Automation in Construction 31 (2013) 65–74

Contents lists available at SciVerse ScienceDirect

Automation in Construction

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

Content-based text mining technique for retrieval of CAD documents

Wen-der Yu *, Jia-yang Hsu

Department of Construction Management, Chung Hua University, Hsinchu 300, Taiwan, ROC

ARTICLE INFO

Article history:
Accepted 25 November 2012
Available online 25 December 2012

Keywords:
CAD
Text mining
Information retrieval
Characteristic document
Construction engineering

ABSTRACT

The computer aided design (CAD) document provides an effective communication medium, a legal contract document, and a reusable design case for a construction project. Due to technological advancements in CAD industry, the volume of CAD documents has been increased dramatically in the database of construction organizations. Traditional retrieval methods relied on textual naming and indexing schemes that require the designers (engineers and architects) to memorize in details the meta-information used to characterize the drawings. Such approaches easily overwhelmed the users' memory capability and thus caused low reusability of CAD documents. In this paper, a content-based text mining technique is adopted to extract the textual content of a CAD document into a characteristic document (CD), which can be retrieved with similarity matching using a Vector Space Model (VSM), so that the automated and expedited retrievals of CAD documents from vast CAD databases become possible. A prototype system, namely Content-based CAD document Retrieval System (CCRS), is developed to implement the proposed method. After preliminary testing with a CAD database with 2094 Chinese annotated CAD drawings collected from two real-world construction projects and a public engineering drawing database, the proposed CCRS is proven to retrieve all relevant CAD documents with relatively high precision when appropriate query is specified. Finally, three search strategies are recommended for the users to narrow down search scope while a target CAD document is desired. It is concluded that the proposed content-based text mining approach provides a promising solution to improve the current difficulty encountered in retrieval and reusability of vast CAD documents for the construction industry.

1. Introduction

The advancement and widespread application of information technologies have been generating vast amount of electronic engineering documents. Among those, the computer aided design (CAD) documents may be the top-ranked in terms of its quantity. The CAD documents are generated by engineers or architects while performing engineering tasks, e.g., conceptual planning, basic design, detailed design, and construction supervision. The quantity of CAD documents being generated usually depends on the type and size of the project. A typical five-floor residential building may require less than 100 CAD drawings. However, a sophisticated mass-transportation project may generate more than 200,000 CAD documents. According to the statistics of the Department of Rapid Transit Systems of Taipei City (http://www.dorts.gov.tw/), the Mass Rapid Transit (MRT) project of Taipei City has generated totally 547,650 CAD drawings (including 235,095 design drawings and 312,555 as-built drawings) up to March 2010 [28] after its commencement in 1979. Due to massive growth of CAD documents, construction organizations are facing with an increasing management costs needed both for storage and retrieval of the electrical CAD documents.

The importance of the CAD document can be viewed from three aspects: (1) it provides an effective communication medium to illustrate the design concept of an engineering product, so that engineers and architects can “visualize” their ideas; (2) it is a legal document that provides a basis for performing, management, and closure of a contract; and (3) it provides a useful library for engineers and architects to reuse previous design models in order to accomplish their design efficiently. In construction practice, the cost items not included in the CAD documents are considered extra work that needs to be tackled with change orders. Moreover, when integrated with construction schedule, the CAD documents provide further help to the construction planner for progress control and dynamic resource allocation [29].

Due to its importance and vast amount, both the client and contractor of a construction project have been devoted in developing approaches for efficient and effective management of CAD documents. Unfortunately, nowadays retrieving CAD drawings is still a slow, complex, and error-prone work; it requires either exhaustive visual examination or a solid memory or both of the designers [12]. Previous efforts mainly focused on creating information management systems for the firm and the project [14]. Such efforts use textual databases to organize the information. Drawings are classified by keywords and additional information, e.g., designer's name, style, date and a textual description. Such kind of Project Management Information System

* Corresponding author. Tel.: +886 3 5186748; fax: +886 3 5370517.
E-mail address: wenderyu@chu.edu.tw (W. Yu).

0926-5805/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.autcon.2012.11.037

(PMIS) approach can be easily overwhelmed by the fast increasing quantity of CAD documents. Although automated document classification systems have been developed [4,5], they were developed for text-based documents, such as plans and final reports. The problem with CAD documents remained unsolved.

Fonseca and Jorge have pointed out in their work [12] that the traditional textual-indexing based approaches for CAD retrievals are not satisfactory due to two problems: (1) they force the designers to memorize in detail the meta-information used to characterize drawings; and (2) they require humans to produce it. As a result, two features need to be provided in an improved method: (1) the free or unstructured form of query—so that the user can use keywords and free-form natural language to search the relevant drawings; and (2) the automated generation of indexing database—in order to process vast amount of CAD documents efficiently.

In response to the above appeals, the present research proposes a content-based text mining technique that extracts the textual characteristic content from the CAD document and stores it into an indexing database; then, a Vector Space Model (VSM) is employed to represent the characteristic content of the CAD document. With the extracted characteristic contents, similarity matching can be conducted between the query description and the indexed CAD documents to identify the most relevant CAD documents for the query. Finally, the most relevant CAD documents are retrieved from the indexed database. With such a content-based text mining technique, the retrieval of CAD document can be expedited and automated.

