005) created a semantic web of biomedical patents based on predefined semantic rules, relations, and vocabulary to aid in information retrieval and knowledge discovery in this area. Murphy et al. (2014) employed a bag-of-words model that considered only functional verbs to represent patents and design problems for conveniently drawing functional analogies from patents. Fu et al. (2013) extracted function (e.g., verbs) and surface (e.g., nouns) terms from patents to measure the analogical distance of patents from a design problem using Bayesian networks to inform the search of patent analogy.

Recently, word embeddings models, which map words or phrases to vectors of real numbers, are employed in semantic query expansion studies. Roy et al. (2016) integrated nearest-neighbours concept with semantic analysis. Diaz et al. (2016) showed that locally trained word embeddings models improve document retrieval performance compared to globally trained word embeddings models. Kuzi et al. (2016) used relevance models and word embeddings models together to improve document retrieval performance. Even though query expansion techniques based on word embeddings demonstrate promising retrieval performance, they mainly focus on increasing the recall performance by forming long queries, which introduces significant noises to both the query and the retrieved patent set. More importantly, these methods may also fail to include the extremely rare keywords related to the focal technology of actual search interests.

Despite the variety of existing patent search methods, they more or less incorporate keyword-based searches. However, the decision of which keywords to use for the search is subject to the searcher's expertise and knowledge and the bias toward better-known terms. Although one can expand and fine tune the keyword set for queries heuristically via reading the patent documents retrieved from a prior set of keyword searches over iterations, such a human reading process is tedious, labour-intensive, prone to human errors, and infeasible for patent retrieval of large systems or technology domains. On the other hand, while the query expansion studies present promising retrieval performance, both the query and retrieved document set may include noisy elements. Our work addresses these challenges by using computers to read patents for discovering additional keywords describing the focal technology.

3 METHODOLOGY

The method utilizes NLP techniques and a semantic engineering knowledge graph (EKG) to automatically extract keyword candidates from the patents retrieved in a prior search and recommend them according to their semantic similarity to the keywords already in the search query to expand the search query. The EKG lies in the heart of the method. It is comprised of the technically meaningful terms extracted from the texts of all patents in the USPTO database and the semantic similarity between them measured based on a word embeddings model trained on the entire patent database. Using the EKG, we aim to generate a medium containing an extensive set of engineering and technology related terms linked by corresponding quantified semantic relations. To date, the patent database which covers all fields of technology has not been utilized to build an ontology database by retrieving engineering and technology-related concepts and relations among them. Hereafter, we first introduce the overall method framework (Figure 1) and then the process of constructing the EKG using NLP techniques and the word2vec algorithm.

As a start, the searcher keys in a seed query consisting of some initial keywords to return a set of patents that contain any of the keywords in their titles or abstracts. Note that, we choose to focus the search on patent titles and abstracts that summarize the disclosed technologies concisely and precisely and avoid main texts and legal claims that contain broad, disguised and noisy texts irrelevant to the technologies themselves. Then, our method extracts word and phrase candidates from the texts of the set of obtained patents through a shallow term retrieval process inspired by the methodology of Rose et al. (2010). Specifically, it collects bi-grams, tri-grams, and four-grams which are delimited by stop-words as candidate phrases, and then selects those contained in the EKG as valid phrases (see Figure 2). The valid phrases and the other single words except the stop-words compose the candidate keyword list. This step is followed by ranking and recommending the candidate keywords according to their relevancies to the search query. The relevancy of a candidate keyword is calculated as the maximum of semantic similarities between the candidate keyword and any of the keywords composing the current search query, as formulated in equation (1)

$$rel_{C_i} = argmax(sim(C_i, q)|\forall q \in Q) \quad (1)$$

where C_i is the i^{th} candidate keyword, and Q is the set of the keywords composing the current search query. Then the searcher may select the relevant keywords from the candidate list to include them in the query for a new search iteration. The search process goes on to return an expanded patent set and thus a new list of candidate keywords extracted from them, for the selection by the searcher to further expand the query keyword set. The heuristic and iterative search process ends when the searcher finds no more relevant keywords from the candidate keyword list to expand the query or when the latest search iteration returns no more new patents and thus no new candidate keywords.

