Automation in Construction 111 (2020) 103048

Contents lists available at ScienceDirect

Automation in Construction

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

Text visualization for construction document information management

Jun Sun[a], Kun Lei[a], Lei Cao[b], Botao Zhong[a,*], Yi Wei[a], Jintao Li[c], Zhiling Yang[a]

[a] School of Civil Engineering & Mechanics, Huazhong University of Science & Technology, Wuhan 430074, China
[b] China Railway Siyuan Survey And Design Group Co., LTD., Wuhan 430074, China
[c] School of Civil Engineering, Architecture and Environment, Hubei University of Technology, Wuhan 430074, China

ARTICLE INFO

Keywords:
Construction document management
Text visualization
Information extraction
Text mining

ABSTRACT

In this study, text mining and visualization technology is applied to extract valuable information otherwise buried in the dense and abstract form of construction report text and process it into simple and intuitive graphics. This framework allows managers to quickly understand and make more informed decisions based upon key information. To extract such key information from a text automatically, the Term Frequency–Inverse Document Frequency algorithm was optimized to the characteristics of engineering texts. The key messages extracted from a case study construction report text in the form of keywords/terms were then visualized using a tag cloud algorithm. Questionnaires completed by construction managers then demonstrated that the proposed information visualization framework can facilitate rapid and informative access to key project information. This framework can thus reduce the workload and time required for construction managers to ascertain and act upon the status of their projects.

1. Introduction

Construction is an extremely information dependent industry in which a project's success largely depends on good access to and management of data. The overall informatization rate of China's construction industry is only 0.027%, far less than the international average of 0.3% [1]. The low efficiency of information transfer associated with low informatization rates leads to reduced productivity and waste of resources. Indeed, Abbaszadegan stated that the lack of information or a delay in the receipt of information may cause 50% to 80% of construction site problems [2]. Accordingly, methods to ensure the accuracy and improve the efficiency of information transmission is a considerable challenge for construction project managers.

In order to eliminate the ambiguity or loss of information in the process of transmission between entities, a more intuitive and effective way to display information is required. Research shows that visualization is one of the more effective ways to improve information management [3]. By presenting information in a pictorial or graphical format, visualization enables decision makers to see the relevant data presented in a way that can improve comprehension of difficult concepts or aid identification of new or changing patterns [4]. Visualization technology can also depict implicit information [5], enabling researchers to see the results of their simulations and calculations and view outcomes and relationships they could not see before. An appropriately designed visual chart can similarly improve the accuracy and comprehensibility of transmitted project information [6], and provide essential information support for managers in making timely and accurate decisions. Research into the application of visualization technology in construction lifecycle management has mainly pursued three objectives [7]: (1) providing interpreted data to the construction team to improve decision making [8]; (2) identifying the disparity between planned and as-built data [9]; and (3) verifying the completeness and accuracy of site data [10].

Because 80% of corporate information is available in a textual format [11], written transfer is an important way to convey project management information. These textual data are large in scale, diverse in format, and scattered in content [12]. Thus, many data management issues can be encountered when working with text data, such as complex data retrieval, difficulties in information reuse, and inefficient interoperability between different management systems [2]. Compared to the tedious and error-prone manual extraction of critical information, visualization technology based on text mining can automatically extract key information from the relevant bodies of text and express it in the form of more intuitive graphics and maps, considerably improving the transmission efficiency of textual information and helping to avoid errors caused by subjectivity in the extraction process.

In the construction industry, text analytics have already been used to classify project documents [13,14], retrieve computer-aided

* Corresponding author.
E-mail address: dadizhong@hust.edu.cn (B. Zhong).

https://doi.org/10.1016/j.autcon.2019.103048
Received 7 May 2019; Received in revised form 25 October 2019; Accepted 5 December 2019
Available online 24 December 2019
0926-5805/ © 2019 Elsevier B.V. All rights reserved.

drawings from databases [15], analyze and predict the occurrence of construction accidents [16], and extract construction regulatory documents [17]. In the field of construction safety management, a natural language processing system was demonstrated by Antoine to extract precursors and outcomes from unstructured injury reports, representing an application of engineering text analysis that does not incorporate any visualization technology [16]. Some research has, however, shown that information visualization approaches can be effectively used to analyze and understand the large amount of data produced during an emergency and to present it on the different devices necessary [18]. By applying text visualization technology, Sizarta provided accurate risk information to petroleum industry managers to aid them in their decision making [19]. Artur verified the advantages of data mining and text visualization technology through an analysis of a large quantity of French and English news articles [20]. Miguel discussed a new research framework applying visual analytics techniques in the safety and risk domains to obtain insight from unstructured text data [21]. Although these methods have taken important steps forward in applications of text mining and visualization, there is currently little research on text visualization related to the identified construction-related objectives, so further research remains necessary.

Indeed, most previous research on text visualization in engineering management has focused on areas such as cost control [3], security management [7,22], emergency management [18], and risk management [19], whereas little research has been conducted on the format of visualizations produced from actual reports generated during the construction process. Such project report text provides a comprehensive description of the construction site conditions over the relevant period of time, and often involves addressing the status of multiple management objectives such as quality, safety, and construction progress. The use of text mining and visualization analysis methods to process and refine such report text can ensure that the project manager is able to effectively obtain the relevant information and has an intuitive and accurate understanding of the state of the project.

Accordingly, in this paper, a framework for visualizing engineering report text is proposed that includes data extraction, visual mapping, and visual display systems. The Term Frequency–Inverse Document Frequency (TF–IDF) algorithm is used to extract the keywords of a project from monthly report material. Then, by using keyword knowledge analysis and multidimensional scaling analysis, the relative importance and mutual clustering relationships of each keyword identified in the text are clarified. Finally, the monthly report material is visualized in the form of a tag cloud by an algorithm to provide an intuitive and easy to understand picture of the current state of the project. The usefulness of the proposed framework to project managers is then evaluated using a set of questionnaires given to construction professionals with various degrees of education and experience.

2. Methodology

The main applications of text visualization technology include: (1) retrieving visual document similarity [23], (2) displaying text content [24], (3) visualizing the sentiments of text [25], (4) exploring document corpora [26], and (5) analyzing various domain-specific rich-text corpora [27]. Because the primary research objective of this study was to help managers quickly obtain effective information by reviewing a single engineering document, a visualization framework for revealing the content of engineering text is proposed. The proposed framework consists of three components: information extraction, visual mapping, and visual display. In this study, the TF–IDF algorithm was used for information extraction and modified with an eye toward improving accuracy when applied to the unique characteristics of engineering texts, and its improvement over the conventional TF–IDF algorithm was experimentally evaluated. Finally, the proposed visualization framework was applied to the visualization of an engineering text corpus case study, and then validated using the results of a set of questionnaires given to construction professionals.

2.1. Construction information visualization framework

To visualize document content, Rusu et al. [28] used a node-link diagram based on a semantic graph extracted from the document. Strobelt et al. [29] introduced a system for transforming a document into a series of cards, in which the content of the document was summarized via keywords and critical figures extracted from the document. Stoffel et al. [30] proposed a technique for producing a thumbnail of a document based on keyword distortion. These techniques help to visualize two critical aspects of a text through visualization technology [31]: (1) word-level content features such as words and figures, and (2) document-level content features such as average sentence length and number of verbs. In this study, to meet construction managers' need for information, an approach that depicts text content at the word level was proposed.

Research [31] has shown that the smallest unit of information transmitted by a text is the keyword (which also includes key terms), so directly illustrating the keywords of a document is the most intuitive approach for succinctly and effectively presenting document content. The accuracy of information acquisition can be improved considerably by using keywords, allowing the user to quickly and efficiently determine whether further investigation of the text is required [32]. In previous research, a support system that does not require either candidate dictionaries or adding key information annotations to the text content was proposed that relies entirely on the system to automatically extract words in the text content as keywords [33]. The implementation of this method can simplify the work of dealing with the large amount of text generated during a project while reducing the time it takes managers to obtain construction information.

In order to accurately extract and present key information from a construction engineering text, three steps must be undertaken: information extraction, visual mapping, and visual display, as shown in the simple construction information visualization framework in Fig. 1.

