Advanced Engineering Informatics 42 (2019) 100959

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

Full length article

An automatic literature knowledge graph and reasoning network modeling framework based on ontology and natural language processing

Hainan Chen[a,b,c], Xiaowei Luo[b,c,*]

[a] School of Intelligent Systems Engineering, Sun Yat-sen University, China
[b] Architecture and Civil Engineering Research Center, Shenzhen Research Institute, City University of Hong Kong, Hong Kong
[c] Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong

ARTICLE INFO

Keywords:
Representation ontology
Natural language processing
Knowledge graph
Knowledge reasoning

ABSTRACT

With the advancement of scientific and engineering research, a huge number of academic literature are accumulated. Manually reviewing the existing literature is the main way to explore embedded knowledge, and the process is quite time-consuming and labor intensive. As the quantity of literature is increasing exponentially, it would be more difficult to cover all aspects of the literature using the traditional manual review approach. To overcome this drawback, bibliometric analysis is used to analyze the current situation and trend of a specific research field. In the bibliometric analysis, only a few key phrases (e.g., authors, publishers, journals, and citations) are usually used as the inputs for analysis. Information other than those phrases is not extracted for analysis, while that neglected information (e.g., abstract) might provide more detailed knowledge in the article. To tackle with this problem, this study proposed an automatic literature knowledge graph and reasoning network modeling framework based on ontology and Natural Language Processing (NLP), to facilitate the efficient knowledge exploration from literature abstract. In this framework, a representation ontology is proposed to characterize the literature abstract data into four knowledge elements (background, objectives, solutions, and findings), and NLP technology is used to extract the ontology instances from the abstract automatically. Based on the representation ontology, a four-space integrated knowledge graph is built using NLP technology. Then, reasoning network is generated according to the reasoning mechanism defined in the proposed ontology model. To validate the proposed framework, a case study is conducted to analyze the literature in the field of construction management. The case study proves that the proposed ontology model can be used to represent the knowledge embedded in the literatures' abstracts, and the ontology elements can be automatically extracted by NLP models. The proposed framework can be an enhancement for the bibliometric analysis to explore more knowledge from the literature.

1. Introduction

Along with the transformation from an industrial society to an information society starting from the 1970s, knowledge becomes one of the key assets for the organizations. In the information society, the creation, distribution, utilization, and manipulation of information become important works, making the importance of intellectual work significantly increased. There is a consensus that effective knowledge management can bring an organization competitive advantage over its rivals [1]. When researchers and practitioners paid much of their attention to the knowledge management to explore its value, more and more documentation has been accumulated to record the generated knowledge. Due to human's limited cognitive capacity, it is a great challenge to retrieve and obtain the knowledge from the accumulated documentation in a human-friendly and efficient way.

According to epistemology, there exist two perspectives on the knowledge related principles for a consensus: objectivist perspective and practice-based perspective [1]. From the objectivist perspective, knowledge is regarded as an entity that can be codified and kept independently among people who possesses and uses it. In contrast, the practice-based perspective assumes that knowledge is embedded in, developed through, and related to people's workplaces and practices [2,3]. The practice-based knowledge can be translated into objective knowledge to some extent through generalizing, summarizing, and abstracting. Literature is one of the most common sources of objective knowledge, as it generally presents the facts with clear and systematical discourses. The methodologies and experiments presented in the literature are reproducible, indicating that embedded knowledge can be

* Corresponding author.
E-mail address: xiaowluo@cityu.edu.hk (X. Luo).

https://doi.org/10.1016/j.aei.2019.100959
Received 26 April 2019; Received in revised form 28 May 2019; Accepted 25 June 2019

transmitted independently. More and more literature is accumulated in the journal databases, offering a valuable way for knowledge management. Up to now, the Web of Science core collection database indexes more than 20,000 journals and 1.4 billion references data; ScienceDirect database indexes over 250,000 academic articles which can be openly accessed; Scopus indexes over 69 million article records and over 1.4 billion references data, which are extremely large knowledge bases.

The accumulated knowledge in a large volume not only significantly enhances science and technology research, but also offers strong support for industry practice and education. Powerful search engines have been designed to support fast retrieval for given reference items or related topic keywords. However, since all the academic literature is unstructured data, the data retrieval mainly relies on the similarity matching of the text features. The similarity matching mechanism would retrieve hundreds and thousands of records for a given keyword, while most of them are unintended records simply because those records contain the text of the given keyword. It is quite a challenging work to construct the reasoning network with unstructured data, in which most of the valuable knowledge are embedded. Without an appropriate reasoning network, it would be quite time-consuming to retrieve the knowledge when needed. Citation data, along with the development of information theory and literature management, is defined as structured features to represent and identify literature. However, the citation data only contains the authors' names, affiliations, source journals of the paper, and very limited knowledge embedded in the title. The paper's abstract, a concise and powerful statement describing the works in the paper, could provide more information about the paper. An abstract usually contains the background, objectives, solutions, and findings of the work. If the abstract data can be characterized into these four aspects, the unstructured abstract data can be presented by structured data, which tells in what background, who targeted at what objectives, used what solutions, and discovered what findings. This structured data can offer fundament knowledge to construct a reasoning network. The key challenges now become how to automatically extract these four aspects from the abstract, and how to connect the tetrad to construct a reasoning network.

Therefore, in addition to the conventional representation of citation data, the abstract data is processed in this study using natural language processing (NLP) technology. An ontology model based on the tetrad structure (background, objective, solution, finding) is proposed to interpret abstract data of the literature. Key phases for each item in the tetrads are extracted as features, which can be quantified using word vectors. A reasoning network is constructed by connecting the tetrads through calculating the correlation and difference. Therefore, based on the citation and abstract data, each knowledge entity in the reasoning network contains six aspects: authors, affiliations, background, objective, solution, and findings. With the constructed reasoning network, the following knowledge can be deduced: (1) for a given background, what objectives can be targeted; (2) for a given objective, what the solutions are and how they perform; (3) for a given objective, who would be the potential experts and organizations to handle the problems. Thus, the values of literature would be optimized.

The remaining of this paper is organized as follows: Section 2 presents and summarizes the background and related works of this study. The proposed automatic knowledge graph and reasoning network modeling framework based on ontology and NLP is presented in Section 3. In Section 4, a case study with the research field of "construction management" is conducted to demonstrate and validate the proposed framework. Section 5 discusses the findings in the case study. Finally, the conclusions and limitations of the study are summarized in Section 6.

2. Research background

2.1. Development of knowledge management

Knowledge is an important asset to organizations. The management of knowledge has aroused widespread attention. Through reviewing the development of the knowledge-based system, the knowledge-based modeling and manipulation can be divided into four categories: (1) ontology; (2) cognitive knowledge base; (3) linguistic knowledge base, and (4) expert knowledge base [4]. The ontology-based models characterize the knowledge as clusters of concepts with corresponding attributes, values, and relations [5]. To define the ontology as a formal and explicit specification of a shared conceptualization, three elements are required: domain concepts, relations, and instances [6]. As domain concept is one of the required elements, the ontology-based model is domain specified and has explicit meanings of the identified concepts. Human knowledge in long-term memory can be represented by an object-attribute-relation model based on the synaptic structure of human memory [7]. For cognitive knowledge-based models, the knowledge is represented by a structure of a dynamic concept network [8]. A concept is a cognitive unit, rather than a cluster of instances in the ontology-based models. The cognitive unit corresponds to a concrete entity in the real-world and an abstract entity in the perceived-world [9]. The cognitive knowledge base is flexible for learned knowledge synergy, but can also be transformed among different knowledge bases. The linguistic and expert knowledge bases are conventional technologies for knowledge modeling and manipulation. The linguistic theories attempt to model human grammar and to divide the human knowledge of grammar into phonology, morphology, syntax, semantics, and lexicon [10]. It is quite intuitive, and the knowledge extraction is efficient provided clear rules. However, along with the growth of the lexicon size, it is quite challenging to capture the full structures [7]. For expert knowledge bases, the knowledge is represented as a set of rules extracted from experts' experiences [11]. The logical and fuzzy rule-based methods are the two primary approaches to establish the expert knowledge base [12]. It is computation and reasoning efficient to develop the expert knowledge base, since the rules have been pre-defined. However, the applications are limited to the expert's domain, and it is hard to transform the expert knowledge base across different domains because of its strong reliance on experts' experiences.

The most common knowledge management tool is the search engine, including the web search engine, academic literature retrieval engine, and the data and document query system inside the organizations. Knowledge retrieval is one primary type of knowledge manipulations. With the advancement of information technologies, the search engine can retrieve data efficiently nowadays. The search engine is based on the similarities and connections strength, rather than the logical relations of the knowledge elements [13]. PageRank algorithm is widely used in search engines [14], and fuzzy search is the dominant method [15]. Consequently, a large amount of unrelated information is retrieved simply because of its phrase similarity. This limitation leads to the exploration in the knowledge graph, which was first proposed by Google to enhance its search engine with semantics [16]. Knowledge graph encodes the semantic information of the knowledge entities, of which the dependency and causality are considered other than the physical connection strength [17]. This idea is used in various knowledge bases or knowledge management applications, including DBpedia, YAGO, Freebase, Wikidata, Spark, Google's Knowledge Vault, Microsoft's Satori and Facebook's entity graph [17–19]. While the knowledge graph requires annotated sources at large scales, the sources should be represented by formatted or formally structured data. Therefore, most of the existing knowledge graphs are established based on Wiki or web pages data, which offers generic knowledge. The generic knowledge graph can be transformed to build industry-specific knowledge graph. Since the industry data are generally un-structured but formatted data,

and most of the concepts in the industry domain are not included in the generic knowledge graphs, the building of industry knowledge graph is quite labor-intensive and time-consuming.

The development of knowledge graph for scientific knowledge mining and management has obtained significant progress, because of the well-formatted citation and description data. Without additional annotation, the connections among the academic literature are measured by the cited and co-cited values. Then a graph is built to represent the knowledge connection of a given topic [20]. The temporal citation frequency is used to discover the research burst; the betweenness centrality is used to detect the significant research, and the variation of citations along time period is employed to evaluate the research trend. Correspondingly, the researchers and organizations can also be deduced according to the citation data [21]. To facilitate the bibliometric science mapping, visualization and automatic tools have been developed, including CiteSpace, HistCite, and VOSviewer [22]. Utilizing the citation data to quantify the connections of the literature is objective and clear, and it is efficient to conduct the trend analysis. The relations among the primary topics contained in the literature can be evaluated using citation analysis together with keyword extraction. However, the citation data only contains the identification information and very limited knowledge about the paper's contents. Thus, the built graph can only support superficial knowledge reasoning.

In summary, knowledge extraction and representation are critical tasks for knowledge management. Graphs are the most promising and widely used tool to represent knowledge. While the knowledge extraction remains the biggest challenge, the extraction should be specifically designed and trained for different types and formats of the data source.

2.2. Ontology model

Different from the commonly used search engine, knowledge management is implemented generally for a specific task in a given domain. As the knowledge management application for the specific task only requires partial knowledge extracted from the raw data, the cognitive-based knowledge management would be too general to extract the required information. On the other hand, the linguistic or expert knowledge bases heavily rely on the rules, becoming difficult for extension [23]. Ontology is an abstract knowledge modeling, which treats the knowledge as concepts, associated attributes, and relations. Ontology can be divided into four categories: application ontologies; domain ontologies; generic ontologies and representation ontologies [4]. Application ontology captures the knowledge for a specific domain and can be applied in specific tasks through extension [24]. Since the application ontology is designed for a specific application, the relations among the concepts are built on specific tasks. The knowledge acquisition can be automatically or semi-automatically conducted using the reasoning mechanisms involved in the application ontologies. Domain ontologies focus on a specific domain for conceptualizations. In this situation, the primary task is to eradicate the misperception among the concepts [25]. Generic ontologies are considered as high-level ontologies. It covers the knowledge in multiple domains by defining the abstract concepts at the high-level, their associated attributes and relations. For example, a generic ontology is used to create models for policy-based regulations, which covers the domains of food and drug administration, financial regulation, contracts and individuals conducting business [26]. Representation ontology is oriented to the knowledge representation languages. It is not restricted to a particular domain, and is primarily used in the semantic web, which provides a platform for automatic data/information processing with extensible metadata [27]. In this study, the main task is to extract the knowledge embedded in the paper abstracts from four aspects (background, objective, solutions, and findings). An ontology is used to define the knowledge representation among these four aspects and the relations among the four knowledge elements are explicit. Therefore, representation ontology will be used in this study.