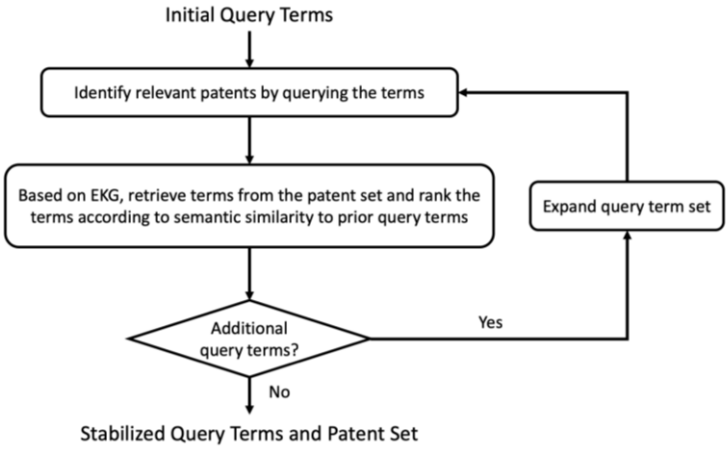

Figure 1. Overall framework for data-driven keyword based patent search

To create the EKG, we processed raw texts of the titles and abstracts of more than 5.7 million utility patents granted from 1976 to October 2017 in the entire USPTO patent database. We again focused on patent titles and abstracts to avoid the noisy information contained in the main texts and legal claims. First, we integrated and structured the texts of all titles and abstracts as a collection of lowercased sentences which are cleared off punctuation. This step yielded a pre-processed corpus of 26.75 million sentences. Then, we adopted a data-driven approach proposed by Mikolov et al. (2013a) to identify phrases by finding words that frequently co-occur and infrequently appear separately. Through such a process, we obtained a vocabulary of 14.1 million terms (words and phrases). Then, a denoising process was conducted to identify and correct noisy phrase formations and detect un-conventional stop words. Correcting the detected noisy phrase formations resulted in a vocabulary consisting of 6.6 million terms. Following a Part-of-Speech (POS) tagging step, we utilized WordNet lemmatizer (Fellbaum, 1998; Steven, Ewan and Edward, 2009) to unify multiple inflected forms of the words and phrases. Lastly, we filtered out the stop-words which were previously identified in the denoising process and are readily available in the Natural Language Toolkit (NLTK). This last step reduced the vocabulary to 5.16 million terms. Moreover, we further filtered out the terms that occurred less than twice in the overall corpus. The resulting vocabulary consists of 4,038,924 terms.

To further calculate the semantic similarity between the terms, we utilized the word2vec algorithm (Mikolov, et al., 2013b) to train a word embeddings model on the pre-processed corpus. The word2vec algorithm trains numeric vectors to represent each term in the vocabulary using a reduced form of neural network language model (NNLM)[1]. The algorithm trains word embeddings by maximizing the probability of correctly classifying terms in the context window of a given target term. The algorithm parses sentences sequentially, forms training samples by selecting target terms and the corresponding neighbouring words in a context window in sentences, and then trains the neural network model using these samples. Herein, we used a context window size of 20 because 90% of the sentences consist of less than 20 words. This guarantees that all terms in the sentence to which the target term belongs are treated as the context of the target term in 90% of the training samples. As a result of the training process, each term in the vocabulary is represented by a numeric vector, which enables the calculation of semantic similarity between the terms. In our case, the word2vec algorithm was tweaked and trained to represent all terms with numeric vectors of 150 dimensions. With these vectors, we calculated their pairwise semantic similarity using cosine similarity. In sum, the EKG consists of 4,038,924 terms[2] mined from the patent titles and abstracts, and 98.5% of their pairwise semantic similarities are larger than 0. We have developed a web-based interface, www.EKGraph.net, to enable the open uses and applications of the EKG. We also shared

[1] Mikolov *et al.* (2013b) modified the classical NNLM structure by removing the hidden layer and leaving only three layers, namely the input layer, the projection layer which directly maps the input layer and the output layer which employs a soft-max regression classifier.

