

Accepted by Elsevier Journal of Automation in Construction for publication on 5 April 2017.  
DOI: 10.1016/j.autcon.2017.04.003.

---

1     **Retrieving similar cases for construction project risk**  
2             **management using natural language processing**  
3                     **techniques**

4     Yang Zou\*, [Yang.Zou@liverpool.ac.uk](mailto:Yang.Zou@liverpool.ac.uk)  
5     School of Engineering, University of Liverpool, Liverpool, UK

6     Arto Kiviniemi, [A.Kiviniemi@liverpool.ac.uk](mailto:A.Kiviniemi@liverpool.ac.uk)  
7     School of Architecture, University of Liverpool, Liverpool, UK

8     Stephen W. Jones, [Stephen.Jones@liverpool.ac.uk](mailto:Stephen.Jones@liverpool.ac.uk)  
9     School of Engineering, University of Liverpool, Liverpool, UK

10    \*Corresponding author: Yang Zou ([Yang.Zou@liverpool.ac.uk](mailto:Yang.Zou@liverpool.ac.uk))

11    **Abstract**

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

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

## 30 **1. Introduction**

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

44 In general, the process of solving new problems based on experience of similar past  
45 problems is known as Case-Based Reasoning (CBR) [8], which examines what has  
46 taken place in the past and applies it to a new situation [9], and could be of particular  
47 help in identifying and mitigating project risks at early stages, e.g. design and  
48 construction planning. In order to facilitate CBR for practical use in the construction  
49 industry, some efforts have been observed in collecting risk cases and establishing a  
50 risk case database. For example, Zhang et al. [10] developed a database containing 249  
51 incident cases to support risk management for metro operations in Shanghai. And there  
52 are more than 600 verified reports about structural risks on the Structural-Safety  
53 website [11] and similarly the National Institute for Occupational Safety and Health

54 (NIOSH) [12] has established a database of over 249 reports on construction accidents.  
55 In addition, for identifying the reasons that contribute to collision injuries, Esmaeili and  
56 Hallowell [13] reviewed and analysed over 300 accident reports. However, as a risk  
57 case database often contains a huge amount of data where reports are written in  
58 everyday language, manually reviewing, analysing and understanding these reports is  
59 a time-consuming, labour-intensive and inefficient work. Failure in extracting ‘correct’  
60 cases and information within a limited time often may mean that the importance of  
61 learning from past experience is missed. Hence, some researchers [7,14,15] pointed out  
62 that a key challenge in current CBR research for project risk management is how to  
63 quickly and accurately retrieve relevant risk case data from the database so that  
64 knowledge and experience could be incorporated into new risk identification and  
65 assessment in a timely manner.

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

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

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

## 88 **2. Background and point of departure**

### 89 **2.1. Current challenges in case retrieval**

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

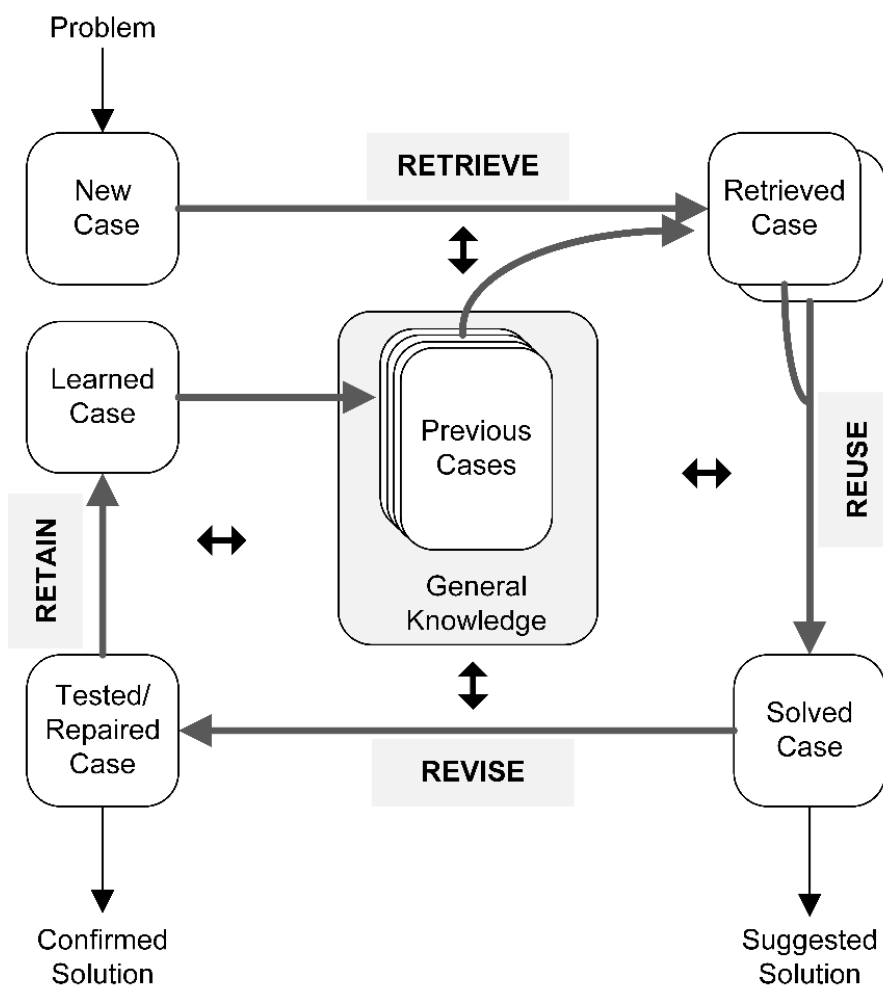


Figure 1 Classical model of a CBR system (Adapted from [24])

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

114 network approach to calculate the similarity value between two accident precursors.  
115 Karim and Adeli [25] collected risk data into Excel tables and developed an attribute  
116 based schema for calculating the similarity between two cases. Goh and Chua [7]  
117 proposed a sub-concept approach based on a semantic network. Other efforts include,  
118 for example, evaluation of attributes [9], taxonomy tree approach [26], ontology-based  
119 method [27].

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

122 (1) Existing studies are very limited in scope. For example, the CBR system developed  
123 by Lu et al. [22] predefined the potential accidents in subway operations and the  
124 similarity calculation is based on attributes that are to some extent subjective. Similarly,  
125 the prototype proposed by Karim and Adeli [25] calculated the similarity index based  
126 on different weights of attributes and is only designed for highway work zone traffic  
127 management.

128 (2) A large amount of pre-processing or preparation work is needed. For instance, the  
129 sub-concept approach [7] needs to establish a semantic network map of variables and  
130 each semantic network is constructed based on analysis of cases and allocation of  
131 weights. Goh and Chua [7] acknowledged that organisations implementing the system  
132 need to consider the cost for establishing and maintaining the semantic networks and  
133 risk cases.

134 (3) Very few studies have been found in addressing the challenge of semantic similarity  
135 in case retrieval. Semantic similarity is defined as “*a metric defined over a set of terms  
136 or documents, where the idea of distance between them is based on the likeness of their  
137 meaning or semantic content as opposed to similarity which can be estimated regarding  
138 their syntactical representation*” [28]. Semantic similarity problems can be observed in,

139 for example, synonyms (e.g. ‘building’ and ‘house’), hyponyms (e.g. ‘structure’ and  
140 ‘bridge’), and even related words (e.g. ‘car’ and ‘bus’). Because risk case reports are  
141 all written in everyday human language and in different ways of expressing meaning  
142 by different individuals or organisations, the outcomes of case retrieval will be  
143 incomplete if a CBR system fails to consider semantic similarity. Therefore, Mantaras  
144 et al. [15] pointed out that improving the performance through more effective  
145 approaches to similarity assessment has been an important research focus in CBR.

## 146 **2.2. Natural Language Processing**

147 Natural language processing (NLP) is an interdisciplinary topic overlapping in  
148 computational linguistics, AI, and computer science that deals with the interactions  
149 between computer and human languages [29]. NLP started its early work in the 1950s  
150 in exploring the fully automatic translation between different languages [30], and in  
151 recent years has seen a rapid increase in use and development in computer science. The  
152 application areas of NLP are very wide including, for example, machine translation,  
153 question answering, speech recognition and information retrieval [31].

154 Information retrieval (IR) refers to the process and activity of extracting useful  
155 information from a collection of information resources [32]. Due to the needs of  
156 managing and using the fast-growing volume of information [33], many IR systems  
157 have been developed and the best examples include web search engines (e.g. Google  
158 and Yahoo), and library resource retrieval systems [34].

159 In the construction industry, even a small project generates a large amount of digital  
160 information such as specifications, computer-aided drawings, and structural analysis  
161 reports [14,35]. In addition, in order to learn from past experience and avoid similar  
162 accidents in new projects, lots of investigations and analysis on previous accidents have  
163 been conducted and the resulting reports and feedbacks are important to improving the

164 existing knowledge and standards [36]. Currently major companies and organisations  
165 are using databases for managing those accident reports [14]. However, new documents  
166 continually need to be added into databases and therefore the size of databases is  
167 increasing. Moreover, these reports are written in human language and in different ways  
168 of expression by different individuals or organisations. As discussed in Section 2.1, a  
169 challenge is how to retrieve valuable and ‘correct’ information from the database  
170 quickly and efficiently.

171 To improve the use and management of ‘human language’ based engineering  
172 documents, a recent research trend is to take advantage of NLP. For example, Yu and  
173 Hsu [16] made the use of the classical VSM and developed a Content-based CAD  
174 document Retrieval System (CCRS) for assisting the management of CAD drawings  
175 and quick retrieval of documents according to given queries. By taking the advantage  
176 of keywords extraction of NLP, Tixier et al. [14] developed a prototype supported by  
177 the R programming language for automatically extracting precursors and outcomes  
178 from unstructured injury reports. Qady and Kandil [17] proposed a method for  
179 automatic clustering of construction project documents based on textual similarity.  
180 Caldas and Soibelman [37] developed a prototype system to automatically classify a  
181 large number of electronic text documents in a hierarchical order in the information  
182 management system. Another study took the advantage of text mining and proposed an  
183 ontology-based text classification method for job hazard analysis [38]. In addition,  
184 Pereira et al. [39] presented a solution to extract valuable information from incident  
185 reports in real time to assist incident duration prediction. However, very few studies  
186 exist in this field and new investigations are still needed.