Information extraction is the most basic step in the proposed process and constitutes the core of construction status information visualization [34]. First, a keyword extraction technique based on the revised TF–IDF algorithm, discussed in Section 2.2, is employed to obtain a core vocabulary that reflects the key information in the engineering text. In this step, critical processes include text preprocessing, word frequency statistics, TF–IDF calculation, candidate word filtering and merging, and keyword screening extraction. Next, visual mapping is conducted to transform the data so that it can be processed into a visual structure [35]. Using visual mapping, the project information extracted by the TF–IDF algorithm and associated processes can be standardized and structured, and can depict more detailed status information by mining the deeper relationships between the keywords. In this study, the social network analysis method is adopted for this process. By analyzing the keyword centrality and the multidimensional scaling and clustering effect, the relationship between the keywords in the engineering text can be determined in order to provide the structure of the extracted status information. The visual display is then used to present the project status information to the manager as human beings are more sensitive to pictures and graphics [36] than to pages and pages of dense and dull words and numbers, and information transmitted through images is more intuitive and lasting. A tag cloud was selected as the visualization method in this study, in which the results of the centrality calculations and the multidimensional scaling and cluster analysis in the visual mapping stage were comprehensively considered.

Three objectives can be achieved using the proposed improved TF–IDF algorithm and multidimensional scaling clustering analysis framework: (1) words are visually and aesthetically presented using a tag cloud that can clearly describe the content of the text; (2) the semantic relationships between words in the text, which are typically not considered in traditional tag cloud technology, are retained,

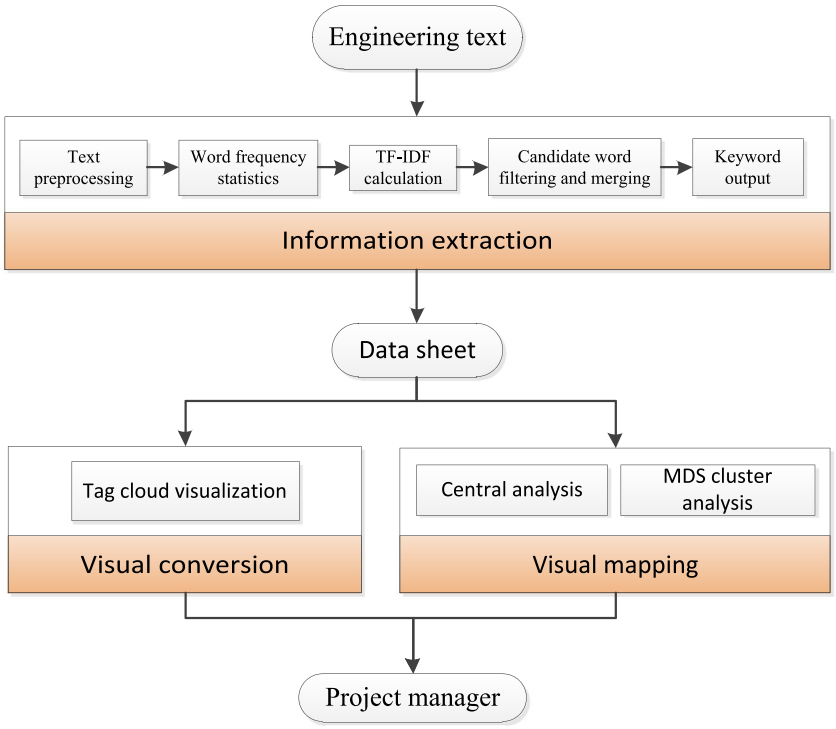

Fig. 1. Simple construction information visualization framework.

summarized, and presented; and (3) the keywords reflecting the main theme of the text are extracted more accurately and word-level patterns such as repetition and co-occurrence between words are clearly depicted.

2.2. Key information extraction based on the improved TF–IDF algorithm

One of the main methods for Chinese keyword extraction is the TF–IDF algorithm [37], consisting of a term frequency component that measures the frequency of terms in the document and an inverse document frequency component that measures the universal importance of these words. The basic concept of TF–IDF is that a word can be considered to be important if it appears frequently in one segment of text while appearing less (or not at all) in others [13]. The complete TF-IDF formula is shown as formula (1):

$$tfidf_{i,j} = tf_{i,j} \times idf_i \tag{1}$$

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2}$$

$$idf_i = \log \frac{N}{n_i} \tag{3}$$

Inverse document frequency idf_i —Assume that if the number of texts containing the word i in a document set is less, the larger the value of idf_i of the word, which means the greater the difference of the word i, so that word i is more likely to be a keyword. The calculation formula is shown as in (3), where N represents the total number of texts in the document set, and n_i represents the number of texts in which the word i appears. $tf_{i,j}$ is the frequency at which the word i appears in the text j. In order to prevent the preference of $tf_{i,j}$ for long text, the formula (2) is generally used to normalize $tf_{i,j}$ in (1). In the formula (2), $n_{i,j}$ is the frequency of the word i appears in the text j, while the denominator $\sum_k n_{k,j}$ represents the sum of times all the words appearing in the document j.

Because the traditional TF–IDF algorithm was not proposed for extracting information from engineering texts, several targeted improvements are required to make it more accurate in this application. Primarily, the structure of the sentence needs to be considered when extracting keywords via text mining [38]. Accordingly, in this study, the basic TF–IDF algorithm was improved to increase the accuracy of keyword acquisition considering the length, position, and lexical characteristics of the engineering application-specific vocabulary. The revised algorithm implementation steps are as shown in Fig. 2.

Generally, the title words or those in the first/last sentences can best express the main content of a document [39]. Therefore, the location information of the title words were used to increase the weight of the title words in the extraction. Using 100 words as the base unit, a weighting formula was accordingly established using the dynamic weighting method as formula (4):

$$\text{weight}_{title}(w_i) = \begin{cases} 0 \; w_i \notin S_{title} \\ 0.5 + \dfrac{100}{\text{docLen}_{d_j}} \end{cases} \tag{4}$$

where S_{title} represents the set of title words in the text and docLen_{d_j} indicates the total number of words in document j. When considering a non-title word, the weight takes a value of 0, and when considering a title word, the weight is set to a reference value of 0.5 and then determined according to the sum of the reference value plus the ratio of each 100 words to the length of document j to determine the weight of the title vocabulary.

Generally, more information can be carried by a long word than by a short word. Berend determined the influence of different word features on the results of keyword extraction, finding that the influence of word length is significantly higher than that of other features [40]. Accordingly, the word length was included in weighting the candidate words using the following equation:

$$weight_{len}(w_i) = \frac{len(w_i)}{\max_{len}} \tag{5}$$

where $len(w_i)$ represents the length of word w_i, which, in order to prevent the excessive dominance of long words, is normalized by the denominator $\max_{len}$, which represents the length of the longest word in

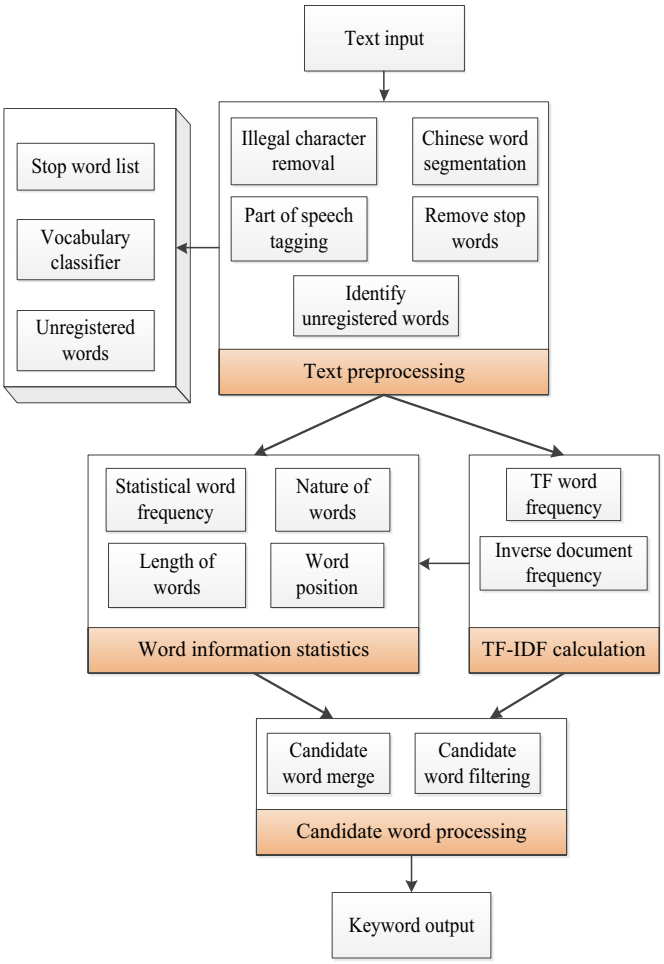

Fig. 2. Modified TF–IDF algorithm.

the subject document.