[2] The vocabulary is shared in a GitHub repository (https://github.com/SerhadS/TKG)

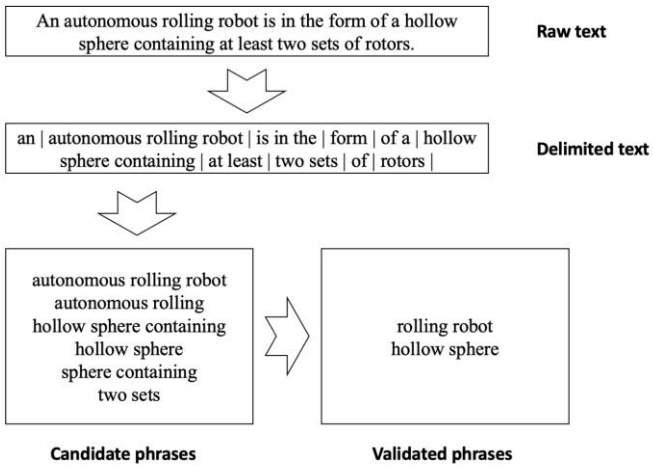

Figure 2. An example run of phrase extraction process

To facilitate the use of the EKG-based keyword discovery and patent retrieval method, we have implemented it in an online interface at www.SerKeys.com. "SerKeys" stands for "search for keywords". Figure 3 shows the screenshots of the iterations during a search process. To use the interface, searchers key in seed search keywords in the search box and click the "Search" button. A logic OR is used between the input keywords to generate the search query. The interface will return a list of terms extracted from the patents containing any of the query keywords and ranked by their relevancy scores (Equation 1), as the candidate keyword list. Searchers can browse the list from top to bottom to identify the relevant keywords and add them into the search box on top by double-clicking them. Then, searchers are suggested to click the "Search" button again to search with the expanded keyword set. Accordingly, the candidate keyword list will be updated with new terms arising from the expanded search query. Searchers can repeat the process of browsing the list and adding additional keywords to the search box for another iteration. When the candidate keyword list stops expanding, and no new keyword can be found, both the keyword set and the resultant patent set converge. The interface also reports the number of EKG-based candidate keywords and patents found via iterations. Especially, the searchers can click the hyperlink of the patent number to get the Google Patent URLs for these patents.

