Cross-domain Knowledge Discovery based on Knowledge Graph and Patent Mining

Fan Ye*, Tie Fu, Lin Gong and Jun Gao
School of Mechanical Engineering, Beijing Institute of Technology, Beijing, China

*Corresponding author e-mail: 3120180417@bit.edu.cn

Abstract. This paper studies an approach on cross-domain knowledge discovery to assist the conceptual stage of the design process related to mechanical engineering. Variable methods and tools are proposed to obtain knowledge within a given domain until now. However, methods on cross-domain knowledge analysis is under-developed. In this paper, domain knowledge graph is built automatically by employing natural language process (NLP) and patent mining. They comprise patent documents obtaining and knowledge extraction. Then according to the international patent classification (IPC), the knowledge elements are divided to some different categories. The elements are stored in Databases and then the given domain knowledge graph is constructed after correlation analyses. The cross-domain knowledge surrounding the given domain knowledge is found by mining the correlation among cross-domain knowledge. The cross-domain knowledge can inspire designers about new design of a given domain. And the efficiency of knowledge reusing can be improved by domain knowledge graphs.

Keywords: Cross-Domain, Knowledge Graph, Natural Language Process (NLP), Patent Mining

1 Introduction
Knowledge is the source of innovation. Patents is an important carrier of knowledge development and innovation. Especially invention and utility model patents contain a lot of product design knowledge. Effective use of patent knowledge can shorten R&D time by 60% and save R&D costs by 40% [1]. Therefore, people pay more attention to economic benefits brought by the knowledge. As patent documents increase, it is difficult for people to acquire the new technology or design knowledge in patents roundly and timely. Employing text mining technology, both the structured and unstructured information can be automatically extracted from patent documents [2]. People can analyze and research these data to provide design knowledge for the conceptual design process.

Conceptual design is an important stage in product design, but the traditional conceptual design process always relies on the professional background and subjective experience of designers. People understand the most important elements in the conceptual design are functional and structural elements by learning some classical design theories. People use computer-aided ways to extract design elements. Scholars have proposed various methods or tools to achieve the extraction of valuable design knowledge
from patents until now. Yoon B, Park Y [3] proposed a network-based patent analysis method. The network is used to analyze the internal structure of the patent with some quantitative analysis. Yang SY et al. [4] proposed to convert claims into conceptual graphs based on syntactic and semantic information. However, they are too complicated and targeted. The study deals with knowledge graph to discover the design knowledge.

2 Related Models

2.1 BERT Model
In 2018, Google proposed a language representation model, called BERT, which stands for Bidirectional Encoder Representations from Transformers [5]. BERT model uses a multi-layer bidirectional Transformer encoder as the core structure of the model. The innovation of the BERT model lies in the pre-training process, which uses masked language model and next sentence prediction to capture word or sentence level relationship features. BERT model has two model sizes, BERT-BASE, BERT-LARGE. BERT-BASE is composed of 12 stacked attention layers, 12 self-attention heads, and a 768-dimensional hidden layer. Through the pre-training process, the model learns grammatical and semantic information of texts, and then uses the labeled data from downstream tasks to fine-tune all parameters. Experiments on several NLP tasks prove that BERT model has good performance.

2.2 Word2Vec Model
Word2Vec is a classical distributed word vector model [6]. The model can use algebraic operations on embedding vectors corresponding to the words to capture the similarity between words. Word2vec model contains two training model architectures, including continuous bag-of-words (CBOW) and continuous skip-gram model [7]. The input of CBOW model is the context of the target word, and the embedding of the target word is predicted by their embedded representation, whereas the skip-gram model is just the opposite. The embedding vectors are trained quickly by Word2Vec model and are used to calculated semantic similarity better.

2.3 Pointwise Mutual Information (PMI)
PMI [8] is employed to measure the related information theoretical association. Herein, PMI can be used to indicate the association of two nodes in the domain knowledge graph, called the structural similarity. If a pair of nodes shows a specific combination habit in the context, then the nodes will get a high PMI value. It is defined as:

\[
PMI(w_i, w_j) = \log_2 \left( \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \right)
\]

where \(p(w_i)\) is a marginal probability of the node \(w_i\) in the patents belonging to a given domain; \(p(w_i, w_j)\) is a joint probability of the nodes \(w_i\) and \(w_j\) in the patents belonging a given domain.