2.3. Natural language processing

Manually extracting the concepts of knowledge in the natural language is time-consuming and labor-intensive, and experts' domain knowledge is required to develop the industry knowledge bases. With the advancement of computer's capability in processing natural language, knowledge extraction becomes more efficient. If the computer system can read and understand the natural language, the knowledge concepts can be automatically extracted. Researchers have paid lots of attention to the field of NLP. As an interdisciplinary research area, NLP involves language and speech processing, human language technology, computational linguistics, speech recognition and synthesis, etc. [28]. Along with the rapid development of information and computation technology, NLP has played significant roles in various applications, including conversational agent or dialogue systems [29], machine translation [30], knowledge mining and reasoning [31], search engine [32], etc. Although great progress has been made in NLP, it remains a great challenge because of the inevitable ambiguity of the representation in natural languages, and the continuous evolvement of the vocabulary and the syntax [28].

The processing of natural language can be divided into three levels: vocabulary processing, syntax processing, and semantic and pragmatic processing. Vocabulary processing, including representation and computing, is the basic operation. Conventionally, Regular Expression [33], Finite-state Automaton [34], Transducer [35], N-gram model [36], Part-of-speech Tagging [37], Hidden Markov Model [38] and Maximum Entropy [39] are the most widely utilized models to handle the vocabulary related issues. For syntax processing, formal grammar and treebank are the most commonly used tools [40]. Later, the probability model and properties constraint-based formalism are proposed and utilized [41]. Meaning representation is the most challenge work in the semantic and pragmatic processing. The First-Order Logic, semantic network, conceptual dependency graph are used to present the semantic information [42]. The conventional methods in the NLP primarily rely on the probability model and the dataset scale, making the modeling generally be the case-specifically designed.

With the improving ability of machine learning, statistic learning was introduced to NLP. Through the co-occurrence matrix and singular value decomposition, the vocabularies are translated into dense vectors. After that, the operations among the semantic elements can be implemented through the operations of the trained vectors. Word2vec, developed by Google [43], is the earliest vector representation for the vocabularies. Through Continuous Bag-of-Words (CBOW) or continuous skip-gram [44], the words are vectorized based on the context. However, Word2vec only considers the neighbor context. Although it can be fast and directly established, its differential ability in the semantics is limited. GloVe is proposed by employing global vectors to represent the words [45]. Through a global co-occurrence matrix, the local and global information is considered in the vectors training. Though GloVe is more robust than Word2vec, it is hard to be adapted to different contexts. To tackle this problem, researchers proposed to employ multiple Long and Short-Term Memory (LSTM) staked neural model to learn the morphology features at different levels (ELMo) [46]. It synthetically considers the demand for dis-ambiguation, part-of-speech and syntax. ELMo performs much better than Word2vec and GloVe. In addition, it can be easily extended to a concrete application through pre-training with fine-tuning. However, the training of ELMo model is sequence procedure, because of the characteristics of LSTM. Thus, the previous contexts have a higher priority than the latter ones. Later, BERT model is proposed [47]. The transformer and attention mechanism are employed to build the bidirectional encoders. Then, the morphemes are learned from multiple attention headers. The BERT model has already been widely validated, and it can be directly used to extract morphemes features from the language texts through pre-

training.

2.4. Applications of NLP in construction management

NLP as one hot topic of Artificial Intelligent research, its application has increased in the field of construction management. In the construction industry, NLP technology is used to interpret the unstructured injury reports, and then build a structured accident database [48,49]. To facilitate the management of construction contract documents, NLP is used to process contract knowledge. A shallow parser-based semantic knowledge extracting system is proposed in Qady and Kandil's work [50]. By evaluating the performance to categorize and retrieve contract document, the NLP-based system can achieve almost 80% of the average kappa score attained by the evaluators and 90% of their F-measure score, proved that the NLP is efficient to process the contract documents. Focusing on the information extraction from construction regulatory documents, the NLP technology is used to build a semantic and rule-based model [51]. Two NLP techniques, Vector Space Model and semantic query expansion, are adopted in Zou et al. research to improve the efficiency and performance of risk case retrieval in construction project risk management [52]. Although NLP technology is used in various kind of situations in construction management, summarizing the advantage of NLP technology, the NLP technology performs well to cope with a large number of text documents which for human' manual operation is quite time and labor consuming.

3. Ontology-based literature knowledge graph and reasoning network framework

Generally, the literature includes four integral parts: citation information, abstract, main body, and references. Conventional journals, representing a large percentage of the academic database, are usually subscription-based, making only the citation information, abstract and references accessible to the public. Researchers have paid lots of attention to the knowledge graph based on the citation and reference data. The hot research topics, research trends, the connections and relationships among the researchers and their affiliations can be identified through citation and reference analysis. However, a large part of the valuable knowledge is neglected because the abstract and main part of the literature is not included in the analysis. Manual review of the literature one by one, labor-intensive and time consuming, is the primary way to summarize the knowledge in the literature. To tackle with the limitations, the proposed framework in this study considers the abstract data, in addition to the citation and reference data, for conducting automatic knowledge extraction and reasoning network building.

Different from the citation and reference data, the abstract essentially is plain text data, with no consistent and standard format. Therefore, a four-elements composed ontology model is designed to represent the knowledge embedded in the abstract. In the proposed framework, the abstract is considered to be composed of sentences. Based on the designed four-elements ontology model, each sentence can be assigned to one of the four elements in the ontology model. By identifying the background, objectives, solutions and findings, the content of abstract data can be expanded into four knowledge domains. For each knowledge domain, an independent network of knowledge topics can be established through correlation analysis. Those networks can be connected according to the causal relationship among these four elements. Four reasoning paths can be established to generate a knowledge reasoning network. The overall framework has three levels as shown in Fig. 1. First, at the ontology modeling level, the abstract data is expanded into four independent domains; second, at the knowledge graph level, the network of knowledge topics are generated for each knowledge domain; finally, at the reasoning network level, the knowledge reasoning network is built according to the causal relationship among the four elements of the ontology model. The detailed procedures are illustrated as follows.

3.1. Natural language processing-based ontology modeling

Ontology is an explicit specification of conceptualization and a formal way to define the semantics of knowledge and data. When mining knowledge in the literature, the abstract data, as the supplementary information of the citation data, gives a clear and concise statement for the research. Because different researchers have different writing styles, the abstract has various kind of discourses, which can be hardly represented using a uniform structure. With natural language processing technology, semantic information can be automatically extracted from the abstract. Then the ontology model can be established at the semantic level, and finally, the literature abstract can be represented using a consistent and clear structure, which is critical for knowledge mining.

3.1.1. Four-element ontology-based abstract knowledge representation for the abstract

The abstract of literature is a concise, complete summary of the research work, helping the reader quickly understand the purpose and contents of the work. The components in the abstract might vary for different disciplines. An abstract in social science, natural science, and engineering should include the research scope, purpose, methodology, and results of the work. While an abstract in humanities may contain the arguments, background, and conclusion. There exists no regulation requiring the abstract should be written in a pre-defined structure, and the journals also do not require that the abstract must have certain content. However, for a logical stringent and systematical document, the abstract must offer readers an overall view of the full content of the work to help the readers decide if the work is valuable. In addition, the abstract is used to index the full document in the online databases. Therefore, although the length of the abstract is limited, it is expected that a brief introduction of the research background, a summary of work conducted in the research, and the major findings or the outcomes of the research should be included in the abstract. Researchers also investigated the structures of abstracts, and found that the background, objective, method, results, conclusions can be clearly identified for most abstracts [53,54].

In this work, the authors assume that an abstract is composed of four components: background, objective, solution, and finding. The background provides a brief introduction to the motivation and point of departure. The objective describes what is expected to achieve by the study. It can be a survey or a review for a specific research topic, a significant scientific or engineering problem, or a demonstration for research theories or principles. The solution part presents the methods, models, or technologies employed in the research to achieve the research objectives. The findings give a summary of the results. A four-elements composed ontology model is designed to represent the contents of the abstract for literature.

$$\text{Abstract} = \{\text{Background}, \text{Objectives}, \text{Solutions}, \text{Findings}\}$$

The abstract can be divided into sentences using periods. Each sentence can be tagged with a specific type among the four elements above, according to its semantic information and contents. The details will be introduced in Section 3.1.2. The matching between the sentence and the four elements can be summarized in a table shown in Fig. 2. In the table, 1 means the sentence is tagged as the element, while 0 means the sentence is not tagged as the element.

3.1.2. Natural Language Processing (NLP)-based automatic ontology elements identification

As a piece of text data, an abstract is composed of ordered sentences, which are composed of a set of ordered vocabulary words. Since manually assigning each sentence to its corresponding ontology element is quite time-consuming and labor-intensive, automatic

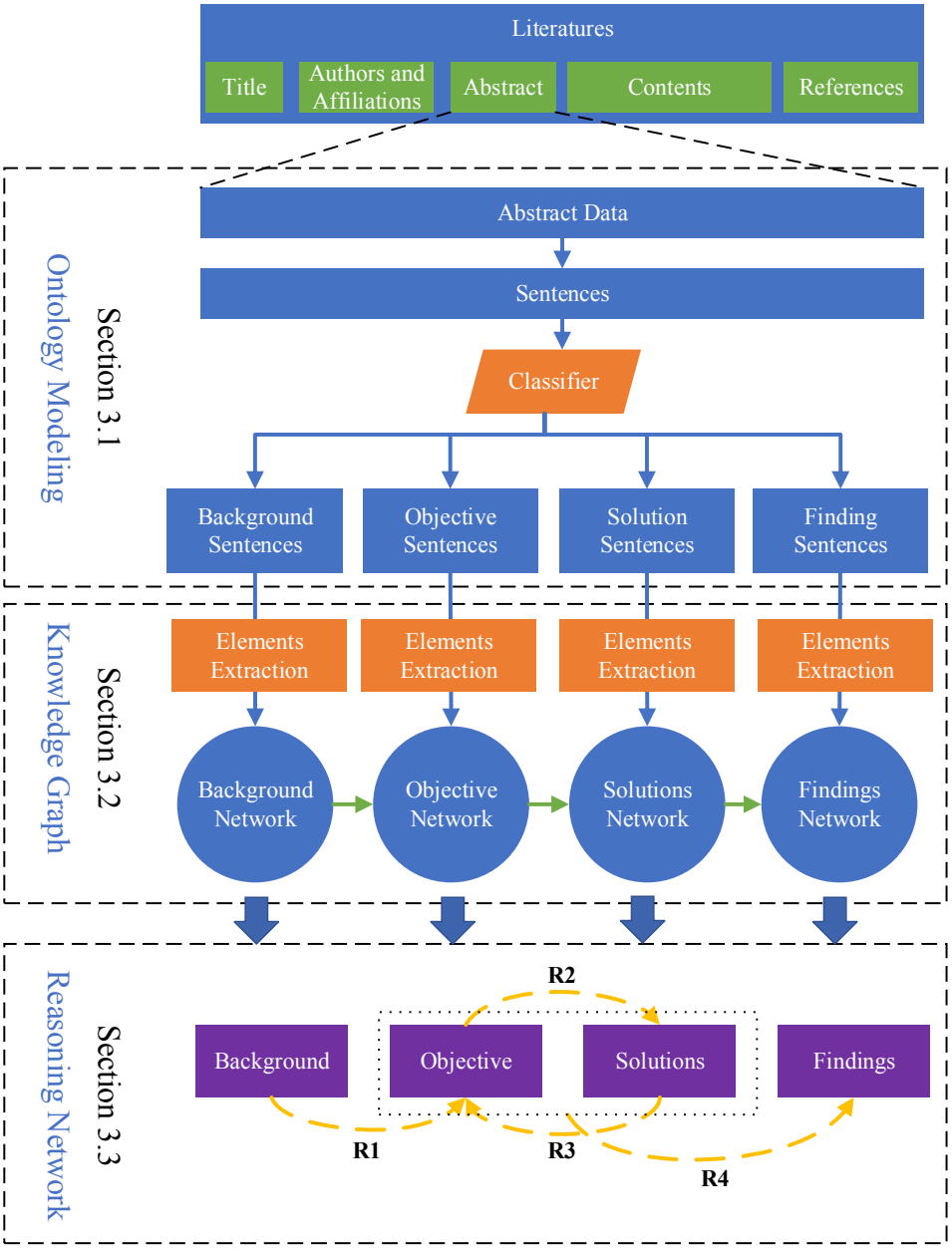

Fig. 1. The three-level framework of the ontology-based literatures' knowledge reasoning network modeling.