Research has also shown that nouns are much more important than verbs or other parts of speech, so text keywords are typically nouns and noun phrases while few are adjectives [41]. Accordingly, considering the semantic information carried by the text, the part of speech should play a decisive role in the extraction of keywords. This can be accomplished using word segmentation. Chinese word segmentation is a processing technology unique to Chinese text that divides a sentence into a series of words. Then, part-of-speech tagging determines the functional nature of each word and tags it according to its context. If the assigned word tag is incorrect, the error will be enlarged in subsequent processing, which will directly affect the extraction results. Therefore, in this study, candidate words were scored according to their different parts of speech, using the part-of-speech category weights shown in Table 1.

In Chinese word tagging categories, there are certain words that can appear as both verbs and nouns in the same text set. For example, the word "建议" can express both "suggest" and "suggestion." Assuming that the word "建议" appears in a document n times as a noun and m times as a verb, if its suitability as a keyword is conventionally weighted according to its part of speech and word frequency, its score would be $0.8*n + 0.5*m$, which yields a lower score than the equivalent noun with a frequency of $n + m$ and a higher score than the equivalent verb with a frequency of $n + m$. In view of this phenomenon, to appropriately consider the lexical problem, the approach proposed in this paper similarly calculates the frequency score for the same words in different parts of speech by modifying the basic term frequency equation to the following:

Table 1
Parts of speech and TF weight setting in keyword extraction.

Vocabulary characteristic	Part of speech	Weight
Verb	COM-VERB, NVERB, NVERB-N	0.5
String	STRING	0.7
Noun	ADJ-NOUN, COM-NOUN	0.8
Unregistered words	SW	1.0
Proper noun	SPNAME, ELSE-PRONOUN, NAME	1.2

$$POSTF\ (w_i) = \sum_{j=0}^{n-1} weight\,(POS_j) * fre\,(POS_j) \tag{6}$$

where $weight\ (POS_j)$ is the jth part-of-speech category weight of candidate word w_i according to the standard weights shown in Table 1, and $fre(POS_j)$ is the number of times the candidate word was marked as belonging to part of speech POS_j.

In this study, we used the part-of-speech weights shown in Table 1 combined with the inverse document frequency characteristics to define the following keyword scoring equation:

$$Score\ (w_i) = POSTF\ (w_i) * IDF * (weight_{title}(w_i) + weight_{len}(w_i)) \tag{7}$$

where IDF represents the inverse document frequency of the word w_i, $POSTF(w_i)$ is the part-of-speech weighted frequency score of w_i given in Eq. (6), $Weight_{len}(w_i)$ represents the word length frequency of w_i, and $weight_{title}(w_i)$ represents the position frequency of w_i. After calculating $Score(w_i)$ of words w_i, the candidate words are sorted according to score and the top N candidate words are selected as the text keywords according to the needs of the analysis.

2.3. Experimental evaluation

In order to verify the modified keyword extraction method, an experiment was designed using 100 Chinese news documents and a word segmentation system provided by the Natural Language Processing & Information Retrieval Sharing Platform.[1] These news documents are selected from the corpus according to the theme to ensure that they are engineering related. Since the corpus provides a corresponding set of subject words, the set of subject words is used as a standard set in the experiment to eliminate the influence of subjective factors and ensure the reliability of the experimental results. The traditional TF–IDF method (baseline) and the revised TF–IDF method proposed in this paper were then each used to identify and filter candidate words and select the top ten keywords.

The precision rate and recall rate were used to quantify the effectiveness of the proposed method. The precision rate is typically used to measure the systematic error between experimental and control results. In the keyword extraction process, the precision rate can be regarded as the ratio of correct words in the set of extracted keywords (determined by manual selection) to all extracted words. The recall rate is calculated as the ratio of the extracted keywords to the manually selected keywords, and is mainly used to evaluate whether the system can extract the correct keywords. The precision rate P and recall rate R were calculated as follows, respectively:

$$P = \frac{|K_{auto} \cap K_{man}|}{|K_{auto}|} \tag{8}$$

$$R = \frac{|K_{auto} \cap K_{man}|}{|K_{man}|} \tag{9}$$

where K_{auto} is the set of automatically extracted keywords and K_{man} is the set of manually selected keywords. When the count of automatically extracted keywords is sufficiently large, the count of correct keywords

[1] http://www.nlpir.org/.

Table 2
Experimental results of traditional and proposed TF–IDF methods.

Method	Average precision rate	Average recall rate	F-value
Traditional TF–IDF method	52.73%	70.24%	60.24%
Proposed TF–IDF method	67.34%	85.79%	75.45%

may increase and lead to an improvement in the recall rate, but the effects of this scenario on the precision rate are less clear as both the numerator and denominator of the precision rate calculation are likely to increase. As a result, the F-value is often used to measure the effectiveness of an algorithm based on the precision and recall rates as follows:

$$F = \frac{2 \times P \times R}{P + R} \tag{10}$$

The first 3% of words obtained using the revised TF–IDF weighting were evaluated accordingly after removing words that were not nouns, adjectives, adverbs, or verbs. The first five of these words were then selected as the keywords, yielding the complete experimental results shown in Table 2, in which it can be seen that the revised TF–IDF method exhibits significantly improved keyword extraction. Compared with the traditional TF–IDF algorithm, the average precision rate and recall rate are increased by 14.61% and 15.55% with the F-value increased by 15.21%. Compared with the conventional TF–IDF method [42], the proposed revised algorithm has the advantages of not requiring that the number of multi-word expressions be specified and of reducing the number of filtering rules required.

3. Visualization process for engineering text

In this section, the proposed engineering text visualization framework described in Section 2 is applied to an engineering case study from the construction of the Second Jingzhou Yangtze River Bridge.

3.1. Case study project profile

The Second Jingzhou Yangtze River Bridge is a double-span cable-stayed bridge located in Jingzhou City, Hubei Province. It has a total length of 6317.8 m, a bridge deck width of 24.5 m, and cost in excess of $26 million. As a major highway crossing over the Yangtze River (the longest river in China), the construction of this bridge is an important transportation project for the development of the ecological and cultural tourism circuit in western Hubei. Therefore, the project is of critical importance to local government departments.

Because of the large scale of this project, there are 15 construction teams and dozens of suppliers. In order to help managers understand the construction site, each subcontractor periodically reports on the construction status of the tender for which it is responsible, including progress, quality, and safety. However, reviewing dozens of such reports undoubtedly represents a burden on managers, so it is necessary to provide a more intuitive and effective way of information transmission. Accordingly, the engineering text visualization framework proposed in this paper, which aims to help managers quickly obtain key information from a report, was applied to a one-month progress report text from this bridge project in order to confirm the potential benefits of visualization in improving the efficiency of information transmission.

3.2. Word-level information extraction

3.2.1. Text preprocessing

The purpose of text preprocessing is to prepare the document so that it meets the input requirements of the extraction model. Many artifacts must be removed in this step, including formatting tags, illegal characters, and stop words, and any unregistered words must be accounted for.

Word documents contain a large number of formatting tags inserted in the process of editing that can introduce considerable inaccuracies in the text statistics and reduce processing efficiency. Furthermore, there are "illegal characters" in most text. These special characters are not related to the text information statistics and include special symbols such as "Y", "?", "#", " > ", "x", "@", punctuation marks, mathematical symbols, and etc. Because they cannot be interpreted as text features, such formatting tags and illegal characters must be removed during the text preprocessing stage.

Stop words appear frequently in a document but are not representative of its subject matter [11]. They include modal particles, adverbs, pronouns, articles, prepositions, and conjunctions. Because of the small amount of information carried by stop words, the efficiency and accuracy of keyword extraction can be improved after deletion of such words.

Unregistered words are those that are not recognized by the custom dictionary and include the names of people, places, terms, and etc. Such unregistered words are unavoidable due to the limited capacity of custom dictionaries, but their presence is an important factor affecting the accuracy of word segmentation. To deal with unregistered words, we created an engineering project text thesaurus by collecting more than 1000 engineering terms related to project management. The words in this thesaurus may not register in the custom dictionary, but will not be removed.

After text preprocessing to remove formatting tags, illegal characters, and stop words, the selected monthly construction report text contained 1734 words. Now, we can calculate the term frequency of any particular keyword. The modified algorithm proposed in this paper then uses the network retrieval data of this text as the corpus for the inverse document frequency calculation, which is reflected by the equation $IDF_i = \log \frac{N}{n_i + 1}$. The product of the term frequency and inverse document frequency components is the TF–IDF score for each high-frequency vocabulary term.

3.2.2. Filtering and merging candidate words

In order to reduce the influence of noise in the process of keyword extraction, in the proposed framework all words are filtered through certain rules before attempting to identify the best candidate keywords. This filtering is performed according to the part of speech and frequency of each word.

It is well known that keywords are typically nouns or verbs rather than adjectives or adverbs. Therefore, the keyword extraction method proposed in this study only considers nouns, verbs, named entities, and their unregistered synonyms while filtering out other parts-of-speech [46]. Additionally, some common high-frequency verbs, such as "go" and "complete," that often appear in documents but have little importance should not be treated as keywords, so we established rules to filter verbs hardly used in headings of with word length less than two.

