Automation in Construction xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Automation in Construction

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

Retrieving similar cases for construction project risk management using Natural Language Processing techniques

Yang Zou[a,*], Arto Kiviniemi[b], Stephen W. Jones[a]

[a] School of Engineering, University of Liverpool, Brownlow Hill, Liverpool, L69 3GH, UK
[b] School of Architecture, University of Liverpool, Leverhulme Building, Liverpool, L69 7ZN, UK

ARTICLE INFO

Keywords:
Risk management
Case-based reasoning (CBR)
Natural Language Processing (NLP)
Vector Space Model (VSM)
Query expansion
Case retrieval

ABSTRACT

Case-based reasoning (CBR) is an important approach in construction project risk management. It emphasises that previous knowledge and experience of accidents and risks are highly valuable and could contribute to avoiding similar risks in new situations. In the CBR cycle, retrieving useful information is the first and the most important step. To facilitate the CBR for practical use, some researchers and organisations have established construction accident databases and their size is growing. However, as those documents are written in everyday language using different ways of expression, how information in similar cases is retrieved quickly and accurately from the database is still a huge challenge. In order to improve the efficiency and performance of risk case retrieval, this paper proposes an approach of combining the use of two Natural Language Processing (NLP) techniques, i.e. Vector Space Model (VSM) and semantic query expansion, and outlines a framework for this Risk Case Retrieval System. A prototype system is developed using the Python programming language to support the implementation of the proposed method. Preliminary test results show that the proposed system is capable of retrieving similar cases automatically and returning, for example, the top 10 similar cases.

1. Introduction

Construction is among the most hazardous and dangerous industries in the world [1]. In the U.S., it is reported that over 157 bridges collapsed between 1989 and 2000 [2], and > 26,000 workers lost their lives on construction sites during the past two decades [3]. Globally, the International Labour Organization (ILO) estimates that approximately 60,000 fatal accidents happen every year [4]. Such serious accidents may not only lead to a bad reputation for the construction industry but also trigger further risks such as project failure, financial difficulty and time overruns. To avoid such serious accidents and improve the performance of risk management in future projects, a few studies [5,6] suggested project practitioners should learn the valuable lessons from previous accidents and embed the consideration of risk management into the development process of a project. Learning from the past is a fundamental process in project risk management that helps individuals and organisations understand when, what and why incidents happened, and how to avoid repeating past mistakes [7].

In general, the process of solving new problems based on experience of similar past problems is known as Case-Based Reasoning (CBR) [8], which examines what has taken place in the past and applies it to a new situation [9], and could be of particular help in identifying and mitigating project risks at early stages, e.g. design and construction planning. In order to facilitate CBR for practical use in the construction industry, some efforts have been observed in collecting risk cases and establishing a risk case database. For example, Zhang et al. [10] developed a database containing 249 incident cases to support risk management for metro operations in Shanghai. And there are > 600 verified reports about structural risks on the Structural-Safety website [11] and similarly the National Institute for Occupational Safety and Health (NIOSH) [12] has established a database of over 249 reports on construction accidents. In addition, for identifying the reasons that contribute to collision injuries, Esmaeili and Hallowell [13] reviewed and analysed over 300 accident reports. However, as a risk case database often contains a huge amount of data where reports are written in everyday language, manually reviewing, analysing and understanding these reports is a time-consuming, labour-intensive and inefficient work. Failure in extracting 'correct' cases and information within a limited time often may mean that the importance of learning from past experience is missed. Hence, some researchers [7,14,15] pointed out that a key challenge in current CBR research for project risk management is how to quickly and accurately retrieve relevant risk case data from the database so that knowledge and experience could be incorporated into new risk identification and assessment in a timely manner.

* Corresponding author.
E-mail address: yang.zou@liverpool.ac.uk (Y. Zou).

http://dx.doi.org/10.1016/j.autcon.2017.04.003
Received 9 September 2016; Received in revised form 9 February 2017; Accepted 5 April 2017

In recent years, with the development and growing use of Natural Language Processing (NLP) in the computer science discipline, some researchers have been trying to introduce NLP into the construction industry to address the analysis and management issues of textual documents, e.g. retrieval of CAD drawings [16], automatic analysis of injury reports [14], and automatic clustering of construction project documents based on textual similarity [17]. It could be seen that NLP is a promising technique in assisting the knowledge and case retrieval of CBR. However, very few studies have been found in this field. In addition, Goh and Chua [7] stated that very few NLP tools nowadays appear to be suitable for the construction industry.

In order to improve the efficiency and performance of risk case retrieval, this paper proposes an approach of combining the use of two NLP techniques, i.e. Vector Space Model (VSM) and semantic query expansion, and outlines a framework for the Risk Case Retrieval System. A prototype system is developed with the Python programming language to support the implementation of the proposed method.

The rest of this paper is organised as follows. Section 2 introduces the background and current challenges of CBR in project risk management, and discusses the potential of integrating NLP in CBR and the motivation of this study. The system architecture and methodologies used in this study are described in Section 3. In Section 4, a prototype system is developed with Python. A simple example is used for illustrating the proposed method, and a preliminary test is conducted to evaluate the system. Finally, the implications, limitations, recommendations for future research and conclusions are addressed in Sections 5 and 6.