Abstract	Background	Objective	Solution	Finding
Sentence 1	0/1	0/1	0/1	0/1
Sentence 2	0/1	0/1	0/1	0/1
…	…	…	…	…
Sentence n	0/1	0/1	0/1	0/1

Fig. 2. Ontology model for literature abstract knowledge representation.

identification ontology element is required to process a large number of literature. In the proposed framework, text classification is used at the sentence level, to achieve the automatic ontology elements identification for sentences.

The text classification relies on the comparison of the text semantics. Because of the ambiguity of the words, it was a great challenge to establish a robust model for semantic understanding, which can compare the semantics in the given context. Along with the development of NLP technology, researchers proposed Word2Vec, ELMo and BERT (Bidirectional Encoder Representations from Transformers) models to describe the words' semantic meanings quantitatively. With those models, the similarity and difference among the words in the given documents can be calculated. The Word2Vec can be regarded as a linear system. Through the pre-training based on a large annotated corpus, the words are projected into a fixed-length vector. Although the computational efficiency is high, the model fails to learn the context of the words. Therefore, a special word vector needs to be re-trained for each situation. By employing stacked bidirectional LSTM, the ELMo can

embed the context relations into vectors. BERT model is the latest text semantic understanding model developed by Google. Based on the newly proposed attention and transformer mechanism, BERT model not only considers the relationship features at the char level, word level, and sentence level, but also consider the relations between the sentences. By pre-training, the BERT can achieve good robustness. Then the BERT model can be transformed into different tasks with the pre-trained model by fine-tuning. Thus, the BERT model is used in the framework to conduct quantification for the semantics.

The BERT model first translates the input sentence into embedding vectors. A pre-established vocabulary set containing 30,522 tokens is used to replace each item (including punctuation marks, delimiters, etc.) of the sentences by a token embedding vector. The maximum length of a sentence is 512, and a positional embedding vector is added to each token to denote its position in the sentence. A segment embedding is used to denotes the tokens belongings to different segmentations, since BERT support two neighbor sentences pairs. By directly adding these three embedding vectors, the input sentences are translated into input features for the BERT model. An encoder is then used to encode the positional relations, referential relations, and semantic relations of the tokens through the attention training. There are 12 stacked attention layers to learn the relations among the tokens. All these learned relations are stored into the embedding vectors as the output of the BERT core module. Through the BERT model, one text sentence is translated into a vector with a size of 768. Then all the semantic manipulation, including similarity or difference calculation, is conducted in the vector domain.

In this work, the output vector of the BERT core model is utilized to conduct classification. A 768×4 fully connected neural network is designed to model classify operation. The output vector of the BERT core model is the input of the classification part. Then through classification, each sentence in one given abstract record is assigned to its corresponding ontology element type. The structure of the BERT-based ontology element identification model is illustrated in Fig. 3.

The Logits-based Binary Cross Entropy is used as the loss function to train the classification part then to conduct ontology element identification. After applying the ontology element identification to the abstract, the matching table between the sentence and the four elements (as shown in Fig. 2) can be obtained.

3.2. Multi-spaces associated knowledge graph

With the Four-Element Ontology model, the abstract is decomposed into four sub-domains of knowledge. The background domain primary contains an introduction and description statement. The objective domain presents the core research topic, challenges or research gaps. The solutions domain contains all kinds of procedures, models used or frameworks developed, including technologies, systems, and tools. The finding domain presents the performance descriptions or outcomes. Corresponding to those four domains, four independent association networks are constructed, with the extracted knowledge elements as the vertexes, and the connections are quantified through the correlations.

According to the linguistics, the noun phrases are the main body of one sentence, while the other parts offer modified, complementary, and emphasized description. Thus, the noun phrases are extracted as the knowledge elements, corresponding to the vertexes (V) in the graph. First, according to the dependency parsing, a syntax tree is established corresponding to each sentence in the abstract. Second, for each word in the sentence, the part of speech (POS) for each word is detected. Then the nouns and numerals are chosen as the probe points. All the sub-syntax trees rooted from the selected nouns and numerals are trimmed out, and only the objective and subjectively related sub-trees are retained. Third, for each obtained sub-tree, the contained texts are segmented into pure noun phrases, with splitting at the positions of the verbs and prepositions. Last, drop the repeated noun phrases for one syntax tree. In this way, the language texts are translated into

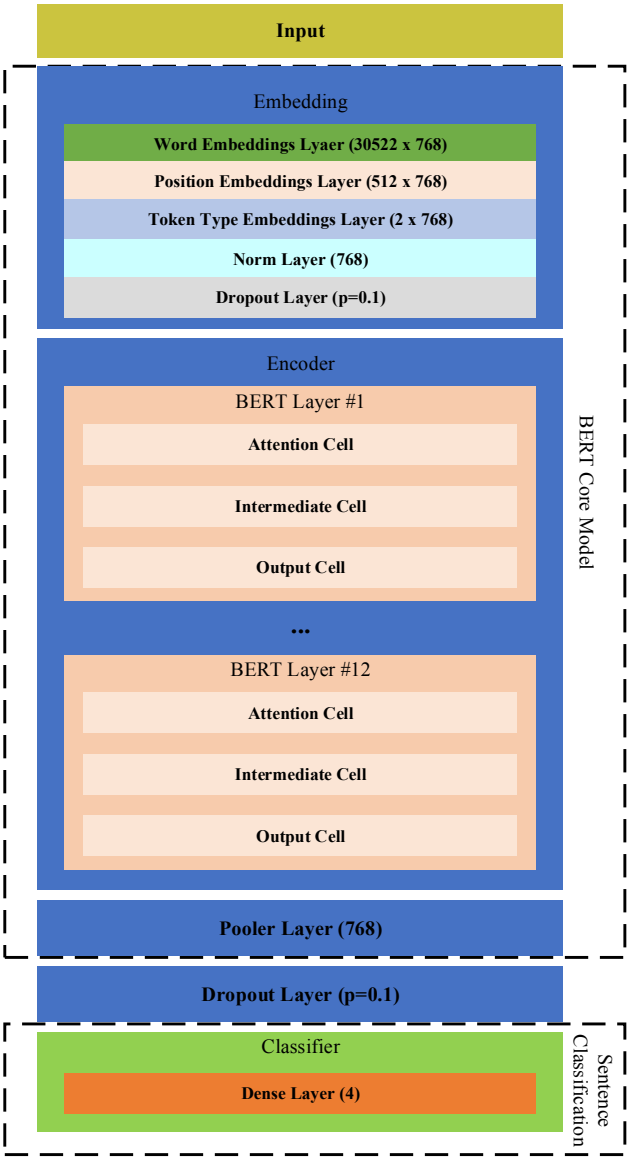

Fig. 3. BERT-based sentence related ontology elements classifier.

individual knowledge elements for each knowledge space. The pseudocode is illustrated in Algorithm 1.

Algorithm 1.

```
Input: Sentence_List = [Background_sentences, Objective_sentences, Solution_sentences, Finding_sentences]
Foreach Sentence in Sentence_List:
  Parse_Tree = Dependency_Parse(Sentence)
  Foreach (Token.POS in ['NOUN', 'NUM']) in Parse_Tree:
    Probe_Points.append(Token)
  Foreach Root_Token in Probe_Points:
    Sub_Tree = Get_Sub_Tree(Root_Token, Parse_Tree)
    Foreach Span in Sub_Tree:
      If Span.DEP in ['OBJ', 'SUBJ']:
        Parse_List.append(Span.text)
Foreach Parse in Parse_List:
  If Parse.is_stop_word():
    Parse_List.drop(Parse)
Return Parse_List
```

Through the knowledge elements extraction, the text is segmented into phrases, and the modifiers are removed. On the other hand,

through the segmentation process, huge number of short nouns are recognized as knowledge elements. Without the context, these nouns contain very limited information. To eliminate the information noise caused by these meaningless nouns, the character-based Shannon Entropy is used to quantify the information in the extracted knowledge elements. The formula of the character-based Shannon Entropy calculation is presented in formula (1).

$$H(X) = -\sum_x P(x)(\log_2[P(x)]) \tag{1}$$

where $P(x)$ is the probability of the character (x)'s belong to a given knowledge element. It shows that the longer knowledge elements contain more information. Thus, the super short knowledge elements can be filtered.

Individual knowledge element gives limited information. Through connecting the knowledge elements, the main concepts can be represented. Link mining refers to data mining techniques that are used to evaluate and investigate the links when building predictive or descriptive models, especially in social network analysis, hypertext mining, and web analysis. PageRank and HITS are the most common link-based object ranking algorithms [55,56], which is quite suitable for web pages. In social network analysis, the relation is quantified primary from two aspects: subordinate relationships; frequency of co-occurrence [57,58].

In this work, a connection factor is designed to quantify the link among the extracted knowledge elements. It involves three aspects of connections: (1) the relative distances (D); (2) semantic similarities between the two knowledge elements (S); and (3) the information content of the target (E). The connection factor (C) is defined as formulas (2)–(5).

$$C(V_i \rightarrow V_j) = 1/D(V_i, V_j) + S(V_i, V_j) + E(V_j) \tag{2}$$

$$D(V_i, V_j) = \begin{cases} 1, & if \ \ V_i \ \ and \ \ V_j \ \ not \ \ in \ \ the \ \ same \ \ paper \\ Distance(V_i, V_j)/Length(text) \end{cases} \tag{3}$$

$$S(V_i, V_j) = Cosine\_similarity(\vec{V_i}, \vec{V_j}) \tag{4}$$

$$E(V_j) = H(V_j) \tag{5}$$

where $D(V_i, V_j)$ is the relative distance of knowledge elements $V_i$and V_j. If V_i *and* $V_j$are not in the same paper, the distance is 1; otherwise, the distance is the ratio of the absolute distance between V_i *and* V_j to the length of the text. It is negatively correlated to the connection factor (C), because the two knowledge elements are stronger associated if they are located closer to each other. $D(V_i, V_j)$ has a value from 0 to 1. $S(V_i, V_j)$ is the semantic similarity coefficient of $V_i$and V_j. The knowledge element texts are encoded into word embedding vectors first, and the cosine similarity is used to quantify the semantic similarities. $S(V_i, V_j)$ has a value from -1 to 1. $S(V_i, V_j) = 1$ indicates that V_i *and* V_j have the same semantic meaning. Thus, the semantic similarity is positively correlated to the connection factors. For individual knowledge elements without additional descriptions, the more similar semantics, the closer of the corresponding vertexes located in the knowledge graph, which also means higher connection values. $E(V_j)$ is the Shannon Entropy value of the target knowledge element V_j. For a given knowledge graph, a random walking starting from a given element towards the target element, the more information contained in the target elements, the more reliable the target is. Thus, the $E(V_j)$ is positive correlated with the connection factors. $E(V_j)$ has a non-negative value. Through quantifying the connections, all the knowledge elements are connected, and a knowledge graph is established.

By network construction, four knowledge graphs are created. The background map ($\boldsymbol{MAP_b}$) illustrates the background information of existing studies; the objective map ($\boldsymbol{MAP_o}$) presents the research topics and challenges; the solutions map ($\boldsymbol{MAP_s}$) collects the proposed methods, systems and tools; and finding map ($\boldsymbol{MAP_f}$) presents the outcomes of the current studies. The knowledge is organized into four spaces, and the reasoning would be conducted following the paths crossing these four domains.

3.3. Ontology-based knowledge reasoning network

According to the proposed ontology model, there is one main reasoning path among the four knowledge graphs. Starting from the background map, with a given topic, the past and current situations can be reviewed through $\boldsymbol{MAP_b}$. Then based on the $\boldsymbol{MAP_o}$, the research challenges and objectives can be identified, and the corresponding solutions can be retrieved from $\boldsymbol{MAP_s}$. Finally, the findings of the implemented solutions to address the identified problems can be directly retrieved from $\boldsymbol{MAP_f}$. The main reasoning path can be described as formula (6).

$$\boldsymbol{MAP_b} \rightarrow \boldsymbol{MAP_o} \rightarrow \boldsymbol{MAP_s} \rightarrow \boldsymbol{MAP_f} \tag{6}$$

Based on the main path, four reasoning rules are designed. Through summarizing the current situations, what are the research challenges (*R1*)? To achieve a given objective, what would be an applicable solution (*R2*)? By giving the solution, what type of issues and objectives can be handled (*R3*)? For a given solution with one research objective, what are the findings (*R4*)? These four-reasoning paths can be presented in formulas (7)–(10).

$$R1: \ \boldsymbol{MAP_b} \rightarrow \boldsymbol{MAP_o} \tag{7}$$

$$R2: \ \boldsymbol{MAP_o} \rightarrow \boldsymbol{MAP_s} \tag{8}$$

$$R3: \boldsymbol{MAP_s} \rightarrow \boldsymbol{MAP_o} \tag{9}$$

$$R4: (\boldsymbol{MAP_o}, \boldsymbol{MAP_s}) \rightarrow \boldsymbol{MAP_f} \tag{10}$$

4. A case study: construction management knowledge model

As a demonstration and validation of the proposed framework, a case study is conducted to build a literature knowledge reasoning network with the topic of construction management. Construction management is considered as a complex interdisciplinary filed involving engineering, scientific, and sociological knowledge. Therefore, efficient knowledge management for construction management research is quite important. The literature data comes from two major online academic databases: Web of Science and Scopus. The keywords "construction management" is used for literature searching. For Web of Science database, only the core collections are used, including SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI. There is no limitation on the time range. 2539 records from Web of Science and 5887 records from Scopus are obtained (the retrieval was conducted in Nov. 2018). Only 3823 English records contain the abstract. Thus, the case study is conducted on these 3823 data records.