In frequency filtering, the inverse document frequency hypothesis states that words that appear in multiple documents are unlikely to represent a single document topic. In order to avoid the influence of high frequency words on the keyword extraction results of the proposed method, words that appear in documents with more than one third of the total number of documents were removed from the extracted keyword set. It's worth noting that the condition is that there are more than 10 documents in the document set. Additionally, considering the relationship between the frequency and position of candidate keywords in

the text, only words with frequencies greater than 1 or appearing in a single instance in the title were retained.

Research by Li has demonstrated that most text keywords are composed of key phrases, and that many existing word segmentation processes may end up separating such key phrases [43]. For example, in typical engineering text extraction, the phrase “Image progress” would be divided into the keywords “image” and “progress,” which could obviously have a considerable impact on the extracted keywords and their inferred meaning. Therefore, after the candidate keyword weights were calculated, instances of merged terms were extracted. Then, the frequencies in which the candidate keywords appeared together in the text were determined and a frequency threshold was set to identify instances of meaningful keyword phrases. The candidate keywords were then merged as follows. Firstly, the first 20 words or phrases with large weight are selected from the candidate vocabulary as the merging object. Secondly, for the candidate keyword phrases, the number of the same words between them and other phrases is counted. When the number is greater than 2 or equal to 2, the phrases with more words are selected as the new candidate keyword phrases. The weight of the new candidate keyword phrases is the sum of the two. Finally, for keywords, we use the common method of two-two combination to calculate the co-occurrence frequency and extract the candidate keyword phrases. The weight of the phrases is also the sum of the two words. The resulting keyword scores are shown in Table 3, and was determined using the Python programming language to implement the proposed modified algorithm.

The text keywords obtained by the information extraction step and their TF–IDF scores reflect the theme of the text to a certain extent, but it remains difficult to express the relationship between these words, undoubtedly leading to the loss of information. However, a chart can be more convincing and descriptive than simply text and data, and can convey more information that can be more intuitively felt. Therefore, in order to present the information in the subject engineering report text more completely, further analyses for subsequent visual depiction are necessary.

3.3. Visual mapping based on keyword relevance

The construction management-relevant information reflected by a single independent keyword is limited. In order to more completely present textual information in a visual form, we need to determine the relationships between the identified keywords. The tightness between keywords can reveal the general theme of the project text as well as the status of the project to a certain extent. Such mapping of the knowledge domain is accomplished using a series of different graphs that depict the relationships between the knowledge development process and document structure. This data mining method uses visualization technology to describe knowledge resources and their interconnections [44].

Table 3
Keyword extraction statistics for case study text.

No.	Vocabulary	Score	No.	Vocabulary	Score
1	Completed output value	0.0871	11	Quality control	0.0625
2	Image progress	0.0829	12	Sand and gravel supply	0.0613
3	Box girder prefabrication	0.0796	13	Production plan	0.0592
4	Subgrade	0.0793	14	Special equipment	0.0577
5	Pile foundation pit	0.0758	15	Duration	0.0574
6	Standardization	0.0725	16	Equipment management	0.0534
7	Safety management	0.0702	17	Pile foundation inspection	0.0532
8	Super bridge	0.0657	18	Rectification notice	0.0507
9	Material procurement	0.0646	19	On-site acceptance	0.0495
10	Measuring output value	0.0644	20	Land acquisition and demolition	0.0492

Managers require project management information mainly to determine the relationship between the objective of the project and the theme of the project text in question. On this basis, the framework proposed in this paper conducts a centrality analysis of the keywords in the project text to determine their relative importance; the greater the centrality of a keyword, the greater its importance in the network, thus the better it reflects the core theme of the text. Thus, using a multi-dimensional scaling and clustering analysis, the framework proposed in this study reveals the clustering relationship of keywords and explores their relationship to the themes of the text.

In the process of information visualization, the first step is to create a “Bipartite matrix” to store the collected analysis data. The rows and columns in the matrix correspond to the nodes in the social network, and the matrix elements corresponding to rows and columns represent the relationship between the nodes. The matrix is a framework in which rows and columns are composed of coded data, so that the matrix is an individual-individual matrix. In the Individual-Individual Data Matrix, rows and lists represent a series of identical individuals, in which each actor is expressed twice, once in a row and once in a column. In this case, the adjacency matrix is a binary matrix. If the element in the matrix is 1, there is a social relationship between the nodes corresponding to the element. On the contrary, if it is 0, there is no relationship between the corresponding nodes.

A network map was then drawn using the NetDraw software to visually describe the distribution of and interrelationships (lines) between the text keywords (nodes). Construction engineering texts generally consist of paragraphs distinguishing different aspects of a project, so keywords that appear in the same paragraph are considered to be more interrelated. However, keywords appearing both in the title and in subsequent paragraphs are not in the same paragraph, but still usually have strong relevance to each other. Therefore, in the matrix describing the interrelationships of the keywords extracted from the engineering text, the elements representing keywords in the same paragraph and those common between the title and the subject paragraph keywords were set as 1, indicating that they have a strong correlation. To extract the monthly project report keywords in this study, the relationships captured in the two-matrix set were analyzed as discussed in Sections 3.3.1 and 3.3.2.

3.3.1. Keyword centrality analysis

Centrality is one of the priorities of a keyword analysis. There are three main indicators of keyword centrality: degree centrality, closeness centrality, and betweenness centrality. The degree centrality of the nodes in a network can be described in absolute and relative terms. Absolute centrality refers to the number of other nodes directly connected to a certain keyword node; a higher value indicates that the keyword is located nearer to the center of the network and thus has a greater ability to affect other keywords. Relative centrality is the ratio of the absolute centrality of the keyword to the maximum absolute center of the network. Betweenness centrality measures the degree to which the keyword node controls the resource, meaning that if a node is on the shortest path between many other node pairs, that node has a higher betweenness centrality and thus importance. Finally, the closeness centrality describes the magnitude of a given keyword's influence on a network by measuring the distance between its node and the other connected nodes. The farther the given keyword node is from the network center, the smaller its closeness centrality, and thus the weaker its influence on the network.

The keyword centrality analysis conducted by UCINET first converted the two-mode network data into the “Bipartite matrix” discussed previously and then analyzed the various centrality indicators of the network. The results of these centrality analyses are shown in Table 4, in which it can be seen that the degree and closeness centralities of the term “image progress” are the highest with values of 0.688 and 0.865, respectively, and its betweenness centrality of 0.007 indicates that this term is relatively close to the center of the network. This suggests that

Table 4
Keyword centrality statistics for case study text.

Keyword	Degree	Closeness	Betweenness
Completed output value	0.313	0.544	0.017
Image progress	0.688	0.865	0.007
Box girder prefabrication	0.438	0.780	0.025
Subgrade	0.500	0.780	0.038
Pile foundation pit	0.438	0.762	0.025
Standardization	0.500	0.800	0.034
Safety management	0.313	0.711	0.010
Super bridge	0.313	0.744	0.016
Material procurement	0.313	0.711	0.010
Measurement payment	0.188	0.627	0.004

the term “image progress” is at the core of the network, so it is used as a core keyword for this monthly project report. The keyword “land acquisition and demolition” exhibits the smallest degree centrality of 0.188, indicating that it has little contact with the other nodes in the network. Its closeness centrality and betweenness centrality values are also relatively small at 0.615 and 0.003, respectively, indicating that “land acquisition and demolition” is also far from the core of the network. Overall, the degree centralities of the terms “quality management,” “captain foundation pit,” “production plan,” and “box girder prefabrication” are larger than those of the other nodes, which are all less than 0.5. Indeed, it can be seen in Fig. 3 that the keyword network of the subject construction engineering document is largely concentrated within a few network nodes, with “land acquisition and demolition” having the weakest influence.

On the basis of this processed data, the NetDraw software was used to draw a keyword map of the text, shown in Fig. 3. Three levels of knowledge discovery are realized in this map: the division of keywords into several smaller classes with intrinsically close associations, allowing for the discovery of new subdivision topics; the ability to judge the relationships between nodes, providing clues as to their relative importance; and the ability to determine the primary content in each topic, revealing the core meaning of the text. Furthermore, visualizing centrality can help the reader to intuitively understand the extent to which each keyword node resides near the core of the network. In the figure, the keyword “image progress” is represented by the largest node and is connected to the most other nodes in the keyword network, indicating that it reflects core information in the text describing the status of the project. The keyword “land acquisition and demolition” is represented by a small node and is at the edge of the network, indicating that it has less effect on the network. The centrality visualization also shows the distribution of keywords in the monthly project report. For example, “quality management,” “captain foundation pit,” “box beam prefabrication,” and “production plan” are at the center of the network with many important connections, indicating that they constitute project information expressed by the core of the text.