187 It is observed that there are two main features in applying NLP into textual document  
188 management in the construction industry:



- 189       • Firstly, most state-of-the-art studies of NLP still lie in the computer science  
190       discipline and most modern applications are often used to treat extremely large  
191       volumes of data e.g. extracting online information [40] and library management  
192       [32]. In contrast, the sizes of electronic data in any construction project and risk  
193       cases in any database are relatively small. Hence, there is a need to select the  
194       appropriate methods and techniques for specific purposes. For example, Tixier  
195       et al. [14] pointed out one difficulty in implementing machine learning for  
196       automatic safety keywords extraction is that small number of injury reports is  
197       not satisfactory as a training database and therefore they developed a NLP  
198       system based on hand-coded rules.
- 199       • Secondly, unlike online webs containing often several aspects of information,  
200       construction project data and risk cases are relatively restricted to certain topics  
201       and thus there is a need to establish the context or rules in processing them. For  
202       instance, when applying ontology and text mining into job hazard analysis, the  
203       authors predefined the list of potential safety hazards and emphasised the  
204       importance of defining the knowledge and resource scope into the construction  
205       safety domain [16].

### 206   **2.3. Motivation and aim of this study**

207   As discussed in Sections 2.1 and 2.2, some existing efforts [14,16,17] have shown that  
208   the application of NLP techniques in managing textual data is a new research trend in  
209   the construction industry and NLP has the potential to address the current challenges of  
210   case retrieval of CBR. However, very limited numbers of studies have been found in  
211   this area. In order to further improve the efficiency and performance of risk case  
212   retrieval, this paper proposes an approach of combining the use of two NLP techniques,  
213   i.e. VSM and semantic query expansion, and outlines a framework for the risk case  
214   retrieval system. The idea was motivated by the following observations:

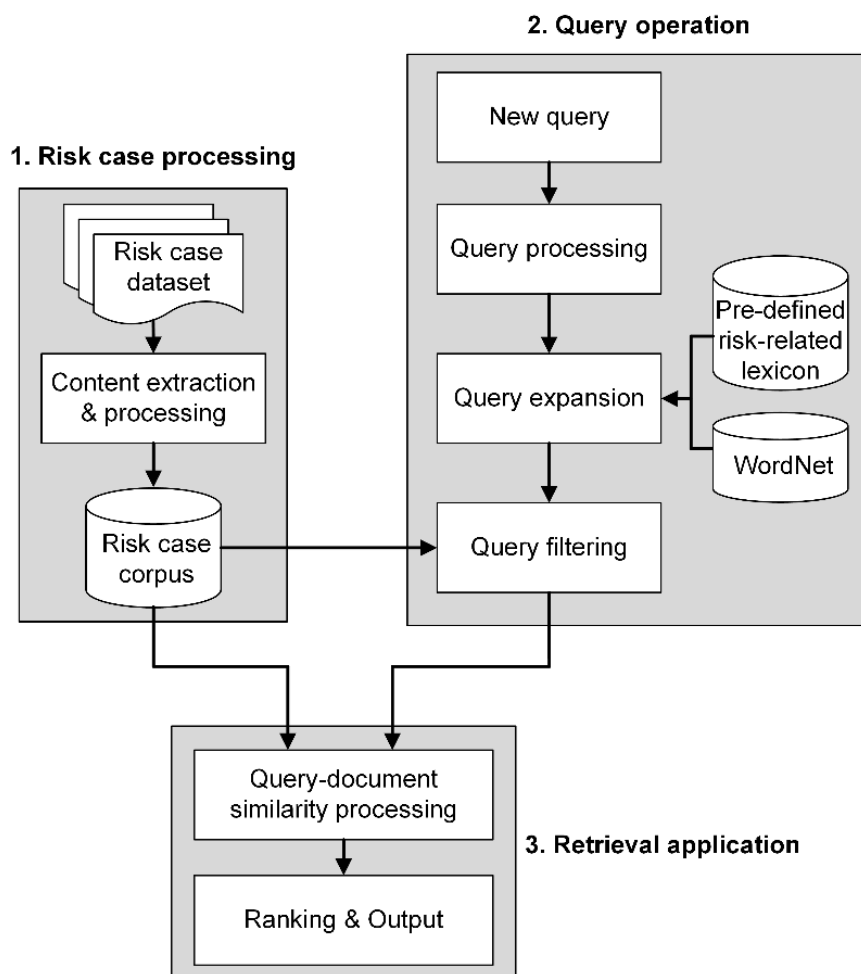
- 215       • VSM is known as one of the most important IR models [32] and it can be used  
216       for information extraction, indexing and relevancy ranking, etc. For example,  
217       Caldas and Soibelman [37] used VSM for characteristic information extraction  
218       and automatic classification of project documents. Similarly, Yu and Hsu [16]  
219       embedded VSM as a core technique in their retrieval system of CAD drawings.  
220       Hence, VSM is potentially helpful in evaluating the relevance between user  
221       need and risk cases in a CBR system.
- 222       • Understanding the relations between words (e.g. hyponymy, synonymy) is an  
223       important step in fully using the concept of semantic similarity [31]. Thus, some  
224       individuals and organisations have started to establish lexical ‘dictionaries’ that  
225       pre-defined the semantic relationships between words, where the most  
226       commonly used resource for English sense relations is the WordNet lexical  
227       database [31,41]. So far a number of studies [42,43] have used WordNet for  
228       improving web retrieval through expanding the query terms using related words  
229       in WordNet and have proved this approach could partially address the semantic  
230       similarity issues and improve the performance and completeness of information  
231       retrieval. Therefore, the basic principle of semantic query expansion is also  
232       applicable for improving the completeness and quality of case retrieval.

### 233   **3. Framework and methodology**

234   The overall framework and methodologies used in this study are described in this  
235   section. Specifically, the system architecture of the proposed Risk Case Retrieval  
236   System (RCRS) is presented in Section 3.1, and the three major modules of RCRS are  
237   described in detail in Sections 3.2, 3.3 and 3.4.

### 238 **3.1 System architecture of the Risk Case Retrieval System**

239 The system architecture of the proposed RCRS is illustrated in Figure 2. The system  
240 consists of three major modules, i.e. (1) Risk case processing, (2) Query operation, and  
241 (3) Retrieval application. Firstly, the risk case processing module automatically extracts  
242 the textual information from a targeted collection of risk cases. It processes the  
243 collected textual information by a defined Sequence of Actions (SoA), i.e. tokenisation,  
244 converting all words into lowercase, lemmatisation, and removing stop words to  
245 establish a risk case content corpus. The SoA is a general approach in current NLP for  
246 processing textual documents [31]. Secondly, the query operation module reads and  
247 processes the given query by SoA. The processed query is prior scanned to match its  
248 expansion of related words in the pre-defined risk-related lexicon. The terms not found  
249 in the pre-defined risk-related lexicon are expanded by using synonyms in WordNet.  
250 Then the system scans the terms in both the original query and the expanded query, and  
251 removes those terms that do not exist in the risk case content corpus. Thirdly, the  
252 retrieval application module combines the queries and risk case corpus together and  
253 performs the query-document similarity calculations. After this, the system ranks all  
254 documents according to their similarity scores and finally returns, for example, the top  
255 10 documents to the users.



256

257

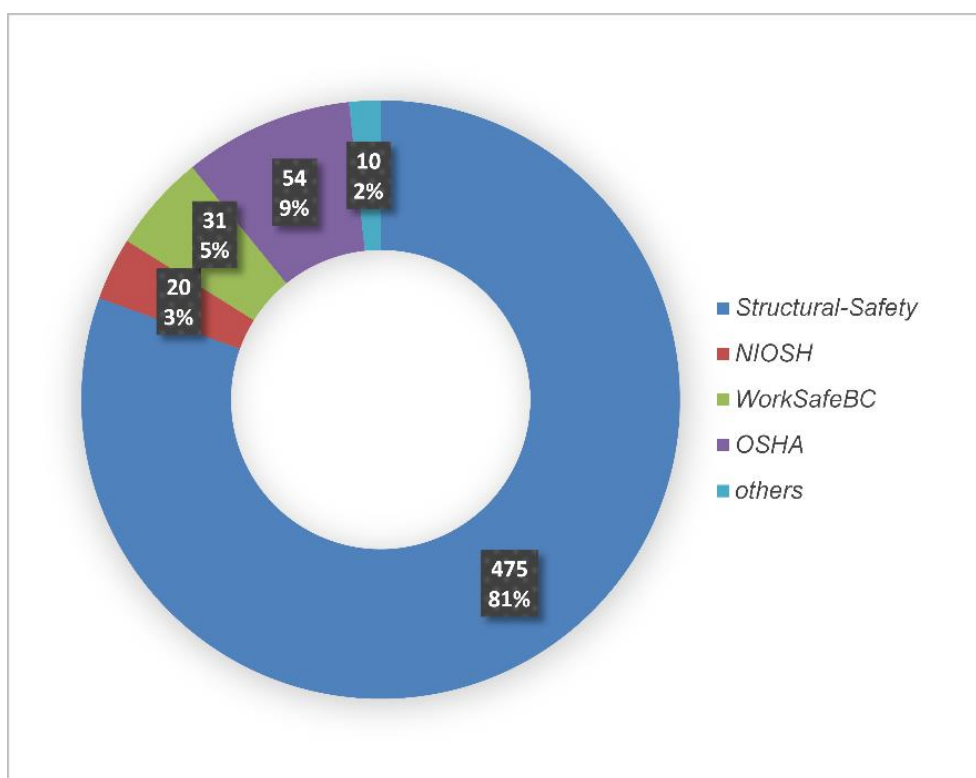
Figure 2 System architecture of RCRS

### 258 3.2 Risk case processing workflow

259 The first step in the risk case processing module is to collect risk cases through a web  
260 search method. In total 590 risk cases were collected from the following major  
261 organisational and governmental construction accident databases: (1) Structural-Safety  
262 [11], (2) the National Institute for Occupational Safety and Health (NIOSH) [12], (3)  
263 WorkSafeBC [44], (4) Occupational Safety and Health Administration [45], and (5)  
264 others (e.g. some published papers that document construction accidents). The source  
265 distribution of collected risk cases is shown in Figure 3 and the category distribution is  
266 presented in Figure 4. Although collecting as many risk cases as possible from every  
267 category of project risks could improve the reliability of the proposed approach, this

268 study stopped collecting more cases due to the following reasons: (1) the authors have  
269 only limited research time and the main focus of this study is developing a NLP based  
270 general approach for risk case retrieval instead of establishing a complete risk case  
271 database; (2) it is observed that some risks (e.g. collapse of structure, loss of life) that  
272 may lead to severe consequences attract more attention while there are very few detailed  
273 reports available on those risks that are not so dangerous, e.g. financial loss, time  
274 overrun.