4.1. Dataset summary

The dataset comes from 1050 publication sources, covers the subject areas of engineering, science, and education. The document type includes journal articles, conference proceedings, review papers, book chapters, and editorial materials. The publication period spans from 1991 to 2019. The distribution of the literature type is summarized in Table 1. As shown in Table 1, research journals and conference papers

Table 1
Literature type count.

Type	Count	Proportion
Article	2743	71.6%
Proceeding	1036	27.1%
Review	34	0.9%
Others	15	0.4%

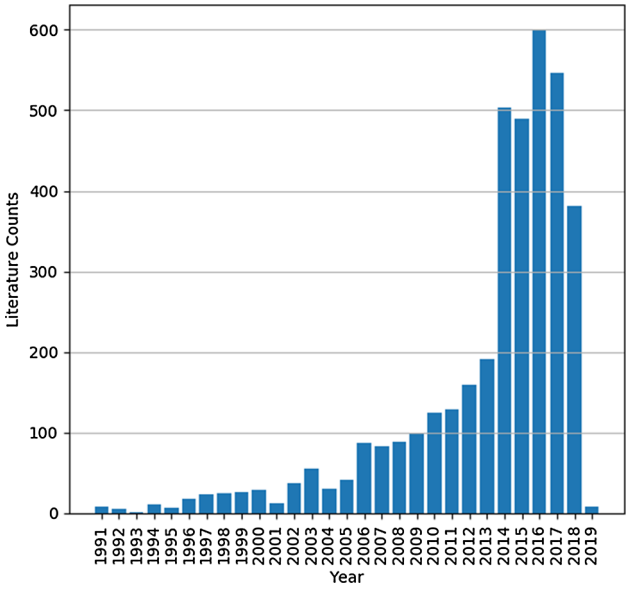

Fig. 4. Involved literature counts for years.

Table 2
Top 10 involved academic journals.

Journal Name	Count	Proportion
Journal of Construction Engineering and Management	339	8.9%
Automation in Construction	140	3.7%
Journal of Management in Engineering	131	3.4%
Engineering, Construction and Architectural Management	87	2.3%
Journal of Computing in Civil Engineering	84	2.2%
Journal of Professional Issues in Engineering Education and Practice	64	1.7%
Journal of Civil Engineering and Management	62	1.6%
Construction Management and Economics	57	1.5%
Canadian Journal of Civil Engineering	36	0.9%
International Journal of Construction Education and Research	29	0.8%

are the main types of literature data. Fig. 4 illustrates the temporal distribution of the literature. The top 10 source journals are summarized in Table 2. Though the stricter search keyword limits the number of literature extracted from the journals, the wide acknowledged academic publications in the field of "construction management" are covered. Summarizing the dataset, it contains multiple topics and offers substantial topics to construct a literature knowledge reasoning network for construction management.

4.2. Validation of automatic ontology elements identification

702 literature are randomly selected to validate the NLP-based automatic ontology elements identification. Among those 702 literature, 70% is employed as training data, and 30% is used as test data. 7006 sentences are extracted from the abstracts of these 702 papers. Each sentence is manually annotated according to its belongings to the ontology elements, and 1327 sentences are labeled as "Meaningless", which generally only offer copyright information (e.g., "© 2013 American Society of Civil Engineers") from the publications. 258 sentences have multiple labels: for example, some authors use complex sentence structures, in which one sentence may contain the research background, objective, and even mention the solutions. After annotation, among those 7006 sentences, the proportion of the background, objective, solution, and finding elements are 23%, 14%, 26% and 22%, respectively. It also should be noted that there are 23% sentences labeled as meaningless. It is because that the sentences segmentation is based on pre-defined rules which defines all the non-ASCII are regarded as the split mark, the copyright mark caused additional segmentation. To enhance the robustness of the model, the meaningless sentences are also included in the training and test. On the other hand, since only 3% of the sentences have multi-labels, all the sentences are regarded as a single-label classification for the ontology-elements identification in this work. The manually tagged labels are used as the ground truth for performance evaluation.

The uncased BERT-Base model with 12-layer, 768-hidden, and 12-head is used in this case study. Through fine-tuning, a 768-hidden unit composed full connected neural network is used to train the classification based on the features extracted by the BERT-Base model. The training performance is summarized in Table 3. The training precision of the identification is 0.90 for background, 0.76 for the objective, 0.93 for solutions, and 0.91 for findings. The recall is 0.94, 0.80, 0.84, and 0.95 respectively. For the testing performance, the precision is 0.75 for backgrounds, 0.56 for objectives, 0.75 for solutions, and 0.75 for findings. The recall is 0.76, 0.64, 0.71 and 0.73 respectively. The testing performance of the automatically ontology elements identification is summarized in Table 4.

Through the training and testing validation of the automatic ontology elements identification model, it shows that there exist apparent patterns in the abstract structure, although there is no strict requirement for the writing of the abstract. The background, solution, finding and meaningless situations can be accurately detected (precision $>$ 0.9) in the training process, and the precisions are higher than 0.7 in the testing. These high precision rates indicate that the researchers follow a general structure to cover the background, objectives, solutions, and findings in the abstract. Therefore, the trained model can be used to extract the knowledge from the literature abstracts and classify them into four aspects.

4.3. Knowledge graph construction

Through the BERT-based ontology elements identification model, 5938 sentences are identified. According to the identified ontology element labels, four individual collections are generated corresponding to the four ontology elements: background, objective, solution, and finding. There are 1582 sentences identified to be background, and 947, 1859, 1550 to be objectives, solutions, and findings, respectively. Using the algorithm described in Algorithm 1, the knowledge elements are extracted from the corresponding sub-collections. The scale of identified sentences and extracted knowledge elements are summarized in

Table 3
Training performance of automatic ontology elements identification.

		Prediction				
		Meaningless	Background	Objective	Solution	Finding
Ground Truth	Meaningless	919	3	3	2	1
	Background	3	1036	43	14	11
	Objective	1	75	530	42	14
	Solution	1	25	106	1094	75
	Finding	1	11	16	24	1033

Table 4
Testing performance of automatic ontology elements identification.

		Prediction				
		Meaningless	Background	Objective	Solution	Finding
Ground Truth	Meaningless	385	5	1	3	5
	Background	4	359	54	28	29
	Objective	1	49	182	39	14
	Solution	3	33	61	397	64
	Finding	1	34	27	64	339

Table 5
Scale summary of the established knowledge graph.

Sub-space Name	Count of Identified Sentences	Count of Extracted Knowledge Elements
Background	1582	17,930
Objective	947	13,137
Solution	1859	19,106
Finding	1550	15,550

Table 5.

Fig. 5 shows the distribution of the entropy values of the knowledge elements corresponding to the four sub-spaces. It shows that the contained information content for each sub-space is generally subject to normal distribution.

To simplify the illustration of the constructed knowledge graph, only ten randomly selected papers are selected to conduct the visualization. Table 6 shows the titles and source journals of those ten papers.

Figs. 6–9 present the constructed knowledge graph corresponding to the knowledge space of the background, objective, solutions, and findings. Through investigating the connections of the extracted knowledge elements in these four knowledge graphs, the most connected elements can be identified. For those four knowledge graphs, the most connected items are summarized in Table 7.

Summarizing the background knowledge graph for the selected ten demo literature, primarily the project management, safety, and green building are involved. For project management, the extracted knowledge elements include physical infrastructures, coordination process, regulatory and contractual compliance checking, etc. For safety, automated worker action recognition, productivity, safety, health issues, and construction accidents, etc. are concerned. The greenhouse gases (GHG), materials, construction, maintenance, rehabilitation, highway infrastructure, construction rehabilitation operations, etc. are considered with the background of green building.

For the objective knowledge space, corresponding to the description in the background, the construction projects delay, feasibility, net

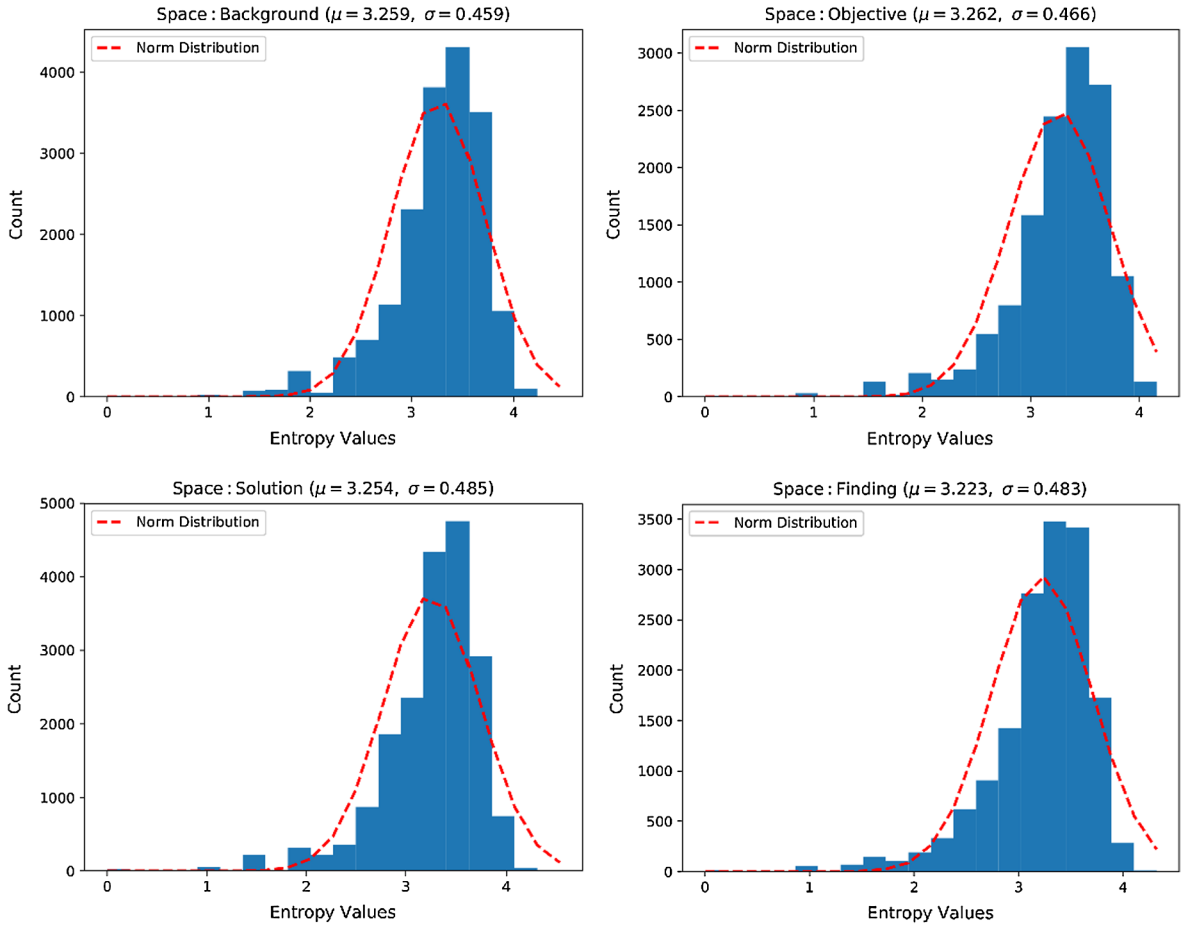

Fig. 5. The entropy values distribution for the four sub-spaces.

Table 6
Randomly selected 10 papers to illustrate the knowledge reasoning graph.