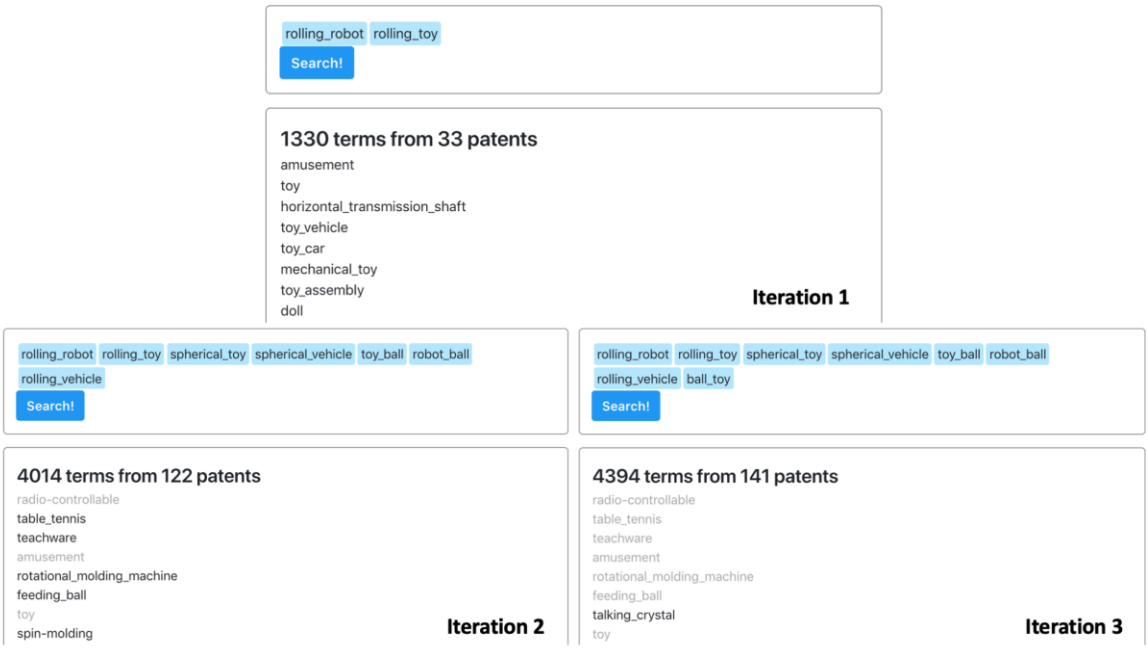

Figure 3. Screenshots from the keyword discovery and patent retrieval tool. Candidate keywords found in prior iterations are in grey, whereas new keywords are in black

4 CASE STUDY: SEARCH FOR SPHERICAL ROLLING ROBOT PATENTS

In this section, we demonstrate the proposed method via a case study of the patent search for spherical rolling robots (SRRs) from the USPTO patent database. SRRs are spherical shape robots that can propel themselves to roll around on the ground (Bicchi et al., 1997; Kim et al., 2016), and have been developed for defence and surveillance applications (Wu et al., 2017) and as toys and other consumer applications. An of example SRR is presented in Figure 4. An SRR patent set has been reported in a recent study (Song and Luo, 2017), which aimed to exhaustively retrieve patents related to SRRs from the USPTO database. Herein, we use the same case because of our rich experience in SRR patent search and for the convenient validation of the new EKG-based search keyword discovery and patent retrieval method.

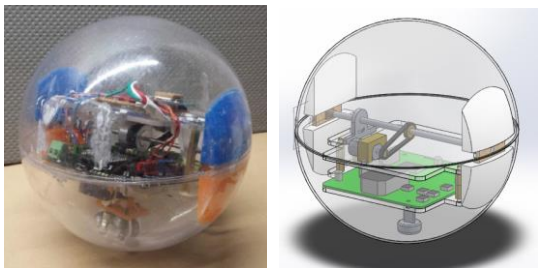

Figure 4. An example SRR (courtesy of Weng Ho Lok, Jin Sen Tan, Paiputra Fleming, Huishan Tan and Cheng Yuen Ong in the SUTD Engineering Design Innovation class in 2015)

We started with the seed search query used in the prior study (Song and Luo, 2017), including 2 keywords, "rolling robot" and "rolling toy", which were selected based on the knowledge and vocabulary of the searcher. The initial search returned 33 patents. From the 33 patents, the EKG-based method automatically extracted 1,330 terms (candidate keywords). These terms were automatically sorted according to their relevance to the search query, i.e., the maximum of their semantic similarities to any of the 2 initial query keywords. Scrolling down the candidate keywords list, we identified 5 relevant keywords to our interest to expand the search query, namely "spherical toy", "spherical vehicle", "toy ball", "robot ball" and "rolling vehicle". Thus, the query keyword set was expanded to 7 terms. We entered the second iteration of search with the new query and obtained 89 new patents, which contributed 2,684 new candidate keywords. From the updated candidate list, only 1 term, "ball toy", was identified as relevant to our interest (see Table 1). The search query set was expanded to include 8 terms. With the expanded query in the third iteration, 19 new patents were returned, resulting in 380 new candidates. We found no more new keywords that are relevant to our interest. The query set converged. In brief, starting with 2 seed query keywords, over 3 iterations, 6 additional keywords were identified for SRRs and 141 patents were retrieved in a few minutes. Screenshots in Figure 3 show these 3 iterations.