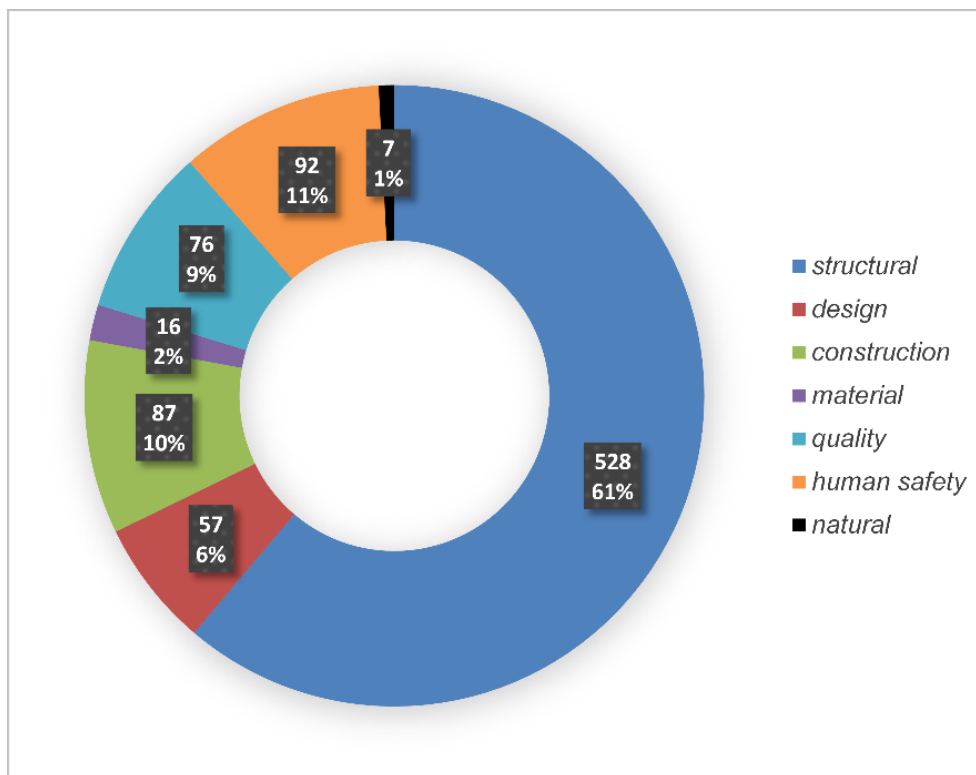
275



276

277

Figure 3 Source distribution of collected risk cases



278

279

Figure 4 Category distribution of collected risk cases

280 The second step is to extract the textural information from the collected reports and  
281 process them to be a risk case content corpus, which goes through the following  
282 processes:

- 283
- **Tokenisation:** this is a process of chopping a document up into pieces (known  
284 as ‘tokens’) and discarding certain characters, such as punctuation [46]. An  
285 example is illustrated in Figure 5.

**Input:** Building , site , construction , safety ?

**Output:**

Building	site	construction	safety
----------	------	--------------	--------

286

287

Figure 5 An example of tokenisation

- 288
- **Converting words into lowercase:** this is a simple task to convert tokens into  
289 lowercase, which could improve the search results [46]. For instance, the term  
290 “Building” is converted to be “building”.

- 291       • **Lemmatisation:** it “usually refers to doing things properly with the use of a  
292       *vocabulary and morphological analysis of words, normally aiming to remove*  
293       *inflectional endings only and to return the base or dictionary form of a word,*  
294       *which is known as the lemma”* [46]. For example, the base form “walk” may  
295       appear as “walk”, “walked”, “walks”, or “walking” in the main text, and the  
296       process of lemmatisation is to convert those words to their base forms.
- 297       • **Stop words removal:** stop words are those extremely common words which  
298       have little value in helping match documents [46]. Removal of those  
299       meaningless words could largely reduce the size of collection and improve the  
300       retrieval efficiency. The stop words used in this study are presented in Table 1  
301       which consists of two sub lists. The first list of stop words is identified by the  
302       Natural Language Toolkit (NLTK) [47], which is a suite of libraries and  
303       programs for symbolic and statistical NLP for English written in the Python  
304       programming language [48]. The second list comes from a manual selection  
305       from the top 100 words that have the most occurrences in the risk case content  
306       corpus but are identified with little value. For example, ‘fig 1’ has an extremely  
307       high occurrences in the whole risk case collection but its tokens (i.e. ‘fig’ and  
308       ‘1’) are of little help to the risk case retrieval. Because there are still some  
309       limitations in current NLP techniques [16], some meaningless words are  
310       produced after Tokenisation, e.g. the symbol underline and the letter “j”.  
311       Removal of these manually selected meaningless words with the highest  
312       numbers of occurrence could effectively reduce the size of data and this method  
313       has been adopted in some previous studies, e.g. Fan and Li [49].
- 314       • **Establishing the risk case corpus:** corpus in the NLP context refers to a large  
315       collection of texts [31] and this process is to combine the processed textual  
316       information into a corpus for further use in the query operation and retrieval  
317       application.

318 Table 1 Stop words used in this paper

Stop words identified by NLTK					Manually selected stop words
the	his	off	him	about	number
couldn	ain	with	doesn	re	15
shan	were	m	an	our	20
between	very	but	who	both	could
any	there	own	was	he	14
himself	while	for	during	this	16
a	hers	is	once	until	f
at	over	too	other	am	b
after	myself	just	ll	no	12
will	then	i	again	mightn	fig
ma	it	wasn	being	hadn	11
its	against	by	yourselves	through	-
o	these	how	not	because	0
what	ve	them	can	out	e
don	her	in	up	if	would
does	are	from	on	mustn	also
didn	wouldn	under	having	below	j
most	theirs	down	of	shouldn	may
same	whom	only	each	aren	r
their	s	where	y	do	10
and	you	all	nor	isn	9
did	now	haven	herself	have	1
your	as	yourself	t	yours	c
which	won	into	should	above	7
further	itself	been	she	me	1
few	needn	d	ours	my	6
to	or	such	weren	here	5
so	why	had	than	more	4
they	before	some	that	themselves	3
those	be	we	hasn		2
when	doing	ourselves	has		

319 **3.3 Query operation process**

320 A basic semantic similarity problem is often observed that terms of the original query  
 321 are different to the ones used in the documents in describing the same semantics [42].

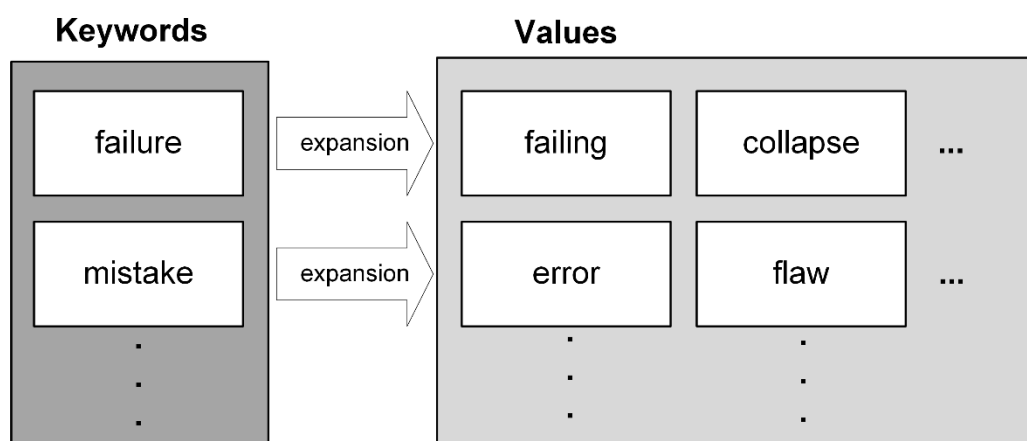


322 To deal with the mismatching problem, a promising solution is to use query expansion  
323 [42,50,51]. In definition, query expansion is a process of reformulating or expanding a  
324 seed query using semantically related words (e.g. hyponyms, synonyms) to improve the  
325 retrieval performance of IR systems [52]. Many web IR efforts have adopted this  
326 approach and a common way is to extract the semantically related words from WordNet  
327 [41-43], a lexical database for the English language.

328 Because the collected risk cases are in different styles of expression by different  
329 individuals or organisations, the above problem also commonly exists in the risk case  
330 database, e.g. “structural failure” and “structure collapse”. Therefore this paper  
331 integrates query expansion into the RCRS for this mismatching problem. However,  
332 WordNet is a relatively complete lexical database for the whole English environment  
333 and contains too much data which is not useful for the risk case retrieval context. For  
334 example, the synonyms of “failure” are “nonstarter”, “loser” and “unsuccessful person”  
335 which are not related to project risk management. In addition, no such dictionary or  
336 database has been found for defining the semantically related words in a risk  
337 management context. Hence, this paper established a small risk-related lexicon to  
338 overcome this limitation and combines the use of this risk-related lexicon and WordNet.

339 The pre-defined risk-related lexicon is a dictionary consisting of 107 key words, which  
340 are most commonly used in the risk management context, and their expansion  
341 suggestions. An example is shown in Figure 6. To develop the lexicon, three major  
342 steps were used. Firstly, the 107 key words (e.g. “building”, “risk”, “collapse”,  
343 “change”, “safety”) were manually selected from all risk factors in a risk database  
344 established by a previous study [53]. The second step performed a deep learning  
345 approach to find out the most related words ( i.e. “Values” in Figure 6) of 107 key  
346 words by using Word2vec [54,55], a deep learning algorithm developed by a research  
347 group led by Tomas Mikolov at Google. Word2vec is an unsupervised learning tool for

348 obtaining vector representations for words and could be used for finding out most  
349 similar or related words in an N-dimensional vector environment. The collected 590  
350 risk cases were initially used for training but it was quickly realised the size of data was  
351 so small that the performance of calculation is not as good as the authors expected.  
352 Then, the free and open Wikipedia content database [56] is used as a supplement for  
353 calculating the most similar words. In the third step, similar words calculated by using  
354 both risk case content corpus and Wikipedia content database are gathered together and  
355 a manual selection process based on knowledge and experience is conducted to delete  
356 words that are not related to the risk management context.