Authors	Title of the Papers	Source of Journal/Conference
Mbala M, Aigbavboa C, Aliu J.	Causes of delay in various construction projects: A literature review	International Conference on Applied Human Factors and Ergonomics
Ryu J H, Seo J O, Jebelli H, et al.	Automated Action Recognition Using an Accelerometer-Embedded Wristband-Type Activity Tracker	Journal of Construction Engineering and Management
Shafahi A, Haghani A.	Project selection and scheduling for phase-able projects with interdependencies among phases	Automation in Construction
Park J, Cho Y K, Ahn C, et al.	A Wireless Tracking System Integrated with BIM for Indoor Construction Applications	Proceedings of the Construction Research Congress
Alaloul W S, Liew M S, Zawawi N A.	Coordination process in construction projects management	Engineering Challenges for Sustainable Future
Salama D M, El-Gohary N M.	Semantic Text Classification for Supporting Automated Compliance Checking in Construction	Journal of Computing in Civil Engineering
Kartelj A, Šurlan N, Cekić Z.	Case-based reasoning and electromagnetism-like method in construction management	Kybernetes
Li H, Lu M, Hsu S C, et al.	Proactive behavior-based safety management for construction safety	Safety science
Hollar D A, Rasdorf W, Liu M, et al.	Preliminary Engineering Cost Estimation Model for Bridge Projects	Journal of construction engineering and management
Cass D, Mukherjee A.	Calculation of Greenhouse Gas Emissions for Highway Construction Operations by Using a Hybrid Life-Cycle Assessment Approach: Case Study for Pavement Operations	Journal of Construction Engineering and Management

present value (NPV), future investments, etc. are the primary research topic for the project management. For the construction safety background, the motion capture and behavior detections, and the application of sensors are most considered objectives. Correspondingly, the life-cycle emissions, long - term environmental impacts, state agencies are the main topics for the green building research.

In the solution knowledge graph, the mixed integer programming (MIP) model; binary classification; multilabel classification; semantic model; normative reasoning; favorability measures; sensitivity analysis; statistical models; complex tracking scenarios; the k-nearest neighbor, multilayer perceptron, decision tree, and multiclass support vector machine are the primary solutions, systems or algorithms for the identified objectives.

The finding knowledge graph presents that through literature review, poor site management; shortage of skilled labor; unrealistic project scheduling; labor absenteeism; design changes/rework due to the construction errors and accidents due to poor site safety are the major causes of delay in the timely delivery of construction projects. In addition, it contains the information that multiclass support vector machine with a 4-s window size showed the best accuracy (88.1%) to classify four different subtasks of masonry work. Correspondingly, the Bluetooth low energy (BLE) sensors and motion sensors with BIM (Building Information Modeling) are found efficient for indoor construction applications. Although only the keywords are offered, the original sentences containing the keywords can be retrieved to get more information if needed.

4.4. Reasoning network construction

The established knowledge graph offers a container for knowledge reasoning. The connections among the knowledge elements are the indicators while conducting deductions. As mentioned in Section 3.3, following the main knowledge reason path defined according to the characteristics of the ontology elements, there can be four knowledge

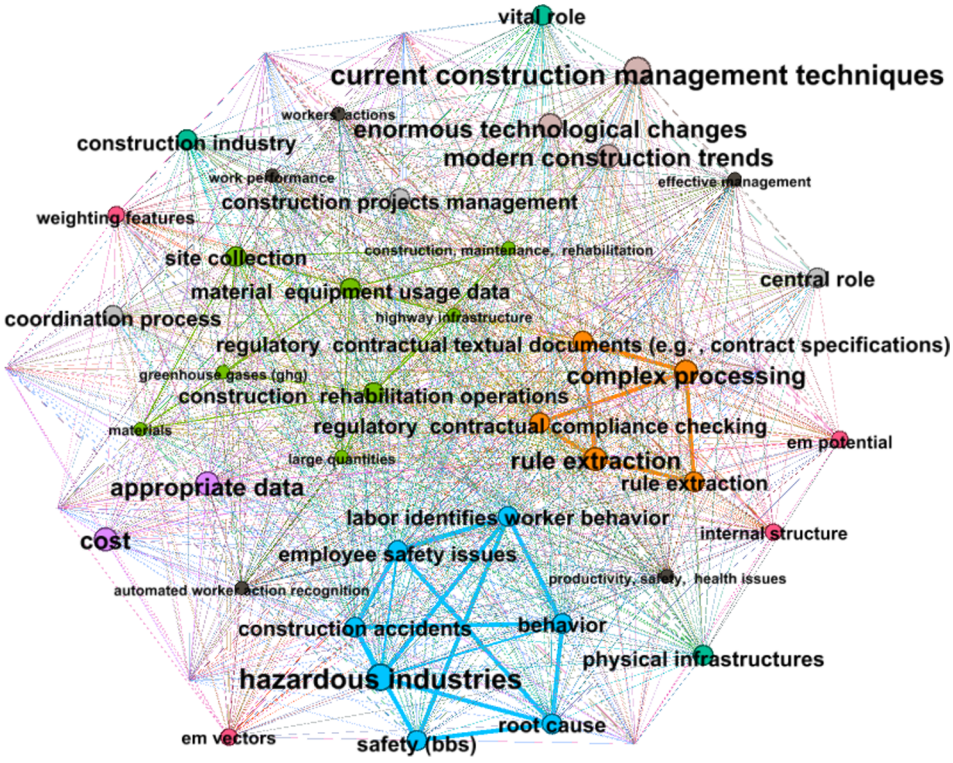

Fig. 6. Knowledge graph of the background space.

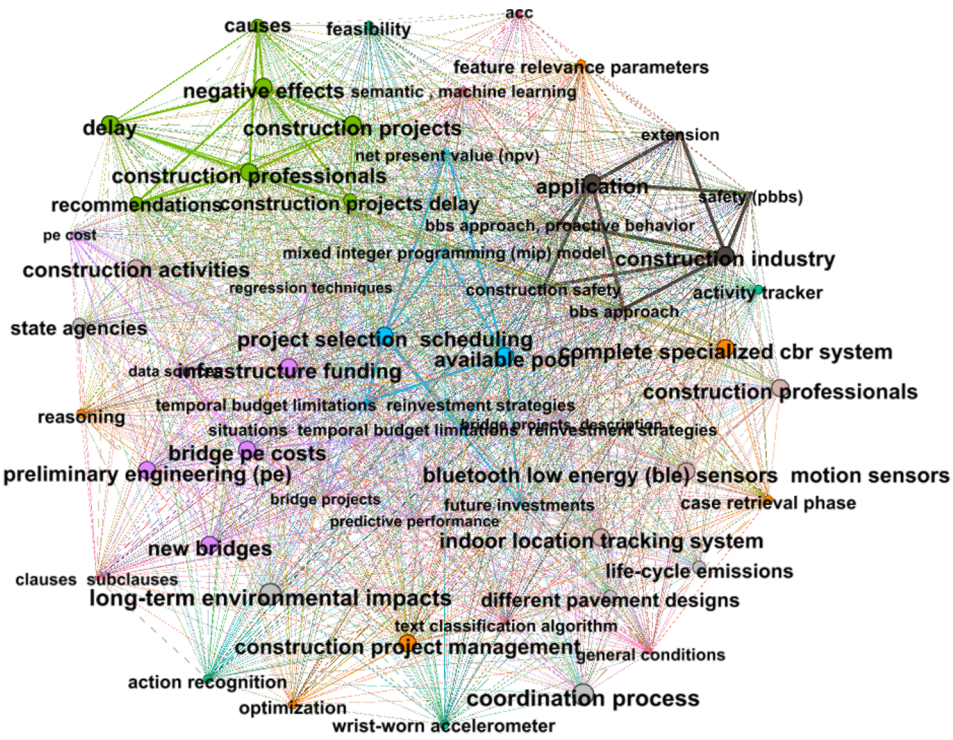

Fig. 7. Knowledge graph of the objective space.

reasoning routes.

(1) *R*1: $\boldsymbol{MAP_b \rightarrow MAP_o}$

In the background knowledge graph ($\boldsymbol{MAP_b}$), given one knowledge element (e.g., "modern construction trends" (BE1 in Table 8)), there is one induction: {BE1} → {OE1, OE2, OE3}, because those four elements (BE1, OE1, OE2, and OE3) exist in the same abstract record. This induction indicates that with the modern construction trends, researchers aim to improve the construction activities and apply the indoor location tracking system. The construction professionals are one key elements in the research of modern construction trends.

Starting from the above three elements in the objective space, the top three objective elements associated with those elements can be deduced based on the connection strengths:

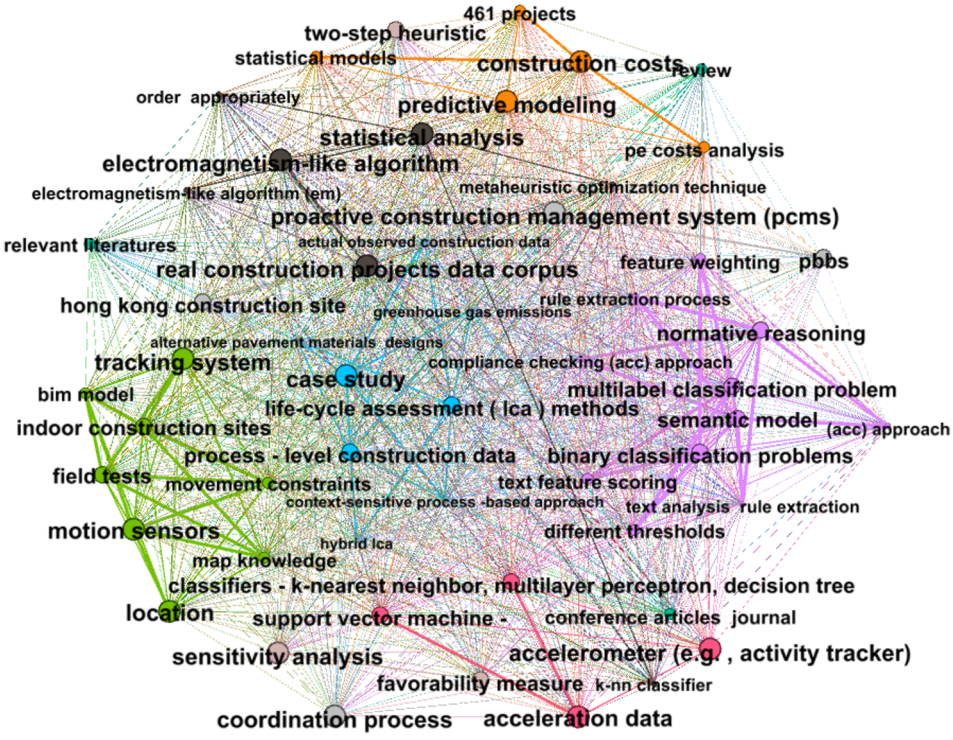

Fig. 8. Knowledge graph of the solution space.

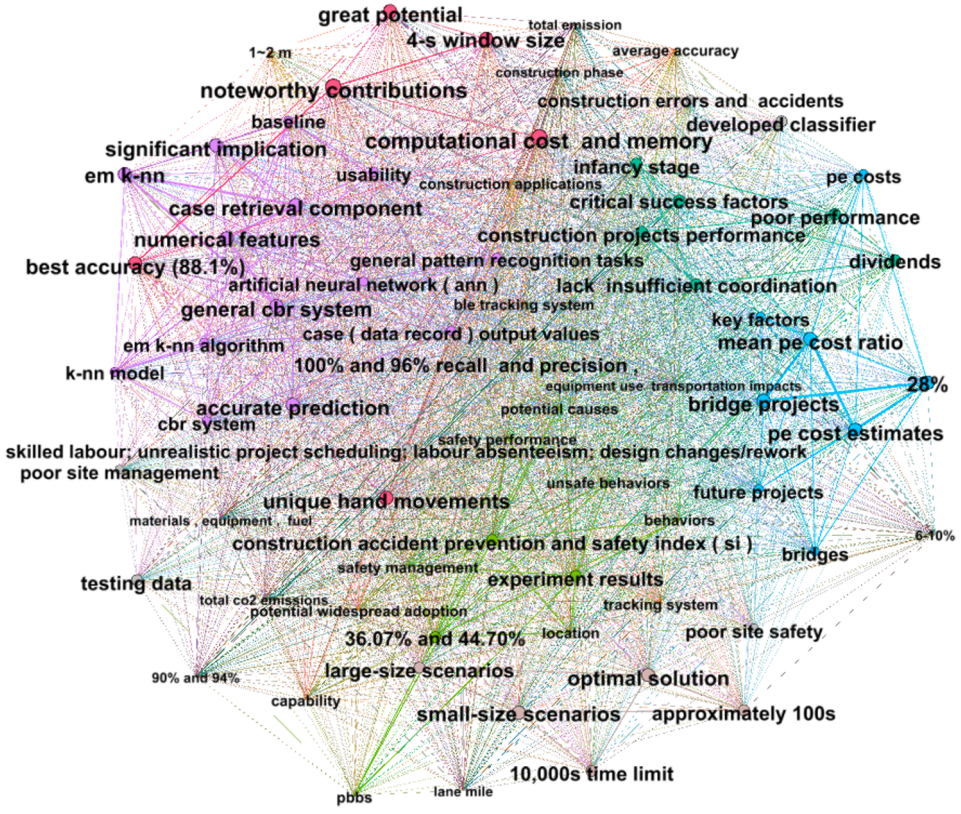

Fig. 9. Knowledge graph of the finding space.