The keyword centrality map in Fig. 3 visually expresses project management information using different element intensities, represented by node size and location in the network, to reflect the degree of linkage between keywords. Keywords displayed as nodes with a higher intensity reflect their importance in the project text and thus the status characteristics of the project. Additionally, the linkages between keywords allow the current status and trend of the project to be explored, and scattered keywords can be combined to express more complete and accurate information. For example, in Fig. 3 the nodes representing “quality management” and “image progress” are larger and in the center of the network, indicating that they have the most connections with other keywords. Based on this observation, we can reasonably speculate that this project report text is centered on quality management and image progress, with related secondary topics including “pile foundation inspection,” “box girder prefabrication,” “pile foundation,” and “production plan.” Through the centrality relationships indicated by the direction of the arrows connecting nodes, information such as the complete topic and its related relationships in the engineering text can be clarified, but the relationship between the keywords cannot be determined. Therefore, further cluster analysis is necessary using multidimensional scaling.

3.3.2. *Multidimensional scaling and clustering analysis of keywords*

The overall multidimensional scaling analysis, also known as a “multivariate scaling analysis,” or the “multidirectional measurement method,” is a multivariate analysis method that attempts to transform individual dissimilarity and similarity data into a multidimensional

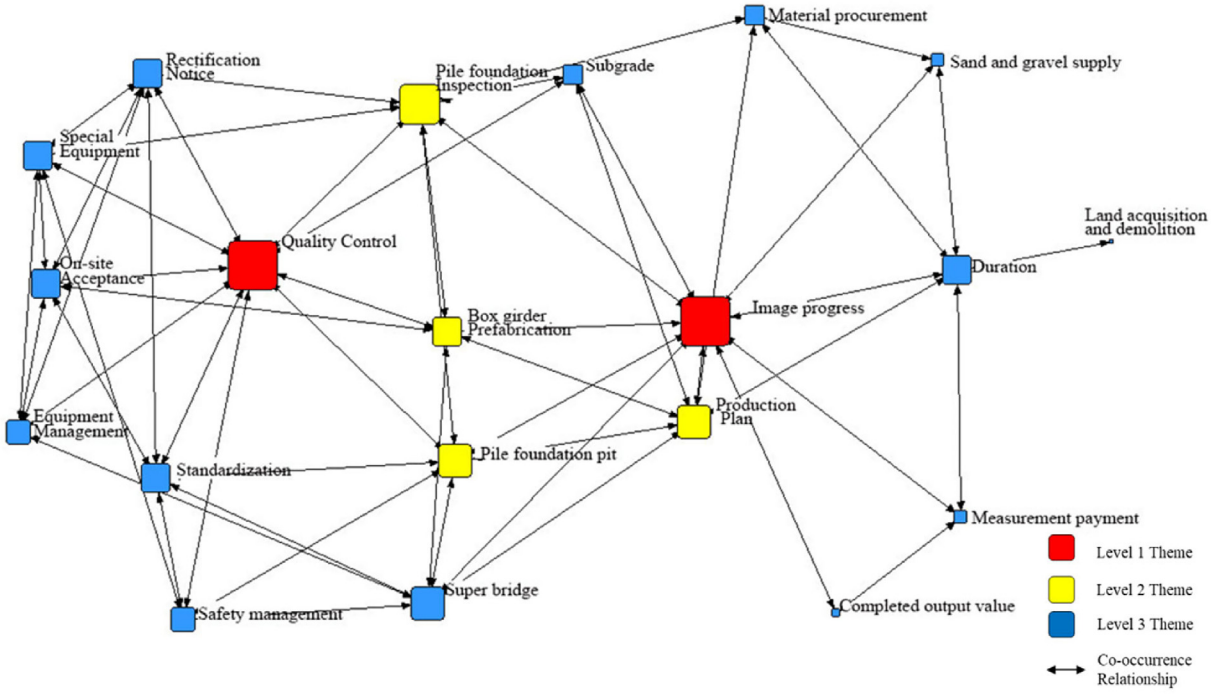

Fig. 3. Keyword centrality analysis map.

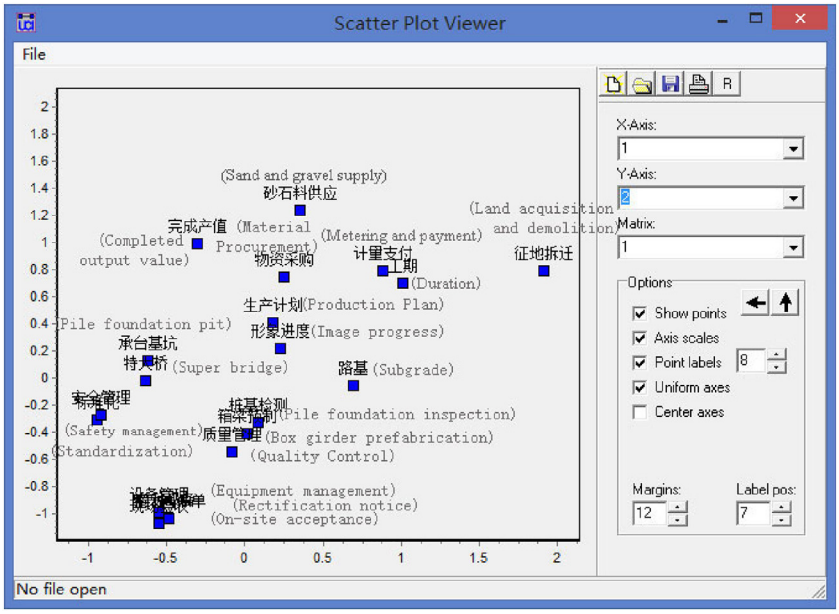

Fig. 4. Coordinate display of keywords from the multidimensional scaling analysis.

spatial map while retaining the relative relationships between raw data points to the extent possible. When faced with of large quantities of raw data in the form of text and symbols, multidimensional scaling can help to process data by finding their structural relationships. Therefore, multidimensional scaling can visually represent the relationship network of the extracted engineering construction project text keywords, and can analyze the assignment data and cluster the keywords to display their relationships as distances in multidimensional space. After importing the two data matrices, setting the research dimension number (as discussed below), and defining the transformation criterion for the multidimensional scaling analysis (also discussed below), the dimensional coordinate data can be obtained to draw the relationship map. When using multidimensional scaling to spatially cluster keywords and extract potentially important topics, managers can employ multi-level design and strategy selection to evaluate topics and their connections at different levels of observation according to management needs, and to explain relationships between identified topics.

In this paper, the analysis dimension was set to 2 and similarity was used as the conversion criterion. The multidimensional scaling coordinate data was then obtained as shown in Figs. 4 and 5 below. It can be clearly seen in Fig. 4 that a considerable clustering effect exists for four groups of terms in the lower left corner. For example, the coordinates of "safety management" and "standardization" are (−0.921, −0.274) and (−0.948, −0.311), respectively; this close distance indicates that these two keywords have considerable relationships in the network, and demonstrates that there is a great deal of association between the two terms in the project text, thus they must reflect the same type of information. Other keyword groups with notable clustering effects are "box girder prefabrication", "pile foundation detection", and "quality management"; "bearing foundation pit" and "special bridge"; "equipment management" and "remediation notice"; and "on-site acceptance" and "special equipment." It can also be seen that "land acquisition and demolition" is far away from the other keywords, indicating that it expresses management information different from the other core keywords.

After the centrality and multidimensional scaling and clustering analyses, the data was imported to provide a multidimensional scaling visualization of the monthly project report text keywords using the NetDraw software, shown in Fig. 5. It can be seen from the figure that the "image progress" and "quality management" terms are more central, thus indicating the two focal points of the text. Thus, the key status information in the subject project report text is mainly image progress information and quality management information. The analysis results show that "special equipment," "rectification notice," "site acceptance," and "equipment management" are relatively close in multidimensional space, and that these words are spatially related to "quality management," "safety management," and "pile foundation," indicating that these four keywords reflect a specific problem occurring in the project. Based on this observation, it is reasonable to speculate that one of the main pieces of information contained in the text is that when performing equipment management on special equipment, on-site acceptance is required. If special equipment that does not meet the standard is found, a rectification notice should be issued in time for rectification. The related keywords are mainly "pile-based inspection," "quality management," and "safety management," which suggests that special equipment management is a focus of quality and safety management objectives, and that "pile-based inspection" and "special equipment" are required for the work.