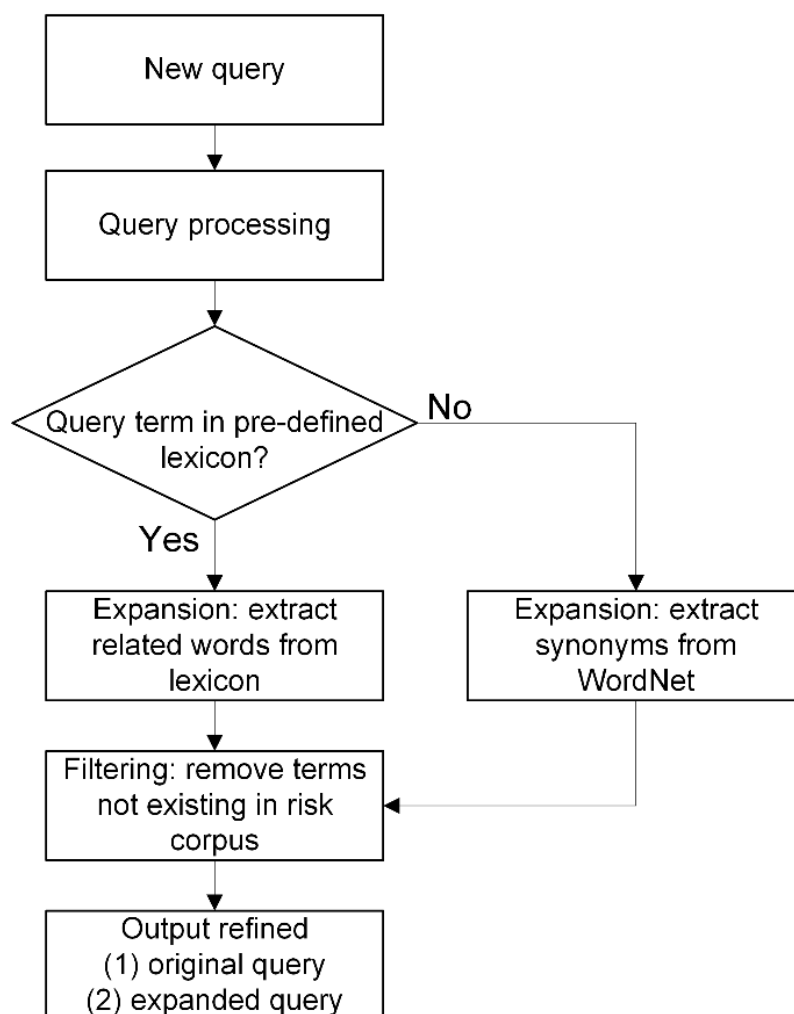
357

358

Figure 6 Example of risk-related lexicon

359 The work flow of query expansion is shown in Figure 7. Specifically, a new query is  
360 firstly read and processed by SoA. Secondly the processed query terms are prior  
361 scanned to match its expansion of related words in the pre-defined risk-related lexicon.  
362 If any terms are not found in the pre-defined risk-related lexicon, they are expanded by  
363 using synonyms in WordNet. After this, there are two queries, i.e. original query,  
364 expanded query. With the observation that original query could mostly reflect a user's  
365 need for case retrieval, this paper keeps the original query and expanded query as two  
366 separate queries. Thirdly, the system scans the terms in both original query and  
367 expanded query, and removes terms that do not exist in the risk case content corpus.

368 Lastly, the system outputs both refined original query and expanded query for further  
369 use in retrieval application.



370

371

Figure 7 Work flow of query expansion

### 372 3.4 Retrieval application process

#### 373 3.4.1 The classical Vector Space Model (VSM)

374 In definition, the VSM is an algebraic model for representing textual documents as  
375 vectors of identifiers and assigning non-binary weights to index terms in queries and in  
376 documents, which is broadly used to compute the degree of similarity between each  
377 document and the query [32,57,58]. The classical VSM is described as follows [32]:

378 Query  $q$  and document  $d_j$  can be represented as  $t$ -dimensional vectors, as shown in  
379 Equations (1) and (2). For the vector model,  $t$  is the total number of index terms and  
380 each dimension corresponds to a separate index term. The elements  $w_{i,j}$  in each vector  
381 is the weight associated with a term-document pair  $(k_i, d_j)$  and  $w_{i,j} \geq 0$ .

$$382 \quad \vec{q} = (w_{1,q}, w_{2,q}, \dots, w_{t,q}) \quad (1)$$

$$383 \quad \vec{d}_j = (w_{1,j}, w_{2,j}, \dots, w_{t,j}) \quad (2)$$

384 In the classical VSM,  $w_{i,j}$  is known as the Term Frequency-Inverse Document  
385 Frequency (TF-IDF) weight. If the weight vector model for a document  $d_j$  is  $\vec{d}_j$ , the  
386 document's TF-IDF weights can be quantified as:

$$387 \quad w_{i,j} = (1 + \log f_{i,j}) \times \log \left( \frac{N}{n_i} \right) \quad (3)$$

388 where  $f_{i,j}$  is the frequency of index term  $k_i$  in the document,  $N$  is the total  
389 number of documents in the document set, and  $n_i$  is the number of documents  
390 containing the term  $k_i$ .

391 Through using the VSM and TF-IDF model, the degree of similarity  $sim(d_j, q)$   
392 between the document  $d_j$  and the query  $q$  can be quantified as the cosine of the angle  
393 between the vectors  $\vec{d}_j$  and  $\vec{q}$ :

$$394 \quad sim(d_j, q) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^t w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^t w_{i,j}^2} \times \sqrt{\sum_{i=1}^t w_{i,q}^2}} \quad (4)$$

395 where  $|\vec{d}_j|$  and  $|\vec{q}|$  are the norms of the document and query vectors, and  $\vec{d}_j \cdot \vec{q}$   
396 is the inner product of the document and query vectors.

397 3.4.2 The proposed score strategy and computational process

398 A number of existing studies [43,59] have validated that query expansion could  
399 effectively improve the IR performance and a common method for query expansion is  
400 to use WordNet or other lexical databases. WordNet has pre-defined the basic semantic  
401 relationships between words, e.g. hypernym, synonym, hyponym. Gong et al. [42,60]  
402 pointed out these different semantic relations between words for query expansion will  
403 lead to different effects on the IR performance and an easy and effective approach to  
404 distinguish their effects is to give different weighting coefficients to the expanded terms.

405 After considering the effect of the expanded query  $q_e$ , this study takes the classical  
406 VSM as a starting point and proposes the following method to compute the similarity  
407 between the query and risk case:

408 
$$score = sim(d_j, q_o) + \lambda \times sim(d_j, q_e) \quad (5)$$

409 where  $\lambda$  is the coefficient for the effect of  $q_e$  and  $0 < \lambda < 1$ , and this study  
410 takes  $\lambda = 0.7$ .

411 The reasons are discussed as follows:

- 412 • The basic assumption of this study is that the original query and expanded query  
413 will cause different effects on the retrieval results. The original query by the  
414 user could mostly reflect a user's searching need for the risk case retrieval, and  
415 expanded terms using pre-defined risk-related lexicon or WordNet are more or  
416 less different with the original query in semantics. Therefore an optimal solution  
417 to distinguish the effects of the original query and the expanded query is to keep  
418 the original query and expanded query as separate operations (i.e. two queries  
419  $q_o$  and  $q_e$ ), and allocate different coefficients for them [42]. The expanded  
420 query  $q_e$  can be considered as an additional interpretation for the original

421 query  $q_o$ . If the coefficient for  $q_o$  is 1, then it is clear that the coefficient for  
 422  $q_e$  should be less than 1.

423 • As discussed in Section 3.3, this paper combines the use of a pre-defined risk-  
 424 related lexicon and synonyms in WordNet as the databases for query expansion.  
 425 The suggested expansion terms in the risk-related lexicon are “synonyms” of  
 426 the keyword in the project risk management context. Therefore, all expanded  
 427 terms can be considered similarly as “synonyms” of the original query. A  
 428 previous study by Gong et al. [42] tested the performance of a web IR system  
 429 using the different semantic relations between words of WordNet for query  
 430 expansion, and demonstrated that the optimal value of coefficient for synonyms  
 431 is 0.7. Hence this study takes  $\lambda$  as 0.7 for practical implementation.

432 The computational process is illustrated as follows. Assume there are totally  $k$  risk  
 433 case documents in the risk case database, a term-document weighting matrix can be  
 434 constructed as shown in Figure 8, where the two queries are extended as the last two  
 435 “documents”. For each risk case or document, the TF-IDF weights of all terms are  
 436 presented in a row. If a document contains no specific term, then this term’s weight in  
 437 the document is 0.

	$Doc_1$	$Doc_2$	...	$Doc_j$	...	$Doc_k$	$q_o$	$q_e$
$Term_1$	$W_{1,1}$	$W_{1,2}$	...	$W_{1,j}$	...	$W_{1,k}$	$W_{1,k+1}$	$W_{1,k+2}$
$Term_2$	$W_{2,1}$	$W_{2,1}$	...	$W_{2,j}$	...	$W_{2,k}$	$W_{2,k+1}$	$W_{2,k+2}$
...	...	...	...	...	...	...	...	...
$Term_i$	$W_{i,1}$	$W_{i,2}$	...	$W_{i,j}$	...	$W_{i,k}$	$W_{i,k+1}$	$W_{i,k+2}$
...	...	...	...	...	...	...	...	...
$Term_n$	$W_{n,1}$	$W_{n,2}$	...	$W_{n,j}$	...	$W_{n,k}$	$W_{n,k+1}$	$W_{n,k+2}$

438

439 Figure 8 Term-document weighting matrix

440 For any document  $d_j$ , the similarity between the query  $q$  and  $d_j$  can be computed as:

$$\begin{aligned} \text{score} &= \text{sim}(d_j, q_o) + 0.7 \times \text{sim}(d_j, q_e) \\ &= \frac{\sum_{i=1}^n w_{i,j} \times w_{i,k+1}}{\sqrt{\sum_{i=1}^n w_{i,j}^2} \times \sqrt{\sum_{i=1}^n w_{i,k+1}^2}} + 0.7 \times \frac{\sum_{i=1}^n w_{i,j} \times w_{i,k+2}}{\sqrt{\sum_{i=1}^n w_{i,j}^2} \times \sqrt{\sum_{i=1}^n w_{i,k+2}^2}} \end{aligned} \quad (6)$$