Table 7
Top ten most connected knowledge elements in the established knowledge graphs.

Space	Identified Knowledge Elements
Background	*worker behavior; construction accidents; employee safety issues; hazardous industries; contractual textual documents; regulatory and contractual compliance checking; safety (bbs); rule extraction; current construction management techniques; material and equipment usage*
Objective	*negative effects; temporal budget limitations; reinvestment strategies; construction projects delay; project selection and scheduling; Net Present Value (NPV); infrastructure funding; bridge PE costs; predictive performance; construction activities;*
Solution	*mixed integer programming (MIP) model; binary classification; multilabel classification; semantic model; normative reasoning; favorability measures; sensitivity analysis; statistical models; complex tracking scenarios; classifiers – the k-nearest neighbor, multilayer perceptron, decision tree, and multiclass support vector machine*
Finding	*computational cost and memory; mean pe cost ratio; automatic construction action recognition; single wrist-worn sensor; construction sites; multiclass support vector machine; best accuracy (88.1%); timely delivery; 28%; numerical features*

Table 8
Extracted elements from the objective space with given knowledge in the background space.

Elements in Background Space		Elements in Objective Space	
ID	Element	ID	Element
BE1	Modern construction trends	OE1	Construction professionals
		OE2	Construction activities
		OE3	Indoor location tracking system
		OE4	Bluetooth low energy (BLE) sensors and motion sensors
		OE5	Construction projects delay
		OE6	Project selection and scheduling
		OE7	Temporal budget limitations and reinvestment strategies

{OE1} {OE4, OE5, OE6}

{OE2} {OE4, OE5, OE7}

{OE3} {OE4, OE6, OE7}

These connections can be elaborated as follows. “Bluetooth low energy sensor and motion sensors” (OE4), “construction projects delay” (OE5) and “project selection and scheduling” (OE6) are research objectives closely related to “construction professionals” (OE1). “Bluetooth low energy (BLE) sensors and motion sensors” (OE4), “construction projects delay” (OE5), “temporal budget limitations and reinvestment strategies” (OE7) are research objectives closely related to “construction activities”(OE2). “Bluetooth low energy (BLE) sensors and motion sensors” (OE4), “project selection and scheduling” (OE6), “temporal budget limitations and reinvestment strategies” (OE7) are research objectives closely related to “construction activities” (OE2).

With the connections shown above and the induction, a new induction can be developed as follows:

{BE1} → {OE4, OE5, OE6, OE7}

This indicates that with the modern construction trends, the researcher can also put their efforts on studies aiming to apply BLE sensors and motion sensors [59], reduce the construction delay [60], optimize the project selection and scheduling [61], and solve the temporal budget limitation through reinvestment strategies [62].

(2) *R2*: $\boldsymbol{MAP_o \rightarrow MAP_s}$

Table 9
Extracted elements from the solution space with given knowledge in the objective space.

Elements in Objective Space		Elements in Solution Space	
ID	Element	ID	Element
OE8	Action recognition	SE1	Accelerometer (e.g., activity tracker)
		SE2	Acceleration data
		SE3	Classifiers k-nearest neighbor, multilayer perceptron, decision tree
		SE4	Support vector machine
		SE5	Proactive construction management system (PCMS)
		SE6	Binary classification problems
		SE7	Process-level construction data
		SE8	Life-cycle assessment (LCA) methods
		SE9	Multi-label classification problem
		SE10	Favorability measure

Correspondingly, according to R2 rule, when given "action recognition " (OE8 in Table 9) in $\boldsymbol{MAP_o}$, the elements "accelerometer (e.g. , activity tracker)" (SE1), "acceleration data" (SE2), "classifiers k-nearest neighbor , multilayer perceptron , decision tree" (SE3) and "support vector machine" (SE4) can be directly extracted from the solution space ($\boldsymbol{MAP_s}$), since the elements (OE8, SE1, SE2, SE3 SE4) belonging to one abstract record, thus {OE8} → {SE1, SE2, SE3}. Then through connections retrieving in the solution space with starting from these directly identified elements, four additional connections are obtained. The involved knowledge elements are presented in the Table 9.

The corresponding extended connections are as follows:

{SE1} {*SE*5, *SE*6, *SE*8}

{SE2} {*SE*5, *SE*6, *SE*9}

{SE3} {*SE*5, *SE*6, *SE*7}

{SE4} {*SE*7, *SE*9, *SE*10}

It shows that the "accelerometer" (SE1) is used in the action recognition, the classification algorithms including the "k-nearest neighbor, multilayer perceptron, decision tree" (SE3), and "support vector machine" (SE4) are the main algorithms to do the recognition. In addition, according to the extended connection, "proactive construction management system" (SE5), "binary classification problems" (SE6) and "life-cycle assessment (LCA) methods" (SE8) are solutions closely related to "accelerometer" (SE1). "proactive construction management system" (SE5), "binary classification problems" (SE6) and "Multi-label classification problem" (SE9) are closely related to "acceleration data" (SE2). "proactive construction management system" (SE5), "binary classification problems" (SE6) and "process-level construction data" (SE7) are solutions related to "classifiers k-nearest neighbor, multilayer perceptron, decision tree" (SE3). "process-level construction data" (SE7), "Multi-label classification problem" (SE9) and "favorability measure" (SE10) are solutions closely related to "support vector machine" (SE4). Thus, a new induction can be established.

{OE8} → {*SE*5, *SE*6, *SE*7, *SE*8, *SE*9, *SE*10}

According to the raw abstract record, the "favorability measure" (SE10) and "process-level construction data" (SE7) actually are utilized to cope with the project selection and greenhouse gas emission issues, respectively. Whether these two solutions can be used in the action recognition needs to be verified further. On the other hand, it also shows that the "binary classification" (SE6) or "Multi-label classification" (SE9) can be the solutions for the action recognition, which has already been investigated by some of the studies [63–65]. Therefore, through *R2* rule, in addition to the solutions presented in one abstract record, multiple solutions can be deduced for one given research objective.

Table 10
Extracted elements from the objective space with given knowledge in the solution space.

Elements in Objective Space		Elements in Solution Space	
ID	Element	ID	Element
OE9	Case retrieval phase	SE11	k-nn classifier
OE10	Reasoning		
OE11	Optimization		
OE12	Feature relevance parameters		
OE13	Construction project management		
OE14	Complete specialized CBR system		
OE15	Construction projects delay		
OE16	Indoor location tracking system		
OE17	Project selection scheduling		
OE18	Text classification		

(3) *R*3: $\boldsymbol{MAP_s} \rightarrow \boldsymbol{MAP_o}$

Given one knowledge element from $\boldsymbol{MAP_s}$ "k-nn classifier" (SE11), "case retrieval" (OE9), "reasoning" (OE10), "optimization" (OE11) and "feature relevance parameters" (SE12) are the directly related research objectives, since they are contained in same abstract record. "k-nn classifier" (SE11) also used in "complete specialized CBR system" (OE14) and "construction project management" (OE13) according to the raw abstract record [66]. Thus, the initial induction is {SE11} → {*OE*9, *OE*10, *OE*11, *OE*12, *OE*13, *OE*14}.

Similarity, the connections can be extended according to the connection retrieval in the objective space. The related knowledge elements are presented in Table 10.

The extended connections can be presented as follows:

{OE9} {OE15, OE16, OE17}

{OE10} {OE15, OE16, OE17}

{OE11} {*OE*16, *OE*17, *OE*18}

{OE12} {*OE*15, *OE*17}

{OE13} {OE15, OE16, OE17}

{OE14} {OE15, OE16, OE17}

Which denotes that the research of "construction projects delay" (OE15), "indoor location tracking" (OE16), and "project selection scheduling" (OE17) are research objectives closely related to "case retrieval phase" (OE9), "Reasoning" (OE10), "construction project management" (OE13) and "complete specialized CBR system" (OE14). "Text classification algorithm" (OE18) is a similar research objective related to "optimization" (OE11). It denotes the "k-nn algorithm" (SE11) may be used in these researches. Then, a new induction can be obtained.

{SE1} → {*OE*15, *OE*16, *OE*17, *OE*18}

Through retrieving the literature, there is literature utilizing the "k-nn classifier" (SE11) to cope with the "indoor location tracking" (OE16) [67], construction management related "construction project delay" (OE15), "project selection scheduling" (OE17), and "text classification" (OE18) [68,69]. Thus, through R3 rule, given one solution, not only the directly connected research objectives can be identified, but also the potential objectives can be explored.

(4) *R*4: $(\boldsymbol{MAP_o}, \boldsymbol{MAP_s}) \rightarrow \boldsymbol{MAP_f}$

Similarity, first given one element form $\boldsymbol{MAP_o}$"construction safety" (OE19) and $\boldsymbol{MAP_s}$ "proactive construction management system (PCMS)" (SE5). The finding elements in the same abstract record are "PBBS" (FE1), "location" (FE2), "behaviors" (FE3), "safety performance" (FE5), "potential causes" (FE6), "unsafe behaviors" (FE7),

Table 11
Extracted elements from the finding space with a given knowledge element in the solution space.

Elements in Objective Space		Elements in Solution Space		Elements in Finding Space	
ID	Element	ID	Element	ID	Element
OE19	Construction Safety	SE5	Proactive construction management system (PCMS)	FE1	Proactive behavior-based safety (PBBS)
				FE2	Location
				FE3	Behaviors
				FE4	Safety management
				FE5	Safety performance
				FE6	Potential causes
				FE7	Unsafe behaviors
				FE8	Experiment results
				FE9	Construction accident prevention and safety index (SI)
				FE10	36.07% and 44.70%
				FE11	Skilled labor; unrealistic project scheduling; labor absenteeism; design changes/rework
				FE12	Numerical features
				FE13	100% and 96% recall and precision
				FE14	BLE tracking system
				FE15	Approximately 100 s
				FE16	Computational cost and memory

"safety management" (FE4), "experiment results" (FE8), "construction accident prevention and safety index (SI)" (FE9) and "36.07% and 44.70%" (FE10). Since the modifiers are removed and only the nouns terms are extracted, without the context, it is impractical to interpret the meaning of the finding elements. According to the raw abstract record, with the objective of "construction safety" (OE19) and the solution of "proactive construction management system" (SE5), the finding is that the PBBS (FE1) can monitor the location-based behaviors (FE2, FE3) and the safety performance (FE5), identify the potential causes (FE6) of unsafe behaviors (FE7) and improve the efficiency of safety management (FE4). The experiment results (FE8) showed that the PBBS (FE1) performed well on construction accident prevention and the safety index (SI) (FE9), with increased improvements by 36.07% and 44.70% (FE10) respectively. [70]. Thus, the initial induction is{OE19, SE5} → {*FE*1, *FE*2, *FE*3, *FE*4, *FE*5, *FE*6, *FE*7, *FE*8, *FE*9, *FE*10}.

Then through connection retrieving, starting from the identified knowledge elements, in the finding space, the related findings can be retrieved. The extracted elements are presented in Table 11.

Then the knowledge elements connections can be presented as follows:

{FE1} {FE11, FE12, FE13}

{FE2} {FE11, FE13, FE14}

{FE3} {FE11, FE13, FE15}

{FE4} {FE11, FE13, FE14}

{FE5} {FE11, FE13, FE14}

{FE6} {FE11, FE14, FE16}

{FE7} {FE11, FE13}

{FE8} {FE11, FE13, FE14}

{FE9} {FE11, FE13, FE14}

{FE10} {FE11, FE13, FE16}

The "skilled labor; unrealistic project scheduling; labor absenteeism; design changes/rework" (FE11) actually is identified as the primary cause of delay in the timely delivery of construction projects, according to the raw abstract record [71]. But through the connection retrieving in the find space, it has relation with the "PBBS" (FE1), "location" (FE2), "behaviors" (FE3), "safety performance" (FE5), etc. all the knowledge elements that directly connected to the solution of "proactive construction management system (PCMS)" (SE5) with the background of "construction safety" (OE19). It shows there is a great potential that the "skilled labor, unrealistic project scheduling, labor absenteeism, and design changes/rework" (FE11) has the relation with the safety management in the construction industry. Through literature retrieving, actually there is literature focused on the safety management in construction with considering the skilled labour [72], unrealistic project scheduling [73], labour absenteeism and design changes/rework [74,75] It approves that through the knowledge elements association, the potential findings can be explored. The knowledge elements "100% and 96% recall and precision" (FE13) and "BLE tracking system" (FE14) are also frequently connected to the directly associated finding elements, which can also be considered having relations with the safety management in the construction industry. The knowledge FE13 essentially is a numerical evaluation which has been widely utilized in the expression of the findings. For knowledge elements FE14, the BLE tracking system is a kind of matured area detection system and has been used in the position related management situations [76].