3 Methodology
The research employs NLP to construct domain knowledge graphs and explore the correlation between different fields of patent knowledge. Specifically, the functional and structural elements are extracted from unstructured data, the title and abstract in the patents, by NLP. The knowledge elements are divided to several domains based on the classes of IPC. Then domain knowledge graphs are built by nodes and edges modeling. Finally, the correlations among the graphs are discovered according to the relationship of nodes from different graphs.

3.1 Knowledge Elements Extraction
Chinese patents have three categories, including invention, design and utility patents. The invention and utility patent documents are used to acquire practical design knowledge for assisting conceptual design.
The paper focuses on patent titles and abstracts. Data Source is Chinese patents from CNKI. Patent documents are crawled by web crawlers and the titles and abstracts of patents are acquired by regular expressions.

Functional and structural knowledge elements are extracted from the original unstructured patent data using NLP. The process of extracting the elements is actually to extract words or phrases on semantics, and then classify them. With the development of NLP, scholars have proposed Word2Vec [7], GloVe [9], ELMo [10] and BERT [5] models to quantitatively describe text semantics. The models represent text as a series of vectors that express the semantics of the text. Word2Vec and GloVe models project text into a vector. They have high calculation efficiency, but are not good at learning the context of a word. The vector representation of the same word in different contexts would be the same. ELMo employs bidirectional long short-term memory (LSTM) to learn word features. BERT uses deep bidirectional Transformers to strengthen the generalization ability of embedding vectors. It describes the character level, word level, sentence level and the relationship between sentences, and learns the semantic and context characteristics of words fully. Therefore, in this study, BERT model is used to identify functional and structural knowledge elements.

Patents text is disassembled into a single character, and part of the data is selected to be manually labeled with a BIO scheme. Next, elements recognition process is achieved by BERT model. Training corpus is first converted into embedding vectors. They are the sum of token, segment and position embeddings as input features of the model. An encoder encodes the embeddings by self-attention mechanism. The model learns the positional, referential and semantic relations of tokens through 12 stacked attention layers. The relations are stored into embedding vectors as the output. The transformer from texts to vectors is achieved.

### 3.2 Domain Knowledge Graph Building

Based on knowledge recognition, the recognized patent text is extracted and divided into two dimensions, functional and structural elements. Then they are divided into several domains according to the classes of IPC. A domain knowledge graph is constructed corresponding to the classified domain, in which the extracted functional and structural knowledge elements are nodes and PMI values are calculated as quantitative representation of edges.

First of all, the patent classification numbers are obtained by the crawler. The numbers are matched by regular expressions to obtain the class symbols of IPC, which is a complex hierarchical classification system consisting of sections, classes, subclasses and groups. For example, "B25J15/00" is a patent classification number, where "B25" is the class symbol. According IPC, "B25" stands for "hand tools; portable power-driven tools; handles for hand implements; workshop equipment; manipulators". Herein, the class of IPC stands for a domain. Then all patent knowledge elements are assigned to their corresponding domain and they are in two dimensions, functional and structural dimension. But the knowledge elements without links are difficult to be understood. In general, semantic network achieves link prediction through analysis and reasoning to find the implicit relationship between nodes. However, it is difficult to explain and display the relationships between functions and structures. The relationships are hoped to be expressed quantitatively. Therefore, using PMI values as weights of edges can simplify the representation of links between nodes. The relevant knowledge elements in the same domain are connected based on quantitative links. Thus, a domain knowledge graph is constructed.

### 3.3 Cross-domain Patent Knowledge Association Discovery

The functional and structural knowledge elements of a given domain and their relationships are obtained from the domain knowledge graph. However, the relationship of knowledge elements among multiple domains could not be expressed. Different domains may have the same or similar functions, which have different structures; or they have the same structure, but these structures play different roles. The semantic correlation analysis of nodes and structural correlation of graphs is to find the relationship among knowledge elements in different domains. These cross-domain knowledge...
elements could become a source of inspiration of new design. Specially, the Word2Vec model is used to train embedding vectors instead of the BERT model, because the former can represent vectors by lower dimension which is benefit to the calculation of semantic similarity.