441 Due to the combination effects of  $q_o$  and  $q_e$ , the range of overall similarity is from 0  
442 to 1.7.

## 443 **4. System development and implementation**

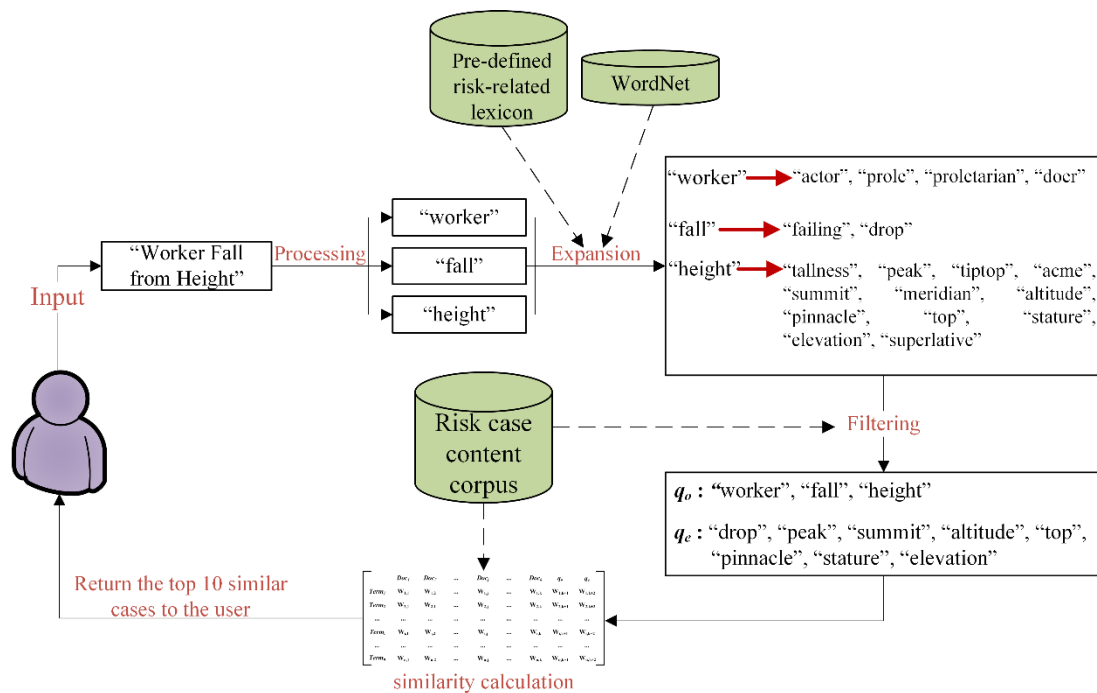
### 444 **4.1 Prototype development**

445 In order to fully implement the proposed RCRS, a prototype was developed using the  
446 Python programming language. Although other programming languages (e.g. R, Java)  
447 could have been used, this study chose Python because:

- 448 • Python is one of most widely used object-oriented programming languages with  
449 lots of features such as free and open source, easy syntax, and good extensibility.  
450 This means a Python program is easily read and understood by others and is  
451 highly extensible.
- 452 • A number of existing tools have been designed to support Python working with  
453 NLP, e.g. NLTK [47], data mining and analysis, e.g. scikit-learn [61]. Therefore  
454 developing the prototype using Python could build on valuable previous work  
455 and avoid repeated modelling work.

### 456 **4.2 Illustrative example**

457 The purpose of this sub-section is to use the example of “Worker Fall from Height” to  
458 illustrate the computational process of the developed prototype system. The overall  
459 computational process is presented in Figure 9.



460

461 Figure 9 Computational process of retrieving "Worker Fall from Height" similar cases

462 The overall computational process can be described as follows:

- 463
- 464
- 465
- 466
- 467
- 468
- 469
- 470
- 471
- 472
- 473
- 474
- 475
- 476
- Before starting risk case retrieval, the system needs to read and process all the risk cases and establish a corpus for further use. As discussed in Section 3.2, a total of 590 risk cases have been collected. The system starts with extracting textual content from each risk case and getting the name list of all risk cases. After reading each case, the system processes its textual content through SoA, and saves the processed case in a temporary file. Then, all temporary files are read according to the sequence of name list and stored in a list where each risk case is a string.
  - If a new query "Worker Fall from Height" is given by the user, the system first processes the query through SoA and obtains the tokens of original query, i.e. "worker", "fall" and "height". Then each token in the processed original query is prior scanned to find out its related words in the pre-defined lexicon. The terms not found in the pre-defined risk-related lexicon are expanded by using synonyms in WordNet. As only "fall" exists in the keyword list of pre-defined



477           lexicon, the pre-defined lexicon is used for expansion of “fall” and the  
478           synonyms of WordNet is used for expansion of “worker” and “height”. The  
479           related words for “fall” are “falling” and “drop”. The related words for “worker”  
480           are “actor”, “prole”, “proletarian” and “doer”. And the related words for “height”  
481           are “tallness”, “peak”, “tiptop”, “acme”, “summit”, “meridian”, “altitude”,  
482           “pinnacle”, “top”, “stature”, “elevation” and “superlative”. Thirdly, the system  
483           filters the original query and expanded query by scanning the risk case content  
484           corpus and deleting those terms that do not appear in the corpus. After filtering,  
485           the original query are “worker”, “fall” and “height” and the expanded terms are  
486           “drop”, “peak”, “summit”, “altitude”, “top”, “pinnacle”, “stature” and  
487           “elevation”.

488           • In the third step, the processed original query and expanded query are first  
489           extended to the corpus as the last two strings in the list. Then the system  
490           performs the calculation of TF-IDF weights and establishes the corresponding  
491           term-document matrix (shown in Figure 8). Finally, the similarity between the  
492           query and each risk case is computed by using Equation (6) and the system  
493           returns the ranked top 10 similar risk cases to the end users. The result is shown  
494           in Table 2.

495 Table 2 Top 10 similar cases of “Worker Fall from Height”

Similarity	Title of risk case	Source	Number
0.355807864882	Young worker falls from third-storey balcony	WorkSafeBC	30
0.350710609398	Fall from roof with too much slack in lifeline	WorkSafeBC	3
0.306337588766	Hispanic laborer dies after falling through a second story floor opening	NIOSH	5
0.286606375085	Worker falls through roof insulation to concrete floor	WorkSafeBC	27
0.282279911804	Worker died after fall from steep-sloped roof	WorkSafeBC	12
0.281084486537	Worker entangled in chain falling from dismantled conveyor	WorkSafeBC	13
0.278102714551	Worker died after being submerged in flooded cranberry field	WorkSafeBC	11
0.277708195414	Workers seriously burned in flash fire	WorkSafeBC	20
0.238392609973	Hispanic worker falls from residential roof	NIOSH	1
0.235168098338	Workers fall when unsecured bin tips off elevated forks	WorkSafeBC	19

496 **4.3 System testing**

497 Although there are a number of matrices that have been proposed to evaluate and test  
 498 IR systems, the most widely used are Precision, Recall and F score [14,16,32] which  
 499 can be calculated with the help of a simplified confusion matrix [32,62] shown in Table  
 500 3. There are four variables in the simplified confusion matrix, i.e. True Positive (TP),  
 501 False Positive (FP), False Negative (FN), and True Negative (TN). Here the terms  
 502 “positive” and “negative” mean the expectation of a retrieval while the terms “true” and  
 503 “false” refer to whether that expectation corresponds to the external judgment. In other  
 504 words, TP means the number of relevant documents retrieved, FP means the number of  
 505 irrelevant documents retrieved, FN means the number of relevant documents not  
 506 retrieved, and TN means the number of irrelevant documents not retrieved.

507 Table 3 Confusion matrix

	Relevant	Not relevant
Retrieved	True Positive (TP)	False Positive (FP)
Not retrieved	False Negative (FN)	True Negative (TN)

508 Precision refers to the fraction of retrieved documents that is relevant and is used to  
509 measure the percentage of relevant documents in all retrieved documents, i.e.

$$510 \quad \textit{Precision} = \frac{TP}{TP+FP} \times 100\% \quad (7)$$

511 Recall is defined as the fraction of relevant documents that has been retrieved and used  
512 for measuring the percentage of retrieved documents in all relevant documents, i.e.

$$513 \quad \textit{Recall} = \frac{TP}{TP+FN} \times 100\% \quad (8)$$

514 Another measure called  $F$  is the harmonic mean of Precision and Recall and is defined  
515 as follows:

$$516 \quad F = \frac{\textit{Precision} \times \textit{Recall}}{\textit{Precision} + \textit{Recall}} \times 100\% \quad (9)$$

517 It is noticed that Precision, Recall, and F value are commonly used for evaluating the  
518 whole retrieval system and it requires an accurate boundary between “retrieved” and  
519 “not retrieved” to calculate the three measures. Here determining the threshold (or cut-  
520 off) is extremely important and its value could in large degree affect the evaluation  
521 results of an IR system. However, there is a need to point out that determining the  
522 threshold value in an IR system is complex and needs a large number of experiments,  
523 which is not within the scope of this study. Unlike web-scale IR, the information in the  
524 construction industry is relatively small-scale and domain-specific and a common  
525 method to evaluate the performance of an IR system for construction projects is through  
526 testing a number of samples and setting user experience based threshold value, e.g.  
527 [16,49]. Besides, with the observation that in the real working environment engineers  
528 often expect to obtain the needed information within a limited amount of time [63] and  
529 the top 10-20 cases would by nature have the most value to the end users [49], the

530 proposed RCRS is designed to return the top 10 most similar cases. Hence, this study  
531 also evaluated the percentage of relevant risk cases among the top 10 similar documents,  
532 which is defined as Precision at 10 (P@10):

$$533 \quad P@10 = \frac{\text{number of relevant documents in top 10}}{10} \times 100\% \quad (10)$$

534 In order to test and evaluate the proposed RCRS, this study took the threshold value as  
535 0.1 from preliminary system use experience and the testing procedure consists of the  
536 following steps:

- 537 • Firstly, a set of key terms (e.g. “bridge”, “fall”, “collapse”, “construction”) that  
538 are relevant to the scope of collected risk cases were selected for making up 10  
539 testing queries. The queries were divided into 3 groups, i.e. “type of risk”,  
540 “object + type of risk”, and “object + type of risk + project phase”, to simulate  
541 the real situations of case retrieval. The “type of risk” group contains three  
542 queries, i.e. “fall from height”, “flood risk”, “design error”. The “object + type  
543 of risk” group consists of 5 queries, i.e. “flood risk of bridge”, “worker fall from  
544 height”, “tower crane collapse”, “bridge failure”, “worker injury”. The “object  
545 + type of risk + project phase” group contains two queries, i.e. “worker die in  
546 construction” and “structure collapse in demolition”;
- 547 • Secondly, each testing query was inputted into the RCRS for query-document  
548 matching and the corresponding output was recorded in an Excel table. As this  
549 paper took an experience-based threshold (or cut-off) value 0.1, those  
550 documents with the similarity score over 0.1 were classified into the “retrieved”  
551 group while those documents with the similarity score which is less than 0.1  
552 were classified to the “not retrieved” group;
- 553 • Thirdly, because the similarity value for those documents containing no terms  
554 of original and expanded queries is 0, then those documents were determined to  
555 be irrelevant directly. Then the results were carefully reviewed to determine if

556 a risk case is relevant to the query by quickly reading and understanding each  
 557 document and analysing the relationship between the query and the document.  
 558 If a document is determined to be relevant to the query, the value “1” was  
 559 labelled for that document in Excel. Otherwise, the value “0” was given. Then,  
 560 TP, FP, FN, TN and P@10 were calculated.  
 561 • In the last step, the calculation of Precision, Recall, and F value for each testing  
 562 retrieval was performed and the testing results are shown in Table 4.