Thus, through the R4 rules, with given one research objective and the corresponding background, the findings can be extended in the constructed finding space. Although because of the elimination of the modifier parts of the contents, it is hard to directly interpret the knowledge elements in the finding space, the extended connections can offer the clues to explore more findings for one given topic.

4.5. The spatial form of the constructed knowledge graph

There are two types of connections among the knowledge elements: (1) connections between the spaces, (2) connections in the space. The connections between the spaces are constructed if the elements are contained in one abstract record. Because for one literature, the background, objective, solutions, and findings are pretty sure to be related. While the connections in the spaces are evaluated through the connection factor (C) defined in formula (2)–(5). Then with one given knowledge elements, the knowledge in the background, objective, solutions and finding spaces can be retrieved through the reasoning network. The overall reasoning graph can be illustrated in Fig. 10.

5. Discussion

5.1. Ontology elements identification

The ontology model is designed to represent the knowledge

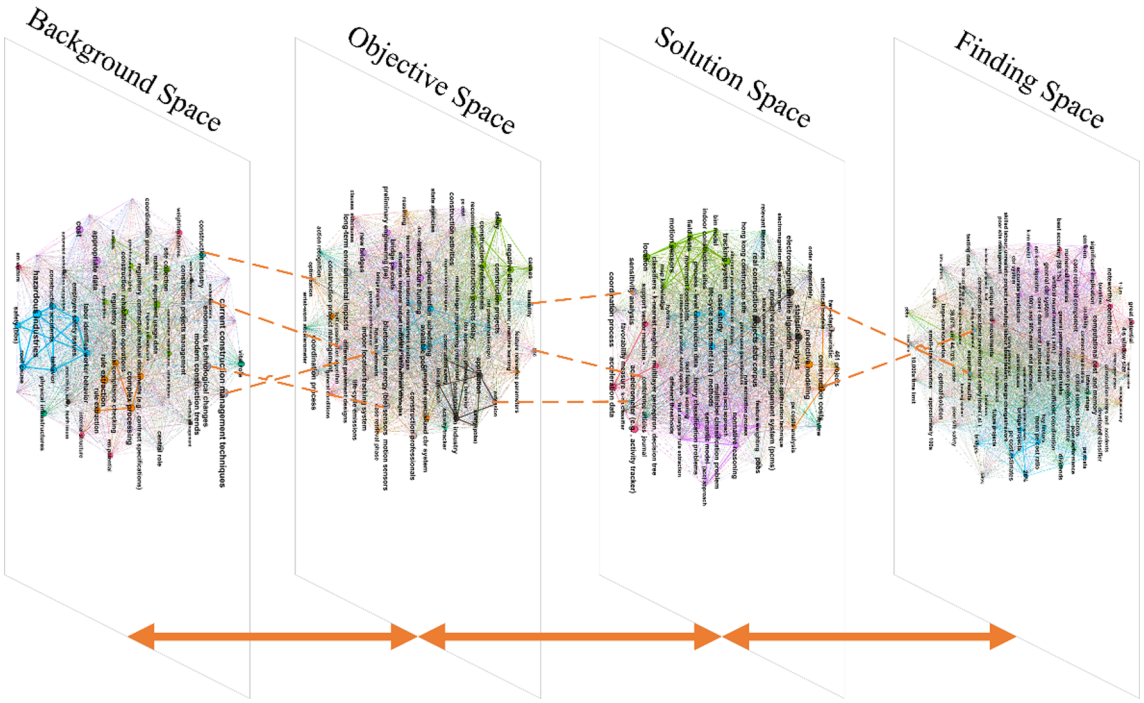

Fig. 10. Multi-space associated knowledge reasoning network.

extracted from the abstracts of literature. It is expected that the abstract data would contain the background, objective, solutions, and findings of the research. This hypothesis has been validated in this study. First through manually annotation, a batch of abstract data, extracted from different databases with different publication sources, are labeled. Each sentence is assigned to its corresponding ontology element. Since the manual label is a subjective detection, a classification model is trained to make automatic identification. The hypothesis is that the NLP-based automatic identification totally relies on the objective algorithms, thus if classification can be trained, then it proves that a pattern exists. Through the classification training and testing, it shows the training process can have a good performance. Thus, the ontology model is validated to be acceptable to represent the knowledge contents of the academic literature abstract data. The testing performance is not as good as the training one, especially for objective content detection. One possible reason is that for most situations, the statements of the objective is the shortest content in the abstract, so the training data for the objective detection is limited. Another possible reason is that the research objective has a strong relationship with the background, so the sentence describing the objective might have multiple labels (background and objective). The results also show that most of the multiple label situations are related to objective statements. Because of the multiple label issues, the detection of the objective part is not as good as the other three parts. However, the overall classification is acceptable according to the training and testing validation results. Therefore, the prosed ontology model is efficient to represent the knowledge contained in the academic literature.

5.2. Semantic expression

The semantic expression is the key point of the text classification. By employing the vectors representation, the string of text can be quantitatively represented. Based on the operations of the vectors, the manipulation of the text can also be quantified. In this work, a length of 768 units vector (for BERT core model) is used to encode the semantic meaning for the classification of ontology elements, and a length of 384 units vector (in spaCy library) is used to code the phrases for the semantic similarity evaluation of the extracted knowledge elements. The longer coding vector, the more semantic meanings can be encoded, but also the more complex of the computation. For this study, only a small knowledge graph is constructed as a demonstration, since this study processes thousands of papers, requiring hundreds of thousands of phrases to be encoded, but computational capacity is limited. On the other hand, the semantic coding is currently based on large generic knowledge corpus training, and it has been widely acknowledged to be efficient to encode the semantic information of the generic knowledge. However, each industry has its own context and professional domain knowledge. Without substantial training for the specific industry context, it is hard to tell if the semantic information of the professional domain knowledge can be correctly encoded. Unfortunately, there exists no completed and annotated corpus in the field of construction management for the training. In this work, a universal functional NLP library is utilized to conduct the vectorization of the semantic information.

5.3. Knowledge graph and knowledge reasoning

The knowledge graph is a kind of fragmented knowledge representation formula. Inspired by the human's fragmented memory, segmenting the knowledge contents into knowledge elements can facilitate knowledge reasoning and connection. In this study, to decrease the scale of the established knowledge graph, only the noun phrases are extracted to be the knowledge elements. Although the noun phrases generally are the main body of the text involved information, a large number of the strength and extent information is also eliminated by dropping the modifiers. Thus, the constructed knowledge graph only contained objective information. This makes the connections of the knowledge elements can only be objective indicators, e.g., distance, entropy, and the similarities of the coding vectors. This is the limitation of the current reasoning in the established knowledge graph.

6. Conclusion

This paper focuses on the knowledge graph building of academic literature abstract data. Summarizing this work, there are three major contributions as follows:

(a) An ontology model is designed to extract the knowledge in the abstracts of literature into background, objective, solution and finding spaces. It can be complimentary of the citation data, parsing the abstracts into a consistent structure to facilitate the further processing of the literature data.
(b) A BERT core model-based ontology elements identification model is trained and validated. Manually annotating the abstract content is time-consuming and labor-intensive, and it is impractical to cope with the large scale of literature data. With the trained classification model, the ontology elements identification can be automatically conducted.
(c) A knowledge graph is established to represent the literature data, and four knowledge deduction rules are proposed for the knowledge reasoning.

However, there are some limitations in this paper, which can be the future works of this study. The current training is based on the single label classification, because of the limited annotation data. While in practice situation, for one sentence, it may contain the background, objective information simultaneously. Thus, the multi-label classification is more practical. The current knowledge graph only considers the objective elements, all the subjective modifier information is eliminated to simplify the graph construction. This simplification might limit the reasoning ability of the knowledge graph. In the future, the strength and extent elements can be considered.

Declaration of Competing Interest

The authors declared that there is no conflict of interest.

Acknowledgement

This work was jointly supported by Shenzhen Science and Technology Innovation Committee Grant #JCYJ20180507181647320 and National Science Foundation of China Grant #51408519. The conclusions herein are those of the authors and do not necessarily reflect the views of the sponsoring agencies.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aei.2019.100959.