The centrality and multidimensional scaling analyses of the keyword network can be used to further explore the engineering text. As highlighted in the box in Fig. 5, the node representing "sand and gravel supply" is closely located to and intimately connected with the node representing "construction duration," indicating that the two are highly correlated in the original text. We can therefore reasonably speculate that in this progress report, issues affecting the construction duration caused by the sand and gravel supply are mentioned. Similarly, "land acquisition and demolition" and "construction duration" also exhibit a close relationship, indicating that land acquisition and demolition work also has an impact on the construction duration, but not as much as the sand and gravel supply, as the relationship between the nodes is not as tight. Thus, in order realize the construction duration goal, the manager should solve issues related to the supply of sand and gravel first, followed by those related to land acquisition and demolition work. Note that although the distance between "land acquisition and demolition" and "sand and gravel supply" is relatively close, there is no arrow between the two, indicating that land acquisition and demolition may affect construction duration, but does not affect the sand and gravel supply, which is in line with common sense judgment. The multidimensional scaling and cluster analysis can clearly improve reading efficiency by emphasizing most important keywords reflecting the theme of the original text. By removing redundant and non-critical information to provide a concise and intuitive view of the text topics,

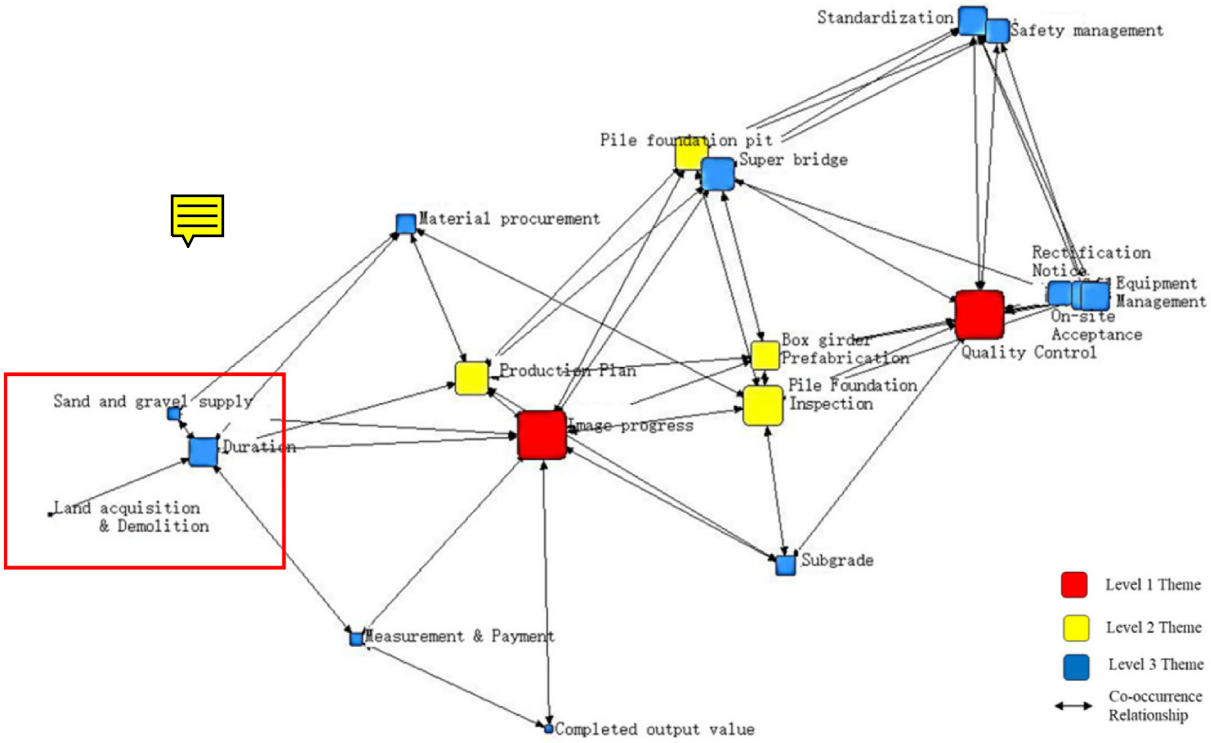

Fig. 5. Keyword multidimensional scaling analysis map.

Fig. 6. Engineering text information visualization tag cloud.

meaningful project management information has been extracted from the report and clearly depicted, providing an effective basis for the visual design of a tag cloud for expressing project status information, as discussed in Section 3.4.

3.4. Visual display (tag cloud) results

In order to clearly and comprehensively reflect the information contained in the text, it is necessary to effectively visualize the data mining results. A tag cloud visualizes keyword content in the form of tags aggregating the content of the subject text, and is a common display format in segment classification. Size, distance, color, and other indicators can be used to clearly reflect different attributes of the displayed keywords. For example, tags with higher frequencies are usually displayed in larger fonts.

To implement the tag cloud, scalable vector graphics technology was used to depict the size of the keyword based on the TF–IDF score and the degree centrality based on color (eye-catching colors like red and yellow indicate large degree centrality and cooler colors such as grey or blue indicate small degree centrality). The keywords were laid out using the two keywords with the largest centrality as the center with the results of the multidimensional scaling and cluster analysis used to determine the distance between these and all other keywords, which were then arranged radially around the center. Fig. 6 shows the visualized project management status information based on the proposed keyword analysis framework.

Using the tag cloud in Fig. 6, a manager could quickly determine that the following management information was contained in the project text:

(1) The “image progress” and “quality management” terms in the picture are bright and colorful in the center of the picture, indicating

Table 5
Questionnaire results.

Investigation item	Mean (construction supervisor/20)	Mean (senior engineers/15)	Mean (project manager/10)	Mean (all respondents/45)
1. Necessity and practicality				4.00
1.1. It is necessary and useful to visualize management information.	3.68	3.86	4.09	3.88
1.2. The method can help users master the necessary on-site information.	3.88	4.04	4.15	4.02
1.3. It is easy to understand and use of the information visualization framework.	4.03	4.01	4.28	4.11
2. Increasing Information comprehension				3.94
2.1. The system can help users to understand the situation on the construction site more intuitively and deeply.	3.83	4.29	4.40	4.17
2.2. Visual display delivers more hidden information to users.	3.67	4.11	4.17	3.98
2.3. Visualized information is easier to understand and transmit.	3.58	3.58	3.84	3.67
3. Decision aid				3.65
3.1. The system can improve on-site management decision-making.	3.50	3.21	3.14	3.28
3.2. The system can assist a user in analyzing the potential problem.	3.61	3.98	3.71	3.76
3.3. Visual information takes less time to review and understand than traditional text formats.	3.82	3.95	3.95	3.91

Note: the mean score is calculated from respondents' feedback on five-scale questionnaire: 1 (strongly disagree), 2, 3, 4 and 5 (strongly agree).

that these terms have a high centrality and weight. It is not difficult to speculate therefore that the core content of the project text is related to management information describing both progress and quality management.

(2) The main work conducted on the project in the month for which the report was written was the foundation pit project because in the tag cloud layout, the term “pile foundation pit” is near the center of the map indicating a greater degree of centrality, and its clustering close to the major terms “image progress” and “quality management” further clarify the main purpose of the text.

(3) The construction duration was affected due to the supply of sand and gravel during material procurement, because the keywords “material procurement,” “sand and gravel supply,” and “duration” are clearly associated by their proximity.

In this way, potential themes in the text are depicted in visual space. The traditional linear text structure is transformed into a visual structure based on spatial display, which is more in line with typical human preferences for and experiences of information acquisition. This improves the ability of project managers to understand the status of their project and identify systematic or tacit knowledge that is difficult to obtain by traditional reading.

4. Evaluation and discussion of results

In order to confirm that the proposed text visualization framework is able to effectively improve the efficiency of construction project information transmission and provide the necessary information to aid managerial decision-making, a quantitative questionnaire and an experiment evaluation, were conducted based on the proposed framework.

For the quantitative questionnaire, respondents were asked to independently score a series of statements regarding their impressions of the text visualization framework proposed in this paper. The respondents were experienced construction site management professionals from several well-known contractors in Wuhan City, Hubei Province. These contractors were selected to evaluate the proposed visualization framework because they usually have more experience in large and complex construction projects that require a high level of standardization and information management, and as such they are typically used to working with information technology automation in the course of construction site management. Thus, the selected respondents were likely to be both willing and qualified to test the visualization framework and provide feedback. During the evaluation, the team introduced the visualization framework, including the source of the visualized data, and explained how to interpret the results. Participants were then asked to complete the questionnaire using a five-point Likert rating scale [45]. The evaluation involved requesting that the invited managers and engineers use the system to complete a questionnaire to provide feedback.