The paper supposes the top \( n \) highest frequency nodes of a given domain \( D_a \) form a set, \( \text{Set}_a = \{ w_1, w_2, ..., w_n \} \), where \( n \) is less than number of all nodes in domain \( D_a \). In order to seek the related nodes from other domains, for each node from the other domain, such as \( D_b \), its semantic relevance score (\( R_{\text{cross}} \)) is calculated as its weighted average semantic similarity to terms of \( \text{Set}_a \):

\[
R_{\text{cross}} = \frac{1}{n} \sum_{i=1}^{n} \text{sim}(w_i, v_j) f_i
\]

\[
r_{\text{cross}} = \frac{R_{\text{cross}}}{\max(R_{\text{cross}})}
\]

In (2), \( w_i \) is the node in \( \text{Set}_a \); \( v_j \) is the node in some other domain; \( \text{sim}(w_i, v_j) \) stands for the semantic similarity of the two nodes; and \( f_i \) is weighting factor, the frequency of the \( i \) th node in \( \text{Set}_a \). In (3), \( \max(R_{\text{cross}}) \) is the maximum of \( R_{\text{cross}} \) for each node from another domain. The \( R_{\text{cross}} \) is standardized by (3).

Based on the semantic relevance score, \( m \) nodes in \( D_b \) are obtained, that are most relevant to \( \text{Set}_a \), as a set, \( \text{Set}_b = \{ v_1, v_2, ..., v_m \} \). Its structural relevance score (\( R_s \)) is calculated as its weighted average PMI value to other nodes around terms in \( D_b \):

\[
R_s = \frac{1}{m} \sum_{j=1}^{m} \text{PMI}(v_j, v_k) \times R_{\text{cross}} / \max(\text{PMI})
\]

where \( m \) is the number of \( \text{Set}_b \) terms; \( v_j \) is a node in \( \text{Set}_b \); \( v_k \) is a node in \( D_b \) apart of nodes in \( \text{Set}_b \); \( R_{\text{cross}} \) is the weighted average structural similarity of node \( k \) except for nodes in \( \text{Set}_b \); \( R_{\text{cross}} \) is the weighting factor here; \( \max(\text{PMI}) \) is maximum of PMI values in \( D_b \).

### 4 Case Study

#### 4.1 Data Source

The case retrieved "machine" in Chinese on the CNKI. There are nearly 10000 pieces of patent information. And public information on the corresponding webpage was crawled, including: title, abstract, publication number and classification number etc. According to the publication number of patent documents, regular expressions were used to filter out design patent documents, leaving 9942 pieces of patent data as the main data source. They were stored in a relational database.

#### 4.2 Experiment

The case chose 100 pieces of patent data from the database randomly. The extra of the text needed to be added to a new sentence because of BERT model limits on sentence length. Then the knowledge elements were labeled in the data manually. The labeling tags include: B-S, I-S, B-F, I-F, O, where “B-” means the beginning of the word, “I-” means the middle of the word, “O” means not related word, “S” and “F” means structure and function. The data annotation format was as Table 1, where the Chinese sentence means that “The tongs acts as a clamp on the outer wall of the thing to be clamped” in Chinese.

| Text          | … | The | tongs | acts | as | a | clamp | on |
|---------------|---|-----|------|------|----|---|-------|----|
| Label         | O | O   | B-S  | O    | O  | O | B-F   | O  |

| Text          | the | outer | wall | of | the | thing | to | … |
|---------------|-----|-------|------|---|-----|-------|----|----|
| Label         | O   | B-S   | I-S  | O | O   | B-S  | O  | O  |

First, the case employed unlabeled titles and abstracts to superimpose pre-training on the basis of the base model for improving the ability to recognize required entities. Then labeled data was used to
construct pre-training model based on the superimposed model. Its parameters were consistent with the default parameters of BERT model.

Three domains which contain over 500 patent documents were obtained, including "B07", "B23" and "B65". The model was successful to recognize and classify functional and structural elements. Table 2 shows the evaluation of the BERT model for the test set.