References

[1] P. Hager, A. Lee, A. Reich, P.O. Toole, E. Sallis, G. Jones, Knowledge Management in Organizations: A Critical Introduction, Oxford University Press, 2013, http://www.citeulike.org/group/420/article/302102.
[2] R.M. Grant, Toward a knowledge-based theory of the firm, Strateg. Manag. J. 17 (1996) 109–122 http://10.0.3.234/smj.4250171110%0Ahttp://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=12493198&site=ehost-live.
[3] A. Yakhlef, The corporeality of practice-based learning, Org. Stud. 31 (2010) 409–430, https://doi.org/10.1177/0170840609357384.
[4] V. Chang, A. Al-Hunaiyyan, A.T. Bimba, R.B. Mahmud, N. Idris, A. Abdelaziz, S. Khan, Towards knowledge modeling and manipulation technologies: a survey, Int. J. Inf. Manage. 36 (2016) 857–871, https://doi.org/10.1016/j.ijinfomgt.2016.05.022.
[5] R. Studer, V.R. Benjamins, D. Fensel, Knowledge engineering: principles and methods, Data Knowl. Eng. 25 (1998) 161–197, https://doi.org/10.1016/S0169-023X(97)00056-6.
[6] D. Sánchez, A methodology to learn ontological attributes from the web, Data Knowl. Eng. 69 (2010) 573–597, https://doi.org/10.1016/j.datak.2010.01.006.
[7] Y. Wang, On a novel cognitive knowledge base (CKB) for cognitive robots and machine learning, Int. J. Softw. Sci. Comput. Intell. 6 (2015) 41–62, https://doi.org/10.4018/ijssci.2014040103.
[8] Y. Tian, Y. Wang, M.L. Gavrilova, G. Ruhe, A formal knowledge representation system (FKRS) for the intelligent knowledge base of a cognitive learning engine, Int. J. Softw. Sci. Comput. Intell. 3 (2012) 1–17, https://doi.org/10.4018/jssci.2011100101.
[9] Y. Wang, Concept algebra: a denotational mathematics for formal knowledge representation and cognitive robot learning, J. Adv. Math. Appl. 4 (2015) 61–86, https://doi.org/10.1166/jama.2015.1074.
[10] T. Pedersen, S. Patwardhan, J. Michelizzi, WordNet: similarity: measuring the relatedness of concepts, Demonstr. Pap. HLT-NAACL 2004 (2004) 38–41.
[11] L. Jiao, T. Denoeux, Q. Pan, A hybrid belief rule-based classification system based on uncertain training data and expert knowledge, IEEE Trans. Syst. Man Cybern.: Syst. 46 (2016) 1711–1723, https://doi.org/10.1109/TSMC.2015.2503381.
[12] N. Walia, ANFIS: Adaptive neuro-fuzzy inference system – a survey, Int. J. Comput. Appl. 123 (2015) 32–38.
[13] C. Antoun, C. Zhang, F.G. Conrad, M.F. Schober, Comparisons of online recruitment strategies for convenience samples, Field Methods 28 (2015) 231–246, https://doi.org/10.1177/1525822x15603149.
[14] A. Srivastava, Parallel PageRank algorithms: a survey, International Journal on Recent and Innovation Trends in Computing and Communication, 2017, pp. 470–473.
[15] J. Alcala-Fdez, J.M. Alonso, A survey of fuzzy systems software: Taxonomy, current research trends, and prospects, IEEE Trans. Fuzzy Syst. 24 (2016) 40–56, https://doi.org/10.1109/TFUZZ.2015.2426212.
[16] B. Haymes, Towards a definition of knowledge, International Converence on Semantic Systems, Leipzig, 2016, pp. 11–15 https://doi.org/10.1007/978-1-349-19066-9_2.
[17] H. Paulheim, Knowledge graph refinement: a survey of approaches and evaluation methods, Semant. Web (2016) 1.
[18] A. Zaveri, D. Kontokostas, U. Leipzig, S. Hellmann, Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO, Semant. Web (2017) 1–53, https://doi.org/10.1016/j.jhep.2015.06.023.
[19] P. Ristoski, H. Paulheim, Semantic Web in data mining and knowledge discovery: a comprehensive survey, J. Web Semant. 36 (2016) 1–22, https://doi.org/10.1016/j.websem.2016.01.001.
[20] Y. Zhang, H. Chen, J. Lu, G. Zhang, Detecting and predicting the topic change of knowledge-based Systems: a topic-based bibliometric analysis from 1991 to 2016, Knowl.-Based Syst. 133 (2017) (1991) 255–268, https://doi.org/10.1016/j.knosys.2017.07.011.
[21] L. Waltman, A review of the literature on citation impact indicators, J. Informetr. 10 (2016) 365–391, https://doi.org/10.1016/j.joi.2016.02.007.
[22] X. Pan, E. Yan, M. Cui, W. Hua, Examining the usage, citation, and diffusion patterns of bibliometric mapping software: a comparative study of three tools, J. Informetr. 12 (2018) 481–493, https://doi.org/10.1016/j.joi.2018.03.005.
[23] D. Dou, H. Wang, H. Liu, Semantic data mining: A survey of ontology-based approaches, Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing, IEEE ICSC 2015, 2015, pp. 244–251 https://doi.org/10.1109/ICOSC.2015.7050814.
[24] M. Savonnet, E. Leclercq, P. Naubourg, EClims: an extensible and dynamic integration framework for biomedical information systems, IEEE J. Biomed. Health. Inf. 20 (2016) 1640–1649, https://doi.org/10.1109/JBHI.2015.2464353.
[25] J. Mesarić, B. Dukić, An approach to creating domain ontologies for higher education in economics, Proceedings of the International Conference on Information Technology Interfaces, 2007, pp. 75–80 https://doi.org/10.1109/ITI.2007.4283747.
[26] M. El Kharbili, P. Stolarski, Building-up a reference generic regulation ontology: a bottom-up approach, Lecture Notes in Business Information Processing, in: 37 LNBIP, 2009, pp. 268–279. https://doi.org/10.1007/978-3-642-03424-4_33.
[27] C. Feilmayr, W. Wöß, An analysis of ontologies and their success factors for application to business, Data Knowl. Eng. 101 (2016) 1–23, https://doi.org/10.1016/j.datak.2015.11.003.
[28] M. Harper, Introducing Speech and Language Processing, Pearson London, 2006, https://doi.org/10.1162/coli.2006.32.1.137.
[29] M. Gašić, D. Hakkani-Tür, A. Celikyilmaz, Spoken language understanding and interaction: machine learning for human-like conversational systems, Comput. Speech Lang. 46 (2017) 249–251, https://doi.org/10.1016/j.csl.2017.05.006.
[30] M. Koponen, Is machine translation post-editing worth the effort? A survey of research into post-editing and effort, J. Special. Transl. 25 (2016) 131–148.
[31] M.-F. Moens, Argumentation mining: how can a machine acquire world and common sense knowledge? Argum. Comput. 9 (2016) 3001, https://doi.org/10.3233/978-1-61499-686-6-4.
[32] M. Zhou, What will search engines be changed what will search engines be changed by NLP advancements, Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval, (2018) pp. 7–7 https://doi.org/10.1145/3234944.3241521.
[33] S. Biswas, D. Sengupta, R. Bhattacharjee, M. Handique, Text manipulation using regular expression, Proceedings – 6th International Advanced Computing Conference, IACC 2016, 2016, pp. 62–67 https://doi.org/10.1109/IACC.2016.22.
[34] A. Maletti, Finite-state technology in natural language processing, Theoret. Comput. Sci. 679 (2005) 2005–2007 http://www.ims.uni-stuttgart.de/institut/mitarbeiter/maletti/pub/mal16.pdf.
[35] J. Chiu, Spoken Term Detection and Spoken Word Sense Induction on Noisy Data, Dublin City University, 2015.
[36] P. Majumder, M. Mitra, B. Chaudhuri, N-gram: a language independent approach to IR and NLP, Knowl. Lang. (2002), http://www.mt-archive.info/ICUKL-2002-Majumder.pdf.
[37] E. Brill, Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging, Comput. Linguist. 21 (1995) 1–37.
[38] J.M. Conroy, D.P. O'leary, Text summarization via hidden Markov models, Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2003, pp. 406–407 https://doi.org/10.1145/383952.384042.
[39] R. Malouf, A comparison of algorithms for maximum entropy parameter estimation, Proceedings of the 6th Conference on Natural Language Learning, vol. 20, 2007, pp.

1–7 https://doi.org/10.3115/1118853.1118871.
[40] L. Shen, L. Champollion, A.K. Joshi, LTAG-spinal and the Treebank, Lang. Resour. Eval. 42 (2007) 1–19, https://doi.org/10.1007/s10579-007-9043-7.
[41] J. Pater, Generative linguistics and neural networks at 60: foundation, friction, and fusion, Language (2019) 1–34.
[42] K. Rayner, C. Clifton, Advances in natural language processing, Science 349 (2015) 261–316.
[43] B. Chiu, G. Crichton, A. Korhonen, S. Pyysalo, How to train good word embeddings for biomedical NLP, Proceedings of the 15th Workshop on Biomedical Natural Language Processing, 2016, pp. 166–174 https://doi.org/10.18653/v1/w16-2922.
[44] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, ArXiv Preprint ArXiv:1301.3781. 1 (2003) 1–12. http://arxiv.org/abs/1301.3781.
[45] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2015, pp. 1532–1543 https://doi.org/10.3115/v1/d14-1162.
[46] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, ArXiv Preprint ArXiv:1802.05365 (2018). doi: 10.18653/v1/N18-1202.
[47] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, ArXiv Preprint ArXiv:1810.04805 (2018). doi: arXiv:1811.03600v2.
[48] A.J.P. Tixier, M.R. Hallowell, B. Rajagopalan, D. Bowman, Automated content analysis for construction safety: a natural language processing system to extract precursors and outcomes from unstructured injury reports, Autom. Constr. 62 (2016) 45–56, https://doi.org/10.1016/j.autcon.2015.11.001.
[49] A.J.P. Tixier, M.R. Hallowell, B. Rajagopalan, D. Bowman, Construction safety clash detection: identifying safety incompatibilities among fundamental attributes using data mining, Autom. Constr. 74 (2017) 39–54, https://doi.org/10.1016/j.autcon.2016.11.001.
[50] M. Al Qady, A. Kandil, Concept relation extraction from construction documents using natural language processing, J. Constr. Eng. Manage. 136 (2010) 294–302, https://doi.org/10.1061/(ASCE)CO.1943-7862.0000131.
[51] J. Zhang, N.M. El-gohary, Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking, J. Comput. Civil Eng. 30 (2016) 1–42, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000346.
[52] Y. Zou, A. Kiviniemi, S.W. Jones, Retrieving similar cases for construction project risk management using Natural Language Processing techniques, Autom. Constr. 80 (2017) 66–76, https://doi.org/10.1016/j.autcon.2017.04.003.
[53] K. Hirohata, N. Okazaki, S. Ananiadou, M. Ishizuka, Identifying sections in scientific abstracts using conditional random fields, Proceedings of the Third International Joint Conference on Natural Language Processing, (2008).
[54] J.-C. Wu, Y.-C. Chang, H.-C. Liou, J.S. Chang, Computational analysis of move structures in academic abstracts, Proceedings of the COLING/ACL on Interactive Presentation Sessions, Association for Computational Linguistics, 2006, pp. 41–44 https://doi.org/10.3115/1225403.1225414.
[55] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank citation ranking: bringing order to the web, World Wide Web Internet Web Inf. Syst. 54 (1998) 1–17 doi: 10.1.1.31.1768.
[56] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, J. ACM 46 (1999) 604–632, https://doi.org/10.1145/324133.324140.
[57] L. Zhang, B. Ashuri, BIM log mining: discovering social networks, Autom. Constr. 91 (2018) 31–43, https://doi.org/10.1016/j.autcon.2018.03.009.
[58] W.M.P. Van Der Aalst, H.A. Reijers, M. Song, Discovering social networks from event logs, Comput. Support. Coop. Work 14 (2005) 549–593, https://doi.org/10.1007/s10606-005-9005-9.
[59] E. Valero, A. Sivanathan, F. Bosché, M. Abdel-Wahab, Analysis of construction trade worker body motions using a wearable and wireless motion sensor network, Autom. Constr. 83 (2017) 48–55, https://doi.org/10.1016/J.AUTCON.2017.08.001.
[60] T. Pourrostam, A. Ismail, Study of methods for minimizing construction delays: evidences from a developing country, Adv. Mater. Res. 201–203 (2011) 2939–2942, https://doi.org/10.4028/www.scientific.net/AMR.201-203.2939.
[61] H. Liu, M. Al-Hussein, M. Lu, BIM-based integrated approach for detailed construction scheduling under resource constraints, Autom. Constr. 53 (2015) 29–43, https://doi.org/10.1016/j.autcon.2015.03.008.
[62] A. Shafahi, A. Haghani, Project selection and scheduling for phase-able projects with interdependencies among phases, Autom. Constr. 93 (2018) 47–62, https://doi.org/10.1016/j.autcon.2018.05.008.
[63] Y. Yang, R. Liu, C. Deng, X. Gao, Multi-task human action recognition via exploring super-category, Signal Process. 124 (2016) 36–44, https://doi.org/10.1016/j.sigpro.2015.10.035.
[64] A. Alhamoud, V. Muradi, D. Böhnstedt, R. Steinmetz, Activity recognition in multi-user environments using techniques of multi-label classification, Proceedings of the 6th International Conference on the Internet of Things, 2016, pp. 15–23 https://doi.org/10.1145/2991561.2991563.
[65] P.E. Taylor, G.J.M. Almeida, J.K. Hodgins, T. Kanade, Multi-label classification for the analysis of human motion quality, Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2012, pp. 2214–2218 https://doi.org/10.1109/EMBC.2012.6346402.
[66] A. Kartelj, N. Šurlan, Z. Cekić, Case-based reasoning and electromagnetism-like method in construction management, Kybernetes 43 (2014) 265–280.
[67] P. Torteeka, X.I.U. Chundi, Indoor positioning based on Wi-Fi fingerprint technique using fuzzy K-nearest neighbor, Proceedings of 2014 11th International Bhurban Conference on Applied Sciences & Technology (IBCAST) Islamabad, 14th–18th January, 2014, 2014, pp. 461–465.
[68] N. Ur-Rahman, J.A. Harding, Textual data mining for industrial knowledge management and text classification: a business oriented approach, Expert Syst. Appl. 39 (2012) 4729–4739.
[69] C.H. Caldas, L. Soibelman, Automating hierarchical document classification for construction management information systems, Autom. Constr. 12 (2003) 395–406.
[70] H. Li, M. Lu, S.C. Hsu, M. Gray, T. Huang, Proactive behavior-based safety management for construction safety improvement, Saf. Sci. 75 (2015) 107–117, https://doi.org/10.1016/j.ssci.2015.01.013.
[71] M. Mbala, C. Aigbavboa, J. Aliu, Causes of delay in various construction projects: a literature review, International Conference on Applied Human Factors and Ergonomics, 2018, pp. 489–495.
[72] Z. Ismail, S. Doostdar, Z. Harun, Factors influencing the implementation of a safety management system for construction sites, Saf. Sci. 50 (2012) 418–423, https://doi.org/10.1016/j.ssci.2011.10.001.
[73] P.X.W. Zou, G. Zhang, J. Wang, Identifying key risks in construction projects: life cycle and stakeholder perspectives related past research and risk classification, Architecture (1993) 1–14.
[74] J.A. Gambatese, M. Asce, M. Behm, J.W. Hinze, Viability of designing for construction worker safety, J. Constr. Eng. Manage. @ ASCE 131 (2005) 1029–1036, https://doi.org/10.1061/(ASCE)0733-9364(2005)131:9(1029).
[75] I. Nahmens, L. Ikuma, An empirical examination of the relationship between lean construction and safety in the industrialized housing industry, Lean Constr. J. 1 (2009) 1–12.
[76] J. Park, K. Kim, Y.K. Cho, Framework of automated construction-safety monitoring using cloud-enabled BIM and BLE mobile tracking sensors, J. Constr. Eng. Manage. 143 (2016) 5016019.