The questionnaire was divided into three sections. Section 1 obtained participant's details. Section 2 evaluated the possible benefits of the proposed system from three aspects: necessity and practicability, increasing information comprehension, and assisting on-site decision-making. In section 3, participants were given the opportunity to make further comments under section 2. The number of the effective respondents was 45 including 20 construction supervisors with nearly five working-years, 15 senior engineers with ten working-years and 10 project managers with fifteen working-years.

The results of the questionnaire are provided in Table 5, in which it can be seen that the overall feedback was on the positive side, with the weighted average scores of all questions above 3.0, indicating that the visualization framework proposed in this paper was generally effective. Critically, the most respondents agreed that visualizing project management information is necessary and useful with the highest scoring 4.00. The reasons mentioned among those who disagreed with the

Table 6
Experimental evaluation results.

	A		B		C	
Respondent	Project manager		Senior engineer		Junior engineer	
Age/working years	45/15		41/8		33/6	
Highest education	Bachelors		High school		Masters	
Material type	OT	VD	OT	VD	OT	VD
Q1:What is the main subject of the material?	9	8	9	7	8	8
Q2:What is the construction phase of the project?	9	9	7	8	7	7
Q3:What are the main problems of the project?	8	8	8	7	6	7
Q4:What are the possible reasons of these problems?	9	8	7	5	8	7
Q5:What actions should be taken in the next phase?	8	8	8	5	7	7
Total score (out of 50)	43	41	39	32	36	36
Time spent (s)	498	372	570	438	516	378

utility of management visualization were mainly along the lines of "I prefer to read unprocessed raw materials compared to processed information, even if it means more time and effort needs to be spent." Furthermore, the system has a good performance in increasing information comprehension especially in helping managers obtain the necessary high-level and more in-depth information describing the situation on the construction site. The lowest score was for whether the proposed visualization system can effectively improve the quality of managers' decisions. Skeptics of this improvement argued that the proposed visualization framework simply presents the extant information in a more concise and lively form, indeed providing information that was originally more difficult to find to aid decision making, but not providing any effective advice on how to make those decisions. On the other hand, some respondents indicated that it takes longer for them to understand the visual information, while others have reservations about the possible loss of information. It is also noted that groups of different backgrounds show differences in results when answering. Overall, project managers with a longer working-years tend to get higher scores, which may benefit from the richer management experience and a broader management perspective, as they often need to be accountable for all aspects of a project. The exceptions appear in question 3.1, and the group of project managers gives the lowest evaluation among the three groups because they already have rich experience in decision making, so the system can provide relatively little help.

For the experimental evaluation, three construction managers were selected to further explore the benefits of the proposed text visualization framework. Managers with different degrees of engineering experience and educational background were selected in order to capture the viewpoints of the different types of people engaged in actual construction project management. For this evaluation test, the visual display (VD) form of engineering text and the original text (OT) were presented. The three respondents were shown the two materials in turn and asked to independently answer five questions related. The time of the entire questionnaire process was recorded, and the accuracy of each respondent's answer to each question was scored on a 10-point scale. The results are shown in Table 6, in which it can be seen that the visual display of engineering information can significantly reduce the time required for information transfer by around 25% on average. In terms of the quality of engineering information transmission, the average reduction in the score of the three respondents was only 7% from OT to VD, while respondents with higher educational experience were clearly better able to interpret the visually displayed information. Clearly the proposed framework for text visualization presents engineering information in a more intuitive way, but also raises the educational requirements for understanding and adapting to changes in the information delivery medium.

In the summary, text visualization methods are generally considered to be effective in improving the efficiency of engineering information transmission. Investigations in this study demonstrated that the use of visualized textual information reduced the time required to understand the information by about 25% while keeping critical information intact. Although the framework proposed in this paper can thus promote the understanding of the text and visually represent its main aspects, it should be noted that the original text remains relevant in practical use because of the loss of information inherent to the visualization process.

There are still some limitations in the visualization framework. Since that the visual information differs from the traditional form in organizational form, it means that the user must first master the method of interpreting the visual text in order to obtain key engineering information. The actual value of the visualization of engineering text still needs further observation and evaluation. It is also noted that the accuracy of the keyword extraction algorithm could likely be further improved. Combining the proposed visualization framework with semantic and grammar knowledge could be one possible way to improve the accuracy of the presented information. Furthermore, the responses to the questionnaires indicated some suggestions for improvement, such as combining the visualized information with decision-making suggestions to improve decision-making efficiency, and the combined use of the text visualization with the original text to reduce the dependence on managerial experience.

5. Conclusions

This paper proposes and evaluates a framework for creating construction engineering text visualizations. Through the information extraction, visual mapping, and visual display steps of this framework, the information conveyed by a construction report text can be displayed in the form of a visual chart. In the information extraction process, factors such as word length, part of speech, and vocabulary position are comprehensively considered. Different extraction weights are then set according to the amount of information that the keyword is capable of reflecting in the engineering text and the importance of that information to improve the accuracy of information extraction and filtering. Then, based on the extracted keywords, a keyword knowledge map analysis is conducted to determine the centrality, multidimensional scaling, and clustering relationships of the keywords so that the information, otherwise transmitted in a scattered manner, can be presented as a clear and complete presentation of project information. Finally, the engineering text information is visualized in the form of a tag cloud so that managers can obtain key information more intuitively and clearly. The results of manager questionnaires show that such automatically extracted and visually depicted information can considerably improve the efficiency and accuracy of this critical information transfer process. Additionally, some information that would not otherwise be easily discovered by a simple data visualization process (or that could only be determined by time consuming in-depth reading) was mined that could be helpful to managers when making decisions.

Overall, the contributions of this paper are as follows. First, this study proposed and confirmed the efficacy of a visual framework for improving the efficiency of management information transmission and assisting decision-making, as information presented in graphical rather

than textual form can be digested in a relatively shorter time and can be more persuasive. Second, unlike the traditional, time-consuming method of reading engineering texts to extract key information, the use of automatic methods to extract information is more objective and consistent in the process of transmission, avoiding prejudice and misunderstanding caused by personal subjectivity. Finally, this combination of engineering text mining and visualization technology can be used to uncover information within the text and display it clearly, providing more comprehensive information with which decision makers can make more timely and correct decisions. The performance of the text mining and visualization framework proposed in this paper together with the results of manager questionnaires demonstrate that visualization technology based on data mining can compensate for the loss of information in the information transmission process between entities and provide a quick and accurate way of gaining insight into construction project status.

Acknowledgments

This research is partly supported by the National Natural Science Foundation of China (Grant No.51878311, No.71732001, No. 51978302)