Table 1. The expansion of query terms and number of retrieved patents over 3 iterations

Iteration	Query terms	Number of new patents	Number of new terms	Number of new relevant terms
1	"rolling toy" "rolling robot"	33	1,330	5
2	"spherical toy" "spherical vehicle" "toy ball" "robot ball" "rolling vehicle"	89	2,684	1
3	"ball toy"	19	380	0
Total	**8**	**141**	**4,394**	**6**

As a comparison, we present the search results of the prior study of Song and Luo (2017) in Table 2. Their search started with the same seed keyword set, "rolling toy" and "rolling robot". Note that, they searched for patents with these 2 terms in the full texts, in contrast to our search only in patent titles and abstracts, and found 169 patents (see Table 2). They manually read the full texts of all these patents,

identifying only 10 relevant patents and 2 additional keywords, “spherical toy” and “spherical robot”. As a second iteration, they used the 2 new keywords to search for patents and found 93 additional patents. Again, by manually reading these patents’ full texts, they verified 13 relevant patents and discovered no more new keywords. To this point, the keyword-based search went through 2 iterations, identified 2 additional keywords and 23 relevant patents in total. The process took a few days since the searcher needed to read the returned patent texts to discover additional keywords. In comparison, the EKG-based method took only a few minutes to iterate until convergence and discover 6 new keywords. In particular, the 6 keywords found by the EKG-based method cover 1 of the 2 keywords discovered by keyword-based search in Song and Luo (2017), while “spherical robot” is missing because it appears in the description part instead of the title and the abstract. For further comparison, two researchers read the titles and abstracts of the 141 patents retrieved by the EKG-based method and identified 26 relevant patents. That is, the precision of the EKG-method, 18% (26/141), is higher than the precision of the human process, 9% (23/262). These 26 patents cover 11 of the 23 relevant patents that Song and Luo (2017) identified via reading the full texts of 262 patents, and also 15 additional patents missed by Song and Luo (2017). The EKG-based method missed the other 12 patents from Song and Luo (2017) because it only searches the titles and abstracts instead of full-texts of patents, and discovered 15 additional relevant patents as a result of the discovery of the 5 additional keywords, namely, “rolling vehicle”, “ball toy”, “toy ball”, “robot ball” and “spherical vehicle”.

Table 2. The expansion of query terms and retrieved patent set over 2 iterations in Song and Luo (2017)

Iteration	Query terms	Number of new patents	Number of new relevant patents	Number of new relevant terms
1	“rolling toy” “rolling robot”	169	10	2
2	“spherical toy” “spherical robot”	93	13	0
Total	**4**	**262**	**23**	**2**

Concerning human errors and cognitive limitations in reading patent texts to discover new relevant terms, Song and Luo (2017) went on to screen patents that cite or are cited by the 23 relevant patents directly and indirectly, and identified 113 relevant patents and 3 more keywords, including “spherical vehicle”, “ball toy” and “toy ball”, via reading the patent texts manually. Adding these keywords into the search query for a new iteration, they found 9 more relevant patents through keyword search and then 3 more through the citations of these 9 newly found patents. Through reading these 12 patents, no new keywords were discovered. To this point, they had found 148 patents and 5 new keywords relevant to SRRs. They went on again to screen the patents that share inventors with the 148 patents, obtaining 2 more relevant patents. These 2 patents did not lead to the discovery of new keywords and relevant patents through citations. In the end, their comprehensive search in patent texts, citation relations, and inventor information led to the discovery of 5 new keywords and 150 patents relevant to SRRs. The whole process took a few additional weeks.

In comparison, these 3 additional keywords discovered through searches via citation relations and inventor information in Song and Luo (2017) are all captured by our EKG-based method only through keyword-based search in a few minutes. Moreover, our method discovered 2 additional keywords, “rolling vehicle” and “ball robot”, which were not captured via human reading in their study. Furthermore, we also used “Google Autocomplete” and the term recommendation field of “Google Image Search” to discover new terms that are relevant to and extend from “rolling toy” and “rolling robot“. “Google Autocomplete” revealed 2 additional keywords, while “Google Image Search” revealed 1 additional keyword of the search interest. These comparisons suggest the EKG-based method may aid searchers in finding non-obvious keywords that describe the technology of interest and completing their keywords for queries for patent search, as well as general searches and data retrieval.