563 Table 4 Testing results

No.	Testing query	Number of retrievals				Performance			
		TP	FP	FN	TN	Precision	Recall	F	P@10
1	fall from height	18	1	18	553	94.7%	50.0%	65.5%	90%
2	flood risk	11	5	0	574	68.8%	100.0%	81.5%	100%
3	design error	22	4	6	558	84.6%	78.6%	81.5%	100%
4	flood risk of bridge	11	30	0	549	26.8%	100.0%	42.3%	100%
5	worker fall from height	25	10	2	553	71.4%	92.6%	80.6%	90%
6	tower crane collapse	18	23	0	549	43.9%	100.0%	61.0%	70%
7	bridge failure	42	16	3	529	72.4%	93.3%	81.6%	100%
8	worker injury	32	3	18	537	91.4%	64.0%	75.3%	100%
9	worker die in construction	30	1	11	548	96.8%	73.2%	83.3%	100%
10	structure collapse in demolition	16	34	0	540	32.0%	100.0%	48.5%	100%

564 The search results show that generally the proposed RCRS is capable of retrieving  
 565 relevant risk cases from the database for a specified query. In particular, the results of  
 566 P@10 are excellent, mostly 100% (7 of 10). Only one testing query had 70% of P@10,  
 567 which also is a satisfactory result. Therefore the top 10 cases returned by the system are  
 568 valuable to the user. The high percentage of P@10 can be explained by the term  
 569 frequency being an important factor in computing the TF-IDF weights and a document  
 570 containing as many query terms as possible is easier to obtain a high similarity score.  
 571 Although the Precision score for several queries were relatively low, this does not mean  
 572 the retrieval results were not good. For example, for the “flood risk of bridge” query,  
 573 41 results were retrieved and only 11 were determined to be similar to the query. Two  
 574 reasons could explain this problem: first, there are a very small number of “flood”

575 related samples in the risk case database; second, because the threshold value 0.1 in this  
576 case is too small and the expanded terms were producing some “noise”. But from its  
577 P@10 score, it can be seen that the top 10 were all similar to the query and nearly all  
578 valuable documents were ranked. Therefore simply increasing the threshold value for  
579 some queries could improve the search results. In addition, some researchers [14,16]  
580 also claim that there are still some technical limitations in the current NLP, which lead  
581 to the conclusion that the search results cannot be perfect. For example, the “flood risk”  
582 here is an entity but the system failed to read it as an entity and split it into two separate  
583 terms “flood” and “risk” for consideration.

## 584 **5. Discussions**

585 The literature shows that CBR is a process of learning from the past, which could  
586 facilitate previous knowledge and experience to be effectively used for risk  
587 management in new projects. In the CBR cycle, RETRIEVE is the first and the most  
588 important step [7,15]. A commonly used traditional way for assessing the similarity  
589 between user need and risk cases is through attaching attribute labels to each risk case  
590 document and allocating different weights to those attributes [9,22,25]. However, as  
591 discussed in Section 2.1, some challenges still exist: (1) traditional methods are very  
592 limited in scope, (2) a large amount of pre-processing or preparation work is needed,  
593 and (3) very few studies have been found to be capable of addressing the challenge of  
594 semantic similarity. In order to overcome the current challenges of case retrieval in  
595 CBR, this paper analysed the potential and benefits of integrating NLP into risk case  
596 retrieval. The idea was motivated by recent research that has introduced NLP into  
597 textual information management into construction industry, e.g. retrieval of CAD  
598 drawings [16], retrieval of relevant information for assisting decision making [64,65],  
599 injury report content analysis [14], and document clustering [17]. It can be seen that the  
600 application of NLP into textual documents analysis and management in the construction

601 industry is a new and promising trend. Some recent studies even extended the use of  
602 NLP into Building Information Modelling (BIM), an emerging digital technology in  
603 the construction industry, for automated code checking [66], processing building  
604 information [67], retrieving online BIM resources [50], etc.

605 A number of recent studies [16,49] successfully used the classical VSM for IR and  
606 document management, and discussed that the semantic similarity is still a huge  
607 challenge in any current application of NLP in the construction industry. To partially  
608 overcome this gap, this paper outlines a framework of combining the use of semantic  
609 query expansion and VSM for retrieval of similar risk cases, and develops a system  
610 prototype with Python to support the proposed approach. The test results show the  
611 proposed system could quickly and effectively retrieve and rank valuable risk cases  
612 when a query is specified. Through implementing the proposed system, end users could  
613 quickly find out risk cases that are valuable references to the new situations or problems  
614 and embed the knowledge and experience of previous accidents into daily work. Any  
615 new cases could be added into the risk case database flexibly for retrieval without pre-  
616 processing work. In addition, because this system prototype is written with Python, the  
617 RCRS could also be easily integrated into software written by other programming  
618 languages. As an example of its practical contributions, the proposed approach can be  
619 embedded into some online risk case databases, e.g. Structural-Safety and NIOSH, as  
620 a semantic searching engine. In the future, the proposed approach can be also expanded  
621 for the wider management of engineering documents and information.

622 Of course, some limitations also exist in this study. These limitations and the  
623 corresponding recommendations for future research are discussed as follows:

- 624 • First, the proposed system is limited in case retrieval within the internal risk  
625 case database and the total number of collected risk cases is still relatively small.  
626 As described in Section 3.2, due to the limited time only 590 risk cases covering

627           7 types of risk were collected. The reasons are: 1) the main purpose of this study  
628           is developing a general approach (i.e. proof of concept) based on NLP for risk  
629           case retrieval instead of establishing a complete risk case database; and 2) there  
630           are relatively few detailed reports on those risks that are not so dangerous or  
631           fatal, e.g. financial loss, time overrun. However, the limited size of the database  
632           will influence the retrieval results and practical applicability. For example, if a  
633           user query is “time overrun” and the database contains no risk cases about “time  
634           overrun”, it will be difficult for the system to return the desired results to the  
635           user. Therefore, future research may consider: 1) how to enrich the risk case  
636           database; 2) how to formulate case retrieval guidelines to the end user according  
637           to the distribution of risk cases; and 3) how to extend the proposed system for  
638           risk case retrieval in external databases and online resources.

639           • Secondly, the semantic similarity problem is still a huge challenge within the  
640           state-of-the-art research of NLP [31], and the query expansion approach  
641           adopted by this study can only address a limited proportion of the problem. In  
642           particular, the proposed system combines the use of a pre-defined risk-related  
643           lexicon and WordNet to deal with the word mismatching problem of case  
644           retrieval. However, the pre-defined lexicon only contains explanations of 107  
645           key terms in the project risk management domain and is not a complete  
646           dictionary. To overcome the shortcoming of the pre-defined lexicon, WordNet  
647           is used as an important supplementary. However, because WordNet is a large  
648           lexical database for the English language and is not specially designed for risk  
649           management, this study found some terms expanded by WordNet are not related  
650           to project risks and have little, or no value in risk case retrieval. Moreover, it  
651           can be seen that human language is still extremely complex and difficult for  
652           computers to understand and process. For example, Caldas and Han [68] made  
653           use of IR and text mining for automatic classification of project documents but



654 found the results were not perfect due to the multiple meanings of words. In  
655 addition, as discussed in Section 4.3, though the pre-defined lexicon and  
656 WordNet can be used for explanation of a single term, it is still difficult for  
657 computer to process the word groups. Hence, one short-term recommendation  
658 for future research may be to establish a comprehensive lexicon for project risk  
659 management which includes the definition of the linked relationships of  
660 common word groups. From a long-term perspective, future research may apply  
661 the state-of-the-art techniques of NLP into addressing the semantic similarity  
662 problem in both risk case retrieval and other fields.

663 • Thirdly, the proposed system has not been put into use and validated in practice.  
664 For better implementation of the proposed approach, the prototype system needs  
665 to be further developed as a tool with easy-to-use user interface and checked by  
666 different scenarios. In addition, as the proposed system was designed to return  
667 the most similar 10 risk cases to the user and the test results presented in  
668 Sections 4.2 and 4.3 are satisfactory, when conducting the preliminary testing  
669 this paper checked the results manually and did not study the best value of the  
670 threshold. Although a number of matrices (e.g. Precision, Recall, F and P@10)  
671 could be used for evaluating an IR system, nearly all of them require a clear  
672 boundary of “retrieved” and “not retrieved”, and “relevant” and “not relevance”.  
673 The threshold value is often used to divide the returned results into “retrieved”  
674 and “not retrieved”; however, Qady and Kandil [17] pointed out the best  
675 threshold value normally lies between 0.05 and 0.95, and determining the best  
676 value needs a large number of experiments. Furthermore, the relevance is by  
677 nature often continuous instead of binary, which leads to the difficulty of  
678 determining if a retrieved document is relevant or not [69,70]. Hence, future  
679 research may further study the threshold value and relevance problem, and test  
680 and improve the proposed approach and system in real practice.