2. Background and point of departure

2.1. Current challenges in case retrieval

CBR is a branch of Artificial Intelligence (AI) and its origin can be traced back to the work of Roger Schank and his students in the early 1980s [15,18,19]. The core philosophy behind CBR is that previous knowledge and experience can be recalled and used as a starting point to solve new problems in many fields. In the project management domain, CBR has been recognised as an important technique for risk identification and analysis [20] and a number of applications have been developed, e.g. construction hazard identification [7,21], safety risk analysis in subway operations [22], and construction supply chain risk management [23]. Fig. 1 shows the classical model of a CBR system adapted from a previous research by Aamodt and Plaza [24]. Basically the implementation cycle of CBR contains four main processes: RETRIEVE, REUSE, REVISE, and RETAIN (known as 'the four REs'), where RETRIEVE is the first and the most important process in any CBR systems [22].

RETRIEVE is a process of searching and determining the most similar and relevant case or cases [15,24], and its importance can be viewed from the following three main aspects: (1) it acts as the only medium for helping individuals extract information from a risk case database; (2) as a risk case database often contains a large number of 'human language' based documents, the performance of case retrieval will have direct influence on the quality and accuracy of retrieved cases; and (3) the inefficiency of case retrieval seriously affects the user experience, which may lead to the importance of previous knowledge and experience being overlooked.

Currently scoring the similarity through allocating weights to factors is the most common method in case retrieval. For example, Lu et al. [22] employed a semantic network approach to calculate the similarity value between two accident precursors. Karim and Adeli [25] collected risk data into Excel tables and developed an attribute based schema for calculating the similarity between two cases. Goh and Chua [7] proposed a sub-concept approach based on a semantic network. Other efforts include, for example, evaluation of attributes [9], taxonomy tree approach [26], ontology-based method [27].

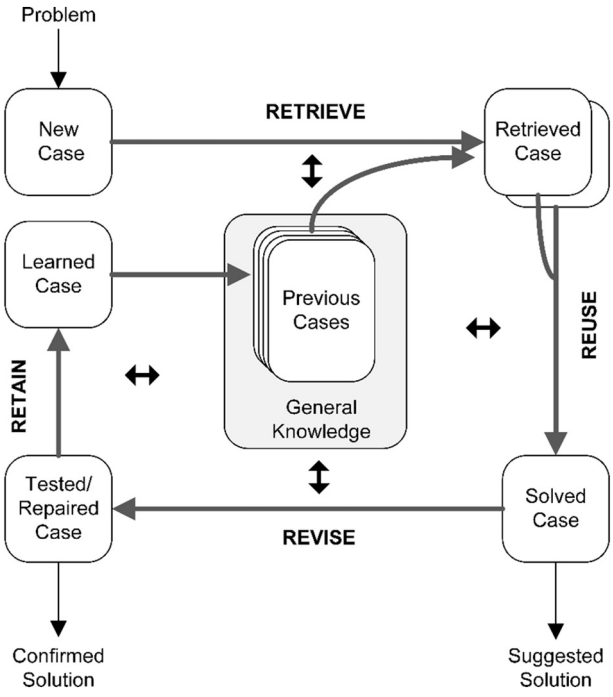

Fig. 1. Classical model of a CBR system [24].

However, challenges and limitations also exist in current efforts, which are summarised as follows:

(1) Existing studies are very limited in scope. For example, the CBR system developed by Lu et al. [22] predefined the potential accidents in subway operations and the similarity calculation is based on attributes that are to some extent subjective. Similarly, the prototype proposed by Karim and Adeli [25] calculated the similarity index based on different weights of attributes and is only designed for highway work zone traffic management.
(2) A large amount of pre-processing or preparation work is needed. For instance, the sub-concept approach [7] needs to establish a semantic network map of variables and each semantic network is constructed based on analysis of cases and allocation of weights. Goh and Chua [7] acknowledged that organisations implementing the system need to consider the cost for establishing and maintaining the semantic networks and risk cases.
(3) Very few studies have been found in addressing the challenge of semantic similarity in case retrieval. Semantic similarity is defined as "*a metric defined over a set of terms or documents, where the idea of distance between them is based on the likeness of their meaning or semantic content as opposed to similarity which can be estimated regarding their syntactical representation*" [28]. Semantic similarity problems can be observed in, for example, synonyms (e.g. 'building' and 'house'), hyponyms (e.g. 'structure' and 'bridge'), and even related words (e.g. 'car' and 'bus'). Because risk case reports are all written in everyday human language and in different ways of expressing meaning by different individuals or organisations, the outcomes of case retrieval will be incomplete if a CBR system fails to consider semantic similarity. Therefore, Mantaras et al. [15] pointed out that improving the performance through more effective approaches to similarity assessment has been an important research focus in CBR.

2.2. Natural Language Processing

Natural Language Processing (NLP) is an interdisciplinary topic overlapping in computational linguistics, AI, and computer science that deals with the interactions between computer and human languages