5 DISCUSSION AND CONCLUDING REMARKS

This paper aims to contribute to the development of patent search methods and tools via a data-driven search query term expansion methodology using NLP techniques and semantic analysis. The method

substitutes the tedious and error-prone human patent-reading efforts required to identify additional keyword terms with NLP using computers, and thus significantly improves the speed, comprehensiveness, and accuracy for patent retrieval. We demonstrated the proposed method using a case study of SRRs and compare the results with that reported in a prior study that aimed to exhaustively search for SRR patents for data-driven design inspiration (Song and Luo, 2017). The comparison validates the effectiveness of the method and highlights its advantage in mining and discovering potential keyword terms over manual reading and heuristics. We have implemented the method for public use at www.serkeys.com.

Despite the effectiveness of the general method, there are limitations regarding EKG. The data-driven phrase detection approach identifies as many statistically significant phrases as possible from the raw patent text. However, it also brought noises to the vocabulary. Meanwhile, although the denoising process reduced the vocabulary size and computational cost significantly by filtering out the identified noises, it is a human-dependent process and subject to human bias and error as well. Therefore, the data-driven phrasing approach should be tailored carefully to obtain a balance between more phrases and less noise. Another limitation is that EKG lacks varied representations of words with multiple meanings. For example, the word "mean" is represented with only one vector despite it has more than one meaning as a noun as well as a verb and an adjective. Thus, the trained vector of a word with multiple meanings would represent the most common use of the word. However, engineering terms, especially phrases, often each point to a very limited number of meanings, mostly unique concepts. Since phrases make up of 80% of our EKG according to our analysis, the noise contribution of multiple meaning issue may not cause significant distortion to the word embeddings model.

Moreover, the returned term list from the EKG may contain some terms that are relevant to the topic of search but cannot be used as standing-alone query terms. For instance, in our search for keywords to describe SRR, we found terms in the candidate term list such as “sphere”, “ball”, “spherical”, “robot”, “self-propelled” and “self-propelling”. Such terms can be combined together for queries. For instance, we tested a few AND queries based on such terms, e.g., {self-propelled AND sphere}, {self-propelled AND ball}, {self-propelling AND sphere}, {self-propelling AND ball}, {spherical AND robot} and {sphere AND robot}, and found additional relevant patents and new keywords from the returned patents, such as “ball toy”, “ball robot” and “robotic ball”. Therefore, it is beneficial to allow searchers (or use a computer algorithm) to combine and recombine the search terms in the recommended list to come up with new composite terms which are not directly from the EKG-based term list. In the next development of the system, we will add the function for such terms to be automatically combined into AND queries to further expand the search.

It is also noteworthy that the proposed method only focused on completing the set of keyword terms describing a focal technology. It does little to ensure the relevance of the retrieved patents. In our case study, only 18% (26 out of 141 patents) of the returned patents are validated as relevant. Despite being higher than the precision (9%) of the pure manual approach, there is great potential to explore algorithms, such as the one used in “Semantic Scholar”, to automate the verification process and build a more intelligent and precise patent search tool. In addition, incorporating additional data that contain technology and engineering design information, such as academic papers, to extend the EKG may further enhance the method for broader uses. Finally, we plan to carry out more case studies in diverse fields to establish more comprehensive understandings of the method for keyword discovery.

In addition to facilitating patent search, the EKG-based keyword discovery method also has other potential applications for engineering design aids. It can serve as a tool for designers to quickly explore the keyword terms that are related to a focal technology but differ from the original terms in the minds of the designers, without reading precedent documents such as patents and papers or other stimulations to human thinking. Such new terms may represent new concepts from other fields and provide inspiration to aid in design by analogy or synthesis. The EKG itself can also be useful for summarizing the key design concepts from lengthy design documents or brainstorming sessions and suggesting relevant concepts to expand the prior designs in the documents or the thought space covered in a brainstorming session. Hence, the proposed method may serve as the basis or engine for the potential development of computer-based tools that aid in engineering design analysis, ideation and concept generation.