This study chose B07 and B23 as the research objects. Domain B07 has 548 functional and 4226 structural nodes; Domain B23 has 757 functional and 4460 structural nodes. The nodes with the top 10 highest frequencies in Domain B07 are reported in Table 3, where Node represents the nodes in Domain B07; f is the frequency of the node, kind is the kind of the node. The nodes are translated from Chinese into in Table 3. According to (2) ~ (4), 20 nodes were obtained in Domain B23 that are most relevant to the 10 nodes in Domain B07 in Table 4, where Node represents the nodes in Domain B23; R is the relevance between the node in B27 and the nodes in Table 3; Kind is the same as the kind in Table 3. In total, the 20 nodes are from cross-domain. They could make designers have a concise understanding of this domain. Designers could obtain the concepts of structure or function in different domains, and expand the design direction.

### Table 2. The evaluation of BERT for test set

| Kind | Precision | Recall | F1-score |
|------|-----------|--------|----------|
| S    | 0.91      | 0.91   | 0.91     |
| F    | 0.71      | 0.80   | 0.75     |

### Table 3. Top 10 terms occurring most frequently in domain B07

| Node  | f  | Kind | Node        | f  | Kind |
|-------|----|------|-------------|----|------|
| sort  | 0.66 | F    | automation  | 0.18 | F    |
| set   | 0.51 | F    | sorting device | 0.16 | S    |
| connect | 0.38 | F    | control     | 0.14 | F    |
| install | 0.24 | F    | rack        | 0.14 | S    |
| detect | 0.20 | F    | fix         | 0.12 | F    |

### Table 4. Top 10 terms most relevant in domain B23

| No.2 clamp nut | Rs | Kind | Node         | Rs | Kind |
|----------------|----|------|--------------|----|------|
| brake          | 0.4165 | S    | steering knuckle | 0.4165 | S    |
| forward lead   | 0.4165 | S    | screw sleeve  | 0.4164 | S    |
| groove         | 0.4165 | F    | screw shaft   | 0.4164 | S    |
| adjusting nut  | 0.4165 | S    | sliding block | 0.4162 | S    |
|                |       |      | material holding | 0.4162 | S    |

### 5 Conclusion

This paper focuses on cross-domain knowledge discovery based on knowledge graph and patent mining. In a summary, there are two major contributions as followings: (1) Domain knowledge graph is built by unstructured data recognition and extraction automatically. Using domain knowledge graph provides simple and comprehensive design knowledge for the conceptual design process and improves the efficiency of knowledge reusing; (2) The cross-domain knowledge is discovered through correlation analysis of knowledge elements in different domains. The cross-domain knowledge could stimulate the creative thinking and facilitate the design process.

It still needs improvement in some aspects. First, the rule of domain division is not precise enough. The division rule needs to be further refined, such as increase the design object as the division rule. In addition, the patent knowledge is divided into two dimensions: function and structure. However, the functions and structures of a product always contain multiple levels. It is necessary to establish a multi-level mapping framework about functions and structures by construct sub-functions or
sub-structures. More completed conceptual designs are provided to designers with more details. These two points are worth exploring in the future.

Acknowledgments
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).

References
[1] Wang Z X, Qiu Q Y, and Feng P E. Patent knowledge mining for conceptual design, Journal-Zhejiang university engineering science, vol. 42, no. 3, pp. 522-527, 2008. (in Chinese)
[2] Kodratoff, Yves. Knowledge discovery in texts: a definition, and applications, International Symposium on Methodologies for Intelligent Systems, Springer, Berlin, Heidelberg, 1999, pp. 16-29.
[3] Yoon B, Park Y. A text-mining-based patent network: Analytical tool for high-technology trend, The Journal of High Technology Management Research, vol. 15, no. 1, pp. 37-50, 2004.
[4] Yang S Y, Soo V W. Extract conceptual graphs from plain texts in patent claims, Engineering Applications of Artificial Intelligence, vol. 25, no. 4, pp. 874-887, 2012.
[5] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.
[6] Zhang D, Xu H, Su Z, et al. Chinese comments sentiment classification based on word2vec and SVMperf, Expert Systems with Applications, vol. 42, no. 4, pp. 1857-1863, 2015.
[7] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space, 2013.
[8] Bouma G. Normalized (pointwise) mutual information in collocation extraction, Proceedings of GSCL, pp. 31-40, 2009.
[9] Pennington J, Richard S, and Manning C. Glove: Global vectors for word representation, Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014.
[10] Peters M E, Neumann M, Iyyer M and et al. Deep contextualized word representations, 2018.