## 681 **6. Conclusions**

682 This paper introduced an approach of combining the use of two NLP techniques (i.e.  
683 VSM and semantic query expansion) for risk case retrieval and proposed a framework  
684 for the risk case retrieval system. The VSM could represent textual documents as  
685 vectors of identifiers and assigning TF-IDF weights to index terms in both queries and  
686 documents, which could be used to compute the degree of similarity between  
687 documents and the query, while the query expansion could solve the mismatching  
688 problem of terms that have the same semantic meanings through expanding the original  
689 query using related terms defined in a pre-defined risk-related lexicon and synonyms  
690 in WordNet. A prototype system was developed using Python to implement the  
691 proposed approach.

692 Through implementing the proposed system, textual content information is firstly  
693 extracted from the risk case dataset and processed to generate a content corpus. After a  
694 query is inputted by the user, then the system starts to read and process the query,  
695 combines the use of a pre-defined risk-related lexicon or WordNet to expand the  
696 original query, and filters out the query terms that do not exist in the content corpus.  
697 Lastly the system gathers original query, expanded query and content corpus together  
698 for query-document similarity computing and returns the top 10 similar risk cases to  
699 the user. The preliminary test results have demonstrated the system's capacity of  
700 automatically retrieving similar risk cases.

701 Although there are still some limitations of applying current NLP technology into  
702 engineering textual information management, using such a system for managing risk  
703 cases could effectively facilitate the risk identification and communication, and  
704 information management. The suggested future research may include, for example: 1)  
705 to enrich the risk case database and expand the capacity of the proposed system for  
706 accessing both internal database and online risk case resources; 2) to investigate how

707 state-of-the-art NLP can be further developed to address the semantic similarity  
708 problems (e.g. processing word groups); 3) to improve the evaluation methods for  
709 retrieval of small-scale data; and 4) to test and optimise the proposed approach and  
710 system in practice.

711

## 712 **Acknowledgements**

713 This research is jointly funded by the University of Liverpool and China Scholarship  
714 Council (Grant Number: 201408500090).

715 **References**

- 716 [1] R. Sacks, O. Rozenfeld, Y. Rosenfeld, Spatial and temporal exposure to safety  
717 hazards in construction, *Journal of Construction Engineering and Management*.  
718 135 (8) (2009) 726-736, [http://dx.doi.org/10.1061/\(ASCE\)0733-  
719 9364\(2009\)135:8\(726\)](http://dx.doi.org/10.1061/(ASCE)0733-9364(2009)135:8(726)).
- 720 [2] K. Wardhana, F.C. Hadipriono, Analysis of recent bridge failures in the United  
721 States, *Journal of performance of constructed facilities*. 17 (3) (2003) 144-150,  
722 [http://dx.doi.org/10.1061/\(ASCE\)0887-3828\(2003\)17:3\(144\)](http://dx.doi.org/10.1061/(ASCE)0887-3828(2003)17:3(144)).
- 723 [3] S.J. Zhang, J. Teizer, J.K. Lee, C.M. Eastman, M. Venugopal, Building  
724 Information Modeling (BIM) and Safety: Automatic Safety Checking of  
725 Construction Models and Schedules, *Automation in Construction*. 29 (2013)  
726 183-195, <http://dx.doi.org/10.1016/j.autcon.2012.05.006>.
- 727 [4] ILO, Fact Sheet on Safety at Work, International Labour Organization (ILO),  
728 Geneva, Switzerland, 2005.
- 729 [5] Y. Zou, A. Kiviniemi, S.W. Jones, A review of risk management through BIM  
730 and BIM-related technologies, *Safety Science* (In press). (2016),  
731 <http://dx.doi.org/10.1016/j.ssci.2015.12.027>.
- 732 [6] I. Dikmen, M. Birgonul, C. Anac, J. Tah, G. Aouad, Learning from risks: A tool  
733 for post-project risk assessment, *Automation in Construction*. 18 (1) (2008) 42-  
734 50, <http://dx.doi.org/10.1016/j.autcon.2008.04.008>.
- 735 [7] Y.M. Goh, D. Chua, Case-based reasoning for construction hazard  
736 identification: case representation and retrieval, *Journal of Construction  
737 Engineering and Management*. 135 (11) (2009) 1181-1189,  
738 [http://dx.doi.org/10.1061/\(ASCE\)CO.1943-7862.0000093](http://dx.doi.org/10.1061/(ASCE)CO.1943-7862.0000093).
- 739 [8] D.H. Jonassen, J. Hernandez-Serrano, Case-based reasoning and instructional  
740 design: Using stories to support problem solving, *Educational Technology  
741 Research and Development*. 50 (2) (2002) 65-77.
- 742 [9] J. Kolodner, *Case-Based Reasoning*, Morgan Kaufmann, San Mateo, CA, 1993.
- 743 [10] X. Zhang, Y. Deng, Q. Li, M. Skitmore, Z. Zhou, An incident database for  
744 improving metro safety: The case of shanghai, *Safety Science*. 84 (2016) 88-96,  
745 <http://dx.doi.org/10.1016/j.ssci.2015.11.023>.
- 746 [11] Structural Safety database, 2016, available at: <http://www.structural-safety.org/>  
747 [accessed on 20 March 2016]
- 748 [12] National Institute for Occupational Safety and Health (NIOSH) database, 2016,  
749 available at: <http://www.cdc.gov/niosh/> [accessed on 25 March 2016]
- 750 [13] B. Esmaeili, M. Hallowell, Attribute-based risk model for measuring safety risk  
751 of struck-by accidents, in: H. Cai, A. Kandil, M. Hastak, P.S. Dunston (Eds.),  
752 *Construction Research Congress*, American Society of Civil Engineers, West  
753 Lafayette, Indiana, 2012, pp. 289-298.
- 754 [14] A.J.-P. Tixier, M.R. Hallowell, B. Rajagopalan, D. Bowman, Automated  
755 content analysis for construction safety: A natural language processing system

- 756 to extract precursors and outcomes from unstructured injury reports,  
757 Automation in Construction. 62 (2016) 45-56,  
758 <http://dx.doi.org/10.1016/j.autcon.2015.11.001>.
- 759 [15] R.L. De Mantaras, D. McSherry, D. Bridge, D. Leake, B. Smyth, S. Craw, B.  
760 Faltings, M.L. Maher, M. T COX, K. Forbus, Retrieval, reuse, revision and  
761 retention in case-based reasoning, The Knowledge Engineering Review. 20 (03)  
762 (2005) 215-240, <http://dx.doi.org/10.1017/S0269888906000646>.
- 763 [16] J.-Y. Hsu, Content-based text mining technique for retrieval of CAD documents,  
764 Automation in Construction. 31 (2013) 65-74,  
765 <http://dx.doi.org/10.1016/j.autcon.2012.11.037>.
- 766 [17] M. Al Qady, A. Kandil, Automatic clustering of construction project documents  
767 based on textual similarity, Automation in Construction. 42 (2014) 36-49,  
768 <http://dx.doi.org/10.1016/j.autcon.2014.02.006>.
- 769 [18] R.C. Schank, Dynamic memory: A theory of reminding and learning in  
770 computers and people, Cambridge University Press, New York, 1983.
- 771 [19] R.C. Schank, A. Kass, C.K. Riesbeck, Inside case-based explanation,  
772 Psychology Press, New York, 2014.
- 773 [20] D. Forbes, S. Smith, M. Horner, Tools for selecting appropriate risk  
774 management techniques in the built environment, Construction Management  
775 and Economics. 26 (11) (2008) 1241-1250,  
776 <http://dx.doi.org/10.1080/01446190802468487>.
- 777 [21] Y.M. Goh, D. Chua, Case-based reasoning approach to construction safety  
778 hazard identification: adaptation and utilization, Journal of Construction  
779 Engineering and Management. 136 (2) (2009) 170-178,  
780 [http://dx.doi.org/10.1061/\(ASCE\)CO.1943-7862.0000116](http://dx.doi.org/10.1061/(ASCE)CO.1943-7862.0000116).
- 781 [22] Y. Lu, Q. Li, W. Xiao, Case-based reasoning for automated safety risk analysis  
782 on subway operation: Case representation and retrieval, Safety Science. 57  
783 (2013) 75-81, <http://dx.doi.org/10.1016/j.ssci.2013.01.020>.
- 784 [23] V. Kumar, N. Viswanadham, A CBR-based decision support system framework  
785 for construction supply chain risk management, Proceedings of 2007 IEEE  
786 International Conference on Automation Science and Engineering, IEEE,  
787 Scottsdale, AZ, 2007, pp. 980-985.
- 788 [24] A. Aamodt, E. Plaza, Case-based reasoning: Foundational issues,  
789 methodological variations, and system approaches, Artificial Intelligence  
790 Communications. 7 (1) (1994) 39-59, <http://dx.doi.org/10.3233/AIC-1994-7104>.
- 791 [25] A. Karim, H. Adeli, CBR Model for Freeway Work Zone Traffic Management,  
792 Journal of Transportation Engineering. 129 (2) (2003) 134-145,  
793 [http://dx.doi.org/10.1061/\(ASCE\)0733-947X\(2003\)129:2\(134\)](http://dx.doi.org/10.1061/(ASCE)0733-947X(2003)129:2(134)).
- 794 [26] P. Cunningham, A taxonomy of similarity mechanisms for case-based  
795 reasoning, IEEE Transactions on Knowledge and Data Engineering. 21 (11)  
796 (2009) 1532-1543, <http://dx.doi.org/10.1109/TKDE.2008.227>.
- 797