REFERENCES

Alberts, D., Yang, C. B., Fobare-DePonio, D., Koubek, K., Robins, S., Rodgers, M. and DeMarco, D. (2011) Introduction to patent searching. In Current challenges in patent information retrieval. Springer Berlin Heidelberg. doi: 10.1007/978-3-642-19231-9_1.

Altshuller, G. S. and Shapiro, R. B. (1956) 'О Психологии изобретательского творчества (On the psychology of inventive creation)(in Russian).', Вопросы Психологии (The Psychological Issues), 6(Вопросы Психологии (The Psychol. Issues)), pp. 37–49.

Benson, C. L. and Magee, C. L. (2013) 'A hybrid keyword and patent class methodology for selecting relevant sets of patents for a technological field', Scientometrics, 96(1), pp. 69–82. doi: 10.1007/s11192-012-0930-3.

Bicchi, a., Balluchi, A., Prattichizzo, D. and Gorelli, A. (1997) 'Introducing the "SPHERICLE": an experimental testbed\nfor research and teaching in nonholonomy', Proceedings of International Conference on Robotics and Automation. doi: 10.1109/ROBOT.1997.619356.

Cascini, G. and Russo, D. (2006) 'Computer-Aided analysis of patents and search for TRIZ contradictions', International Journal of Product Development, 4(1–2), pp. 52–67. doi: 10.1504/IJPD.2007.011533.

D'hondt, E. (2009) 'Lexical issues of a syntactic approach to interactive patent retrieval', The Proceedings of the 3rd BCSIRSG Symposium on …, pp. 102–109. Available at: http://lands.let.kun.nl/literature/dhondt.2009.1.pdf.

Diaz, F., Mitra, B. and Craswell, N. (2016) 'Query Expansion with Locally-Trained Word Embeddings', in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 367–377. doi: 10.18653/v1/P16-1035.

Fellbaum, C. (1998) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Fu, K., Cagan, J., Kotovsky, K. and Wood, K. (2013) 'Discovering Structure in Design Databases Through Functional and Surface Based Mapping', Journal of Mechanical Design, 135(March 2013), p. 031006. doi: 10.1115/1.4023484.

Fu, K., Murphy, J., Yang, M., Otto, K., Jensen, D. and Wood, K. (2014) 'Design-by-analogy: experimental evaluation of a functional analogy search methodology for concept generation improvement', Research in Engineering Design, 26(1), pp. 77–95. doi: 10.1007/s00163-014-0186-4.

Fujii, A. (2007) 'Enhancing patent retrieval by citation analysis', Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07, pp. 793–794. doi: 10.1145/1277741.1277912.

Gerken, J. M. and Moehrle, M. G. (2012) 'A new instrument for technology monitoring: Novelty in patents measured by semantic patent analysis', Scientometrics, 91(3), pp. 645–670. doi: 10.1007/s11192-012-0635-7.

Graf, E., Frommholz, I., Lalmas, M. and Van Rijsbergen, K. (2010) 'Knowledge modeling in prior art search', Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6107 LNCS, pp. 31–46. doi: 10.1007/978-3-642-13084-7_4.

He, Y. and Luo, J. (2017) 'The novelty "sweet spot" of invention', Design Science. Cambridge University Press, 3, p. e21. doi: 10.1017/dsj.2017.23.

Jeong, Y., Lee, K., Yoon, B. and Phaal, R. (2015) 'Development of a patent roadmap through the Generative Topographic Mapping and Bass diffusion model', Journal of Engineering and Technology Management - JET-M. Elsevier B.V., 38, pp. 53–70. doi: 10.1016/j.jengtecman.2015.08.006.

Kim, G., Kwon, Y., Suh, E. S. and Ahn, J. (2016) 'Analysis of Architectural Complexity for Product Family and Platform', Journal of Mechanical Design, pp. 1–11. doi: 10.1115/1.4033504.

Koch, S., Bosch, H., Giereth, M. and Ertl, T. (2011) 'Iterative integration of visual insights during scalable patent search and analysis', IEEE Transactions on Visualization and Computer Graphics, 17(5), pp. 557–569. doi: 10.1109/TVCG.2010.85.