References

[1] V. Singh, N. Gu, X. Wang, A theoretical framework of a BIM-based multi-disciplinary collaboration platform, Autom. Constr. 20 (2) (2011) 134–144 https://doi.org/10.1109/IntelliSys.2015.7360c1140.
[2] M. Martínez-Rojas, N. Marín, M.A. Vila, The role of information technologies to address data handling in construction project management, J. Comput. Civ. Eng. 30 (4) (2016) 4015064https://doi.org/10.1061/(ASCE)CP.1943-5487.0000538.
[3] A. Abbaszadegan, D. Grau, Assessing the influence of automated data analytics on cost and schedule performance, Procedia Eng. 123 (2015) 3–6 https://doi.org/10.1016/j.proeng.2015.10.047.
[4] K.C. Garwood, C. Jones, N. Clements, V. Miori, Innovations to identifying the effects of clear information visualization: reducing managers time in data interpretation, J. Vis. Lit. (2018) 1–11,https://doi.org/10.1080/1051144X.2018.1435024.
[5] N. Bazurto-Gomez, J.C. Torres, R. Gutierrez, M. Chamorro, C. Bulger, T. Hernandez, J.A. Guerra-Gomez, An information visualization application case to understand the world happiness report ,Communications in Computer and Information Science, 847 (2019) 44–56,https://doi.org/10.1007/978-3-030-05270-6_4.
[6] I. Spence, William Playfair and the Psychology of Graphs, Proceedings of the American Statistical Association, Section on Statistical Graphics, (2006), pp. 2426–2436 http://psych.utoronto.ca/users/spence/Spence%20(2006).pdf.
[7] Y. Zhou, L.Y. Ding, L.J. Chen, Application of 4D visualization technology for safety management in metro construction, Autom. Constr. 34 (2013) 25–36 https://doi.org/10.1016/j.autcon.2012.10.011.
[8] K. Liston, M. Fischer, J. Kunz, Designing and evaluating visualization techniques for construction planning, Comput. Civil Build. Eng. 2000 (2000) 1293–1300 http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.471.3447&rep=rep1&type=pdf.
[9] M.G. Fard, F. Peña-Mora, Application of visualization techniques for construction progress monitoring, Comput. Civil Eng. 2007 (2007) 216–223 https://experts.illinois.edu/en/publications/application-of-visualization-techniques-for-construction-progress.
[10] C. Chiu, A.D. Russell, Design of a construction management data visualization environment: a top-down approach, Autom. Constr. 20 (4) (2011) 399–417 https://doi.org/10.1016/j.autcon.2010.11.010.
[11] N. Ur-Rahman, J.A. Harding, Textual data mining for industrial knowledge management and text classification: a business oriented approach, Expert Syst. Appl. 39 (5) (2012) 4729–4739 https://doi.org/10.1016/j.eswa.2011.09.124.
[12] L. Soibelman, J. Wu, C. Caldas, I. Brilakis, K.Y. Lin, Management and analysis of unstructured construction data types, Adv. Eng. Inform. 22 (1) (2008) 15–27 https://doi.org/10.1016/j.aei.2007.08.011.
[13] C.H. Caldas, L. Soibelman, Automating hierarchical document classification for construction management information systems, Autom. Constr. 12 (4) (2003) 395–406,https://doi.org/10.1016/S0926-5805(03)00004-9.
[14] M. Al Qady, A. Kandil, Automatic clustering of construction project documents based on textual similarity, Autom. Constr. 42 (2014) 36–49,https://doi.org/10.1016/j.autcon.2012.11.041.
[15] J. Davidoff, Cognition Through Color, 217 The MIT Press, Cambridge, MA, US, 1991, p. 217 https://psycnet.apa.org/record/1991-97922-000.
[16] A.J. 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.2012.11.041.
[17] J. Zhang, N.M. El-Gohary, Semantic NLP based information extraction from construction regulatory documents for automated compliance checking, J. Comput. Civ. Eng. 30 (2) (2013) 4015014,https://doi.org/10.1061/%28ASCE%29CP.1943-5487.0000346.
[18] F. Dusse, P.S. Júnior, A.T. Alves, R. Novais, V. Vieira, M. Mendonça, Information visualization for emergency management: a systematic mapping study, Expert Syst. Appl. 45 (2016) 424–437 https://doi.org/10.1016/j.eswa.2015.10.007.
[19] S. Sarshar, S. Haugen, Visualizing risk related information for work orders through the planning process of maintenance activities, Saf. Sci. 101 (2018) 144–154 https://doi.org/10.1016/j.ssci.2017.09.001.
[20] A. Šilić, A. Morin, J. Chauchat, B.D. Bašić, Visualization of temporal text collections based on correspondence analysis, Expert Syst. Appl. 39 (15) (2012) 12143–12157, https://doi.org/10.1016/j.eswa.2012.04.040.
[21] M. Figueres-Esteban, P. Hughes, C. Van Gulijk, Visual analytics for text-based railway incident reports, Saf. Sci. 89 (2016) 72–76 https://doi.org/10.1016/j.ssci.2016.05.009.
[22] S. Azhar, Role of visualization technologies in safety planning and management at construction jobsites, Procedia Eng. 171 (2017) 215–226 https://doi.org/10.1016/j.proeng.2017.01.329.
[23] J.L. Neto, A.D. Santos, C.A. Kaestner, N. Alexandre, D. Santos, Document Clustering and Text Summarization, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.43.4634, (2000).
[24] O. Kaser, D. Lemire, Tag-Cloud Drawing: Algorithms for Cloud Visualization, arXiv Preprint cs/0703109, 2007. https://arxiv.org/abs/cs/0703109.
[25] V. Benjaoran, S. Bhokha, An integrated safety management with construction management using 4D CAD model, Saf. Sci. 48 (3) (2010) 395–403 https://doi.org/10.1016/j.ssci.2009.09.009.
[26] A. Endert, P. Fiaux, C. North, Semantic interaction for visual text analytics, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2012, pp. 473–482,https://doi.org/10.1145/2207676.2207741.
[27] M. Vidal, G.V. Menezes, K. Berlt, E.S. de Moura, K. Okada, N. Ziviani, D. Fernandes, M. Cristo, Selecting keywords to represent web pages using Wikipedia information, Proceedings of the 18th Brazilian Symposium on Multimedia and the Web, ACM, 2012, pp. 375–382,https://doi.org/10.1145/2382636.2382714.
[28] D. Rusu, B. Fortuna, D. Mladenic, M. Grobelnik, R. Sipoš, Document visualization based on semantic graphs, 2009 13th International Conference Information Visualisation, IEEE, 2009, pp. 292–297,https://doi.org/10.1109/IV.2009.57.
[29] H. Strobelt, D. Oelke, C. Rohrdantz, A. Stoffel, D.A. Keim, O. Deussen, Document cards: a top trumps visualization for documents, IEEE Trans. Vis. Comput. Graph. 15 (6) (2009) 1145–1152 https://doi.org/10.1109/TVCG.2009.139.
[30] A. Stoffel, H. Strobelt, O. Deussen, D.A. Keim, Document Thumbnails with Variable Text Scaling, Computer Graphics Forum, vol. 31, Wiley Online Library, 2012, pp. 1165–1173,https://doi.org/10.1111/j.1467-8659.2012.03109.x.
[31] N. Cao, W. Cui, Introduction to Text Visualization, https://doi.org/10.2991/978-94-6239-186-4, (2016).
[32] X. Fei, X. Wu, X. Hu, Keyphrase extraction based on semantic relatedness, International Conference on Cognitive Informatics, IEEE, 2010, pp. 308–312 https://doi.org/10.1109/COGINF.2010.5599721.
[33] G. Ercan, I. Cicekli, Using lexical chains for keyword extraction, Inf. Process. Manag. 43 (6) (2007) 1705–1714 https://doi.org/10.1016/j.ipm.2007.01.015.
[34] J. Yang, E. Kim, M. Hur, S. Cho, M. Han, I. Seo, Knowledge extraction and visualization of digital design process, Expert Syst. Appl. 92 (2018) 206–215 https://doi.org/10.1016/j.eswa.2017.09.002.
[35] R. Glauber, D.B. Claro, A systematic mapping study on open information extraction, Expert Syst. Appl. 112 (2018) 372–387,https://doi.org/10.1016/j.eswa.2018.06.046.
[36] L.M. Bourne, P. Weaver, The origins of schedule management: the concepts used in planning, allocating, visualizing and managing time in a project, Front. Eng. Manag. 5 (2) (2018) 150–166,https://doi.org/10.15302/J-FEM-2018012.
[37] C. Chen, Improved TF-IDF in big news retrieval: an empirical study, Pattern Recogn. Lett. 93 (2017) 113–122 https://doi.org/10.1016/j.patrec.2016.11.004.
[38] X. Gao, M.P. Singh, P. Mehra, Mining business contracts for service exceptions, IEEE Trans. Serv. Comput. 5 (3) (2012) 333–344 https://doi.org/10.1109/TSC.2011.1.
[39] P. Sun, L. Wang, Q. Xia, The Keyword Extraction of Chinese Medical Web Page Based on WF-TF-IDF Algorithm, 2017 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), IEEE, 2017, pp. 193–198,https://doi.org/10.1109/CyberC.2017.40.
[40] G. Berend, R. Farkas, Sztergak: feature engineering for keyphrase extraction, Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2010, pp. 186–189 https://dl.acm.org/citation.cfm?id=1859704.
[41] K. Barker, N. Cornacchia, Using noun phrase heads to extract document keyphrases, Biennial conference of the canadian society on computational studies of intelligence: advances in artificial intelligence, (2000), pp. 40–52, https://doi.org/10.15302/J-FEM-2018012.
[42] D. Huang, S. Wang, F. Ren, Creating Chinese-English comparable corpora, IEICE Trans. Inf. Syst. 96 (8) (2013) 1853–1861, https://doi.org/10.1587/transinf.E96.D.1853.
[43] Z.K. Li J, Keyword extraction based on TF/IDF for Chinese news document, Wuhan Univ. J. Nat. Sci. 12 (5) (2007) 917–921,https://doi.org/10.1007/s11859-007-0038-4.
[44] M. Akiyoshi, Knowledge sharing over the network, Thin Solid Films 517 (4) (2008) 1512–1514 https://doi.org/10.1016/j.tsf.2008.09.042.
[45] Y. Lin, H. Lee, Developing project communities of practice-based knowledge management system in construction, Autom. Constr. 22 (2012) 422–432 https://doi.org/10.1016/j.autcon.2011.10.004.
[46] Z. Botao, X. Xuejiao, L. Peter, W. Xu, L. Hanbin, Convolutional neural network: Deep learning-based classification of building quality problems, Advanced Engineering Informatics 40 (2019) 46–57, https://doi.org/10.1016/j.aei.2019.02.009.