- 798 [27] J. Zhao, L. Cui, L. Zhao, T. Qiu, B. Chen, Learning HAZOP expert system by  
799 case-based reasoning and ontology, *Computers & Chemical Engineering*. 33 (1)  
800 (2009) 371-378, <http://dx.doi.org/10.1016/j.compchemeng.2008.10.006>.
- 801 [28] S. Harispe, S. Ranwez, S. Janaqi, J. Montmain, Semantic similarity from natural  
802 language and ontology analysis, *Synthesis Lectures on Human Language*  
803 *Technologies*. 8 (1) (2015) 1-254,  
804 <http://dx.doi.org/10.2200/S00639ED1V01Y201504HLT027>.
- 805 [29] G.G. Chowdhury, Natural language processing, *Annual review of information*  
806 *science and technology*. 37 (1) (2003) 51-89,  
807 <http://dx.doi.org/10.1002/aris.1440370103>.
- 808 [30] Y. Bar-Hillel, The present status of automatic translation of languages,  
809 *Advances in computers*. 1 (1960) 91-163, [http://dx.doi.org/10.1016/S0065-](http://dx.doi.org/10.1016/S0065-2458(08)60607-5)  
810 [2458\(08\)60607-5](http://dx.doi.org/10.1016/S0065-2458(08)60607-5).
- 811 [31] D. Jurafsky, J.H. Martin, *Speech and language processing : an introduction to*  
812 *natural language processing, computational linguistics, and speech recognition*,  
813 *second ed.*, Prentice Hall, New Jersey, 2009.
- 814 [32] R. Baeza-Yates, B. Ribeiro-Neto, *Modern information retrieval : the concepts*  
815 *and technology behind search*, second ed., Addison Wesley, Harlow, UK, 2011.
- 816 [33] X. Bai, Predicting consumer sentiments from online text, *Decision Support*  
817 *Systems*. 50 (4) (2011) 732-742, <http://dx.doi.org/10.1016/j.dss.2010.08.024>.
- 818 [34] M.N. Murty, A.K. Jain, Knowledge-based clustering scheme for collection  
819 management and retrieval of library books, *Pattern recognition*. 28 (7) (1995)  
820 949-963, [http://dx.doi.org/10.1016/0031-3203\(94\)00173-J](http://dx.doi.org/10.1016/0031-3203(94)00173-J).
- 821 [35] L. Soibelman, J. Wu, C. Caldas, I. Brilakis, K.-Y. Lin, Management and  
822 analysis of unstructured construction data types, *Advanced Engineering*  
823 *Informatics*. 22 (1) (2008) 15-27, <http://dx.doi.org/10.1016/j.aei.2007.08.011>.
- 824 [36] D. Kaminetzky, *Design and construction failures: Lessons from forensic*  
825 *investigations*, McGraw-Hill, New York, 2001.
- 826 [37] C.H. Caldas, L. Soibelman, Automating hierarchical document classification  
827 for construction management information systems, *Automation in Construction*.  
828 12 (4) (2003) 395-406, [http://dx.doi.org/10.1016/S0926-5805\(03\)00004-9](http://dx.doi.org/10.1016/S0926-5805(03)00004-9).
- 829 [38] N.-W. Chi, K.-Y. Lin, S.-H. Hsieh, Using ontology-based text classification to  
830 assist Job Hazard Analysis, *Advanced Engineering Informatics*. 28 (4) (2014)  
831 381-394, <http://dx.doi.org/10.1016/j.aei.2014.05.001>.
- 832 [39] F.C. Pereira, F. Rodrigues, M. Ben-Akiva, Text analysis in incident duration  
833 prediction, *Transportation Research Part C: Emerging Technologies*. 37 (2013)  
834 177-192, <http://dx.doi.org/10.1016/j.trc.2013.10.002>.
- 835 [40] M.K. Khribi, M. Jemni, O. Nasraoui, Automatic recommendations for e-  
836 learning personalization based on web usage mining techniques and information  
837 retrieval, *Proceedings of 2008 Eighth IEEE International Conference on*  
838 *Advanced Learning Technologies*, IEEE, Santander, Cantabria, 2008, pp. 241-  
839 245.

- 840 [41] C. Fellbaum, WordNet: an electronic lexical database, MIT Press, Cambridge,  
841 Mass, 1998.
- 842 [42] Z. Gong, C.W. Cheang, U.L. Hou, Web query expansion by WordNet,  
843 Proceedings of 16th International Conference on Database and Expert Systems  
844 Applications, Springer, Copenhagen, Denmark, 2005, pp. 166-175.
- 845 [43] V. Snasel, P. Moravec, J. Pokorny, WordNet ontology based model for web  
846 retrieval, Proceedings of International Workshop on Challenges in Web  
847 Information Retrieval and Integration, IEEE, Tokyo, Japan, 2005, pp. 220-225.
- 848 [44] WorkSafeBC database, 2016, available at: [www.worksafebc.com/en](http://www.worksafebc.com/en) [accessed  
849 on 28 March 2016]
- 850 [45] Occupational Safety and Health Administration (OSHA) database, 2016,  
851 available at: [www.osha.gov](http://www.osha.gov) [accessed on 1 April 2016]
- 852 [46] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval,  
853 Cambridge University Press, New York, 2008.
- 854 [47] Natural Language Toolkit, 2016, available at: [www.nltk.org](http://www.nltk.org) [accessed on 5  
855 April 2016]
- 856 [48] J. Perkins, Python 3 Text Processing with NLTK 3 Cookbook, second ed., Packt  
857 Publishing Ltd, Birmingham, UK, 2014.
- 858 [49] H. Fan, H. Li, Retrieving similar cases for alternative dispute resolution in  
859 construction accidents using text mining techniques, Automation in  
860 Construction. 34 (2013) 85-91, <http://dx.doi.org/10.1016/j.autcon.2012.10.014>.
- 861 [50] G. Gao, Y.-S. Liu, M. Wang, M. Gu, J.-H. Yong, A query expansion method  
862 for retrieving online BIM resources based on Industry Foundation Classes,  
863 Automation in Construction. 56 (2015) 14-25,  
864 <http://dx.doi.org/10.1016/j.autcon.2015.04.006>.
- 865 [51] F. Colace, M. De Santo, L. Greco, P. Napoletano, Weighted word pairs for  
866 query expansion, Information Processing & Management. 51 (1) (2015) 179-  
867 193, <http://dx.doi.org/10.1016/j.ipm.2014.07.004>.
- 868 [52] O. Vechtomova, Y. Wang, A study of the effect of term proximity on query  
869 expansion, Journal of Information Science. 32 (4) (2006) 324-333,  
870 <http://dx.doi.org/10.1177/0165551506065787>.
- 871 [53] Y. Zou, A. Kiviniemi, S.W. Jones, Developing a Tailored RBS Linking to BIM  
872 for Risk Management of Bridge Projects, Engineering, Construction and  
873 Architectural Management. 23 (6) (2016) 727-750,  
874 <http://dx.doi.org/10.1108/ECAM-01-2016-0009>.
- 875 [54] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word  
876 representations in vector space, 2013, available at:  
877 <http://arxiv.org/abs/1301.3781>.
- 878 [55] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed  
879 representations of words and phrases and their compositionality, 2013, available  
880 at: <https://arxiv.org/abs/1310.4546>.



- 881 [56] Wikipedia content database (English), 2016, available at:  
882 [meta.wikimedia.org/wiki/Data\\_dumps](http://meta.wikimedia.org/wiki/Data_dumps) [accessed on 10 April 2016]
- 883 [57] G. Salton, A. Wong, C.-S. Yang, A vector space model for automatic indexing,  
884 Communications of the ACM. 18 (11) (1975) 613-620,  
885 <http://dx.doi.org/10.1145/361219.361220>.
- 886 [58] K. Sparck Jones, A statistical interpretation of term specificity and its  
887 application in retrieval, Journal of documentation. 28 (1) (1972) 11-21,  
888 <http://dx.doi.org/10.1108/eb026526>.
- 889 [59] T. De Simone, D. Kazakov, Using wordnet similarity and antonymy relations  
890 to aid document retrieval, Proceedings of Recent Advances in Natural  
891 Language Processing (RANLP), Borovets, Bulgaria, 2005.
- 892 [60] Z. Gong, C.W. Cheang, An implementation of web image search engines,  
893 Proceedings of International Conference on Asian Digital Libraries, Springer,  
894 Shanghai, China, 2004, pp. 355-367.
- 895 [61] Scikit-learn toolkit, 2016, available at: [www.scikit-learn.org/stable/](http://www.scikit-learn.org/stable/) [accessed  
896 on 15 April 2016]
- 897 [62] D.L. Olson, D. Delen, Advanced data mining techniques, Springer-Verlag  
898 Berlin Heidelberg, Germany, 2008.
- 899 [63] A.S. Kazi, Knowledge management in the construction industry: A socio-  
900 technical perspective, IGI Global, Hershey, Pennsylvania, 2005.
- 901 [64] X. Lv, N.M. El-Gohary, Enhanced context-based document relevance  
902 assessment and ranking for improved information retrieval to support  
903 environmental decision making, Advanced Engineering Informatics. 30 (4)  
904 (2016) 737-750, <http://dx.doi.org/10.1016/j.aei.2016.08.004>.
- 905 [65] X. Lv, N.M. El-Gohary, Semantic annotation for supporting context-aware  
906 information retrieval in the transportation project environmental review domain,  
907 Journal of Computing in Civil Engineering. 30 (6) (2016) 04016033,  
908 [http://dx.doi.org/10.1061/\(ASCE\)CP.1943-5487.0000565](http://dx.doi.org/10.1061/(ASCE)CP.1943-5487.0000565).
- 909 [66] J. Zhang, N.M. El-Gohary, Integrating semantic NLP and logic reasoning into  
910 a unified system for fully-automated code checking, Automation in  
911 Construction. 73 (2017) 45-57, <http://dx.doi.org/10.1016/j.autcon.2016.08.027>.
- 912 [67] J. Beetz, J. Van Leeuwen, B. De Vries, IfcOWL: A case of transforming  
913 EXPRESS schemas into ontologies, Artificial Intelligence for Engineering  
914 Design, Analysis and Manufacturing. 23 (01) (2009) 89-101,  
915 <http://dx.doi.org/10.1017/S0890060409000122>.
- 916 [68] C.H. Caldas, L. Soibelman, J. Han, Automated classification of construction  
917 project documents, Journal of Computing in Civil Engineering. 16 (4) (2002)  
918 234-243, [http://dx.doi.org/10.1061/\(ASCE\)0887-3801\(2002\)16:4\(234\)](http://dx.doi.org/10.1061/(ASCE)0887-3801(2002)16:4(234)).
- 919 [69] J. Kekäläinen, Binary and graded relevance in IR evaluations—comparison of  
920 the effects on ranking of IR systems, Information Processing & Management.  
921 41 (5) (2005) 1019-1033, <http://dx.doi.org/10.1016/j.ipm.2005.01.004>.



922 [70] J.W. Janes, The binary nature of continuous relevance judgments: A study of  
923 users' perceptions, *Journal of the American Society for Information Science*. 42  
924 (10) (1991) 754-756, [http://dx.doi.org/10.1002/\(SICI\)1097-](http://dx.doi.org/10.1002/(SICI)1097-4571(199112)42:10<754::AID-ASI9>3.0.CO;2-C)  
925 [4571\(199112\)42:10<754::AID-ASI9>3.0.CO;2-C](http://dx.doi.org/10.1002/(SICI)1097-4571(199112)42:10<754::AID-ASI9>3.0.CO;2-C).

926