Kuzi, S., Shtok, A. and Kurland, O. (2016) 'Query Expansion Using Word Embeddings', in Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM '16. New York, New York, USA: ACM Press, pp. 1929–1932. doi: 10.1145/2983323.2983876.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) 'Efficient Estimation of Word Representations in Vector Space'. Available at: http://arxiv.org/abs/1301.3781 (Accessed: 26 November 2018).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013) 'Distributed Representations of Words and Phrases and their Compositionality', pp. 3111–3119. Available at: http://papers.nips.cc/paper/5021-distributed-representations-of-words-andphrases (Accessed: 26 November 2018).

Montecchi, T., Russo, D. and Liu, Y. (2013) 'Searching in Cooperative Patent Classification: Comparison between keyword and concept-based search', Advanced Engineering Informatics. Elsevier, 27(3), pp. 335–345. doi: 10.1016/J.AEI.2013.02.002.

Mukherjea, S. (2005) 'Information retrieval and knowledge discovery utilising a biomedical Semantic Web', Briefings in Bioinformatics, 6(3), pp. 252–262. doi: 10.1093/bib/6.3.252.

Murphy, J., Fu, K., Otto, K., Yang, M., Jensen, D. and Wood, K. (2014) 'Function Based Design-by-Analogy: A Functional Vector Approach to Analogical Search', Journal of Mechanical Design, 136(10), p. 101102. doi: 10.1115/1.4028093.
Nakamura, H., Suzuki, S., Sakata, I. and Kajikawa, Y. (2015) 'Knowledge combination modeling: The measurement of knowledge similarity between different technological domains', Technological Forecasting and Social Change. Elsevier Inc., 94, pp. 187–201. doi: 10.1016/j.techfore.2014.09.009.
Rose, S., Engel, D., Cramer, N. and Cowley, W. (2010) 'Automatic Keyword Extraction from Individual Documents', in Text Mining. Chichester, UK: John Wiley & Sons, Ltd, pp. 1–20. doi: 10.1002/9780470689646.ch1.
Roy, D., Paul, D., Mitra, M. and Garain, U. (2016) 'Using Word Embeddings for Automatic Query Expansion'. Available at: http://arxiv.org/abs/1606.07608 (Accessed: 4 December 2018).
Song, B. and Luo, J. (2017) 'Mining Patent Precedents for Data-Driven Design : The Case of Spherical Rolling Robots', Journal of Mechanical Design, 139(11), p. 111420. doi: 10.1115/1.4037613.
Song, B., Luo, J. and Wood, K. L. (2018) 'Data-Driven Platform Design: Patent Data and Function Network Analysis', Journal of Mechanical Design, In Press.
Srinivasan, V., Song, B., Luo, J., Subburaj, K., Elara, M. R., Blessing, L. and Wood, K. (2018) 'Does Analogical Distance Affect Performance of Ideation?', Journal of Mechanical Design. American Society of Mechanical Engineers, 140(7), p. 071101. doi: 10.1115/1.4040165.
Steven, B., Ewan, K. and Edward, L. (2009) Natural Language Processing with Python. O'Reilly Media Inc.
Takaki, T., Fujii, A. and Ishikawa, T. (2004) 'Associative document retrieval by query subtopic analysis and its application to invalidity patent search', ACM Conference on Information and Knowledge Management, pp. 399–405. doi: 10.1145/1031171.1031251.
Wang, S.-J. (2011) 'The state of art patent search with an example of human vaccines', Human Vaccines, 7(2), pp. 265–268. doi: 10.4161/hv.7.2.14004.
Wu, F., Vibhute, A., Soh, G. S., Wood, K. L. and Foong, S. (2017) 'A compact magnetic field-based obstacle detection and avoidance system for miniature spherical robots', Sensors (Switzerland), 17(6). doi: 10.3390/s17061231.
Xue, X. and Croft, W. B. (2009) 'Automatic query generation for patent search', in Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09. New York, New York, USA: ACM Press, p. 2037. doi: 10.1145/1645953.1646295.

ACKNOWLEDGMENTS

The research is funded by the SUTD-MIT International Design Centre.