The rest of the paper is organized as follows: related work is reviewed in Section 2; the core text mining technology adopted in this research is revisited in Section 3; the system architecture and associated algorithms of the proposed method is described in Section 4; in Section 5, a real world CAD database is selected as case study to test the *Recall* and *Precision* performance indexes of the proposed method; in Section 6, five different searching strategies are tested and the optimum strategies are identified and recommended; issues with the prerequirements and limitations for the proposed method are discussed in Section 7. Finally, conclusions and future recommendations are addressed in Section 8.

2. Literature reviews

This section reviews the work in several fields relevant to the present study, including construction management information system, automated classification of construction documents, and retrieval of CAD documents.

2.1. Construction management information system

The total amount of engineering documents related to a building construction project may be over 10,000 [14]. For a construction firm, the total amount of construction documents can easily exceed 56,000 [10,15]. Such documents include drawings, design specifications, schedules, quantity control reports, and other forms. Construction and project management information systems have been developed to manage such documents effectively.

Early efforts have been devoted to this area in 1990s. For example, the Constructability Lessons Learned Database (CLLD) and Integrated Knowledge-Intensive System (IKIS) were developed by Kartam and his team [16,17]. CLLD & IKIS provided an interactive computerized method for collecting, storing, and making constructability knowledge available. The Architecture and Engineering Performance Information Center (AEPIC) was developed in Maryland University, which focused on the collection, collation, study, and analysis of information relating to structural/functional failure of buildings, civil structures, and constructed facilities [19]. Bechtel On-Line Reference Library (OLRL) was developed by Bechtel, Inc., to reduce the amount of time the employees spent physically searching through reference materials [1], which stored a wide variety of computer generated documents thereby easing their retrieval for employee use, reuse, and modification. Civil Engineering Information System (CEIS) was developed by the Kajima Corp. [17] to provide a storage of information related to advanced construction technologies, construction know-how, construction planning, records of completed projects, and project costs.

Later in 2000s, Hajjar and AbouRizk [14] proposed an integrated approach to facilitate document search based on the concept of specialized construction data models. Their method required that designers first created document templates and was proved to be able to reduce the time and effort for document search on the project- and company-wide bases. Cao et al. [6] proposed an integrated method that combines the On-Line Analysis Processing (OLAP) of the data warehouse system and decision support system, namely Construction Management Decision Support System (CMDSS). Their method separated the analysis database (data warehouse) from the operational database, which renders the decision support process much faster. The adoption of OLAP transforms the data in a relational database into multidimensional cubes that could be observed from all perspectives.

The abovementioned construction management information systems (and their like, e.g., PMIS) are information retrieval systems that provide a database storing construction documents and a searching or indexing system to facilitate information retrievals. The users of the systems should be aware of the subdirectories (usually are also indexed) in which the files are stored. As a result, the naming system of the stored files may affect the search efficiency of the system significantly. As the number of the files grows, the naming/indexing system will become too complicated for the users to memorize. As a result, it becomes less efficient for the designers to retrieve relevant files.

2.2. Automated classification of construction documents

As a large amount of construction project information is exchanged using text documents, including contracts, change orders, field reports, requests for information, and meeting minutes [5], many previously automatic classification systems of construction documents were developed based on text mining techniques ([3–5,26]). Soibelman et al. [26] addressed that managing construction documents with the model-based information systems, such as Industry Foundation Classes (IFC) based building information models (BIMs) was challenging due to difficulties in establishing relations between the documents and project model (CAD) objects. Manually building of the desired connections is impractical since the PMIS typically store thousands of text documents and CAD drawings. Although search engines may be employed, limitations exist due to: (1) multiple words share the same meaning; (2) words have multiple meanings; and (3) relevant documents do not contain the user-defined search terms. Above all, the text information for the content of the CAD files should exist in the first place, and it needs to be generated manually in the traditional management information system.

In order to improve the drawbacks of the traditional management information systems, Caldas et al. [5] proposed a Construction Document Classification System (CDCS) to classify text-type construction documents. In their system, couple of machine learning techniques were adopted including support vector machine (SVM), Rocchio algorithm, naive Bayes, *k*-nearest neighbor, and IBM Miner (by International Business Machine, IBM) for Text. Up to 91.12% of correct retrieval was achieved. A similar work was reported by the same authors to build an automatic hierarchical classification of construction project documents according to project components [4] with similar classification accuracy.

Brilakis and Soibelman [3] developed a content-based search engine for image files. Soibelman et al. [26] further extended it to unstructured construction data types, such as text documents, site images, web pages, and project schedules. The abovementioned systems show a direction of automated classification for construction documents. However, the CAD documents were not tackled; one critical reason may be due to the lack of a tool for content information extraction from the CAD electronic documents.

2.3. Content-based retrieval of CAD documents

Traditional CAD systems rely mainly on conventional database queries and direct manipulation to retrieve relevant files [12]. Recently, there have been more and more interest in querying multimedia databases by content. Most of the work has focused on image databases [7]. Some of them use color and texture as main features to describe image content [24]. The CAD data, however, are stored in structured form (vector graphics), which needs different approaches from the image-based (color, texture) methods.

An early attempt to retrieving CAD data was developed by Mark Gross and Ellen Do in the context of the Electronic Cocktail Napkin [13] that addressed a visual retrieval scheme based on diagrams to indexing databases of architectural drawings. Users draw sketches of buildings, which are compared with annotations (diagrams), stored in a database and manually produced by users. Even though this system works well for small sets of drawings, the lack of automatic indexing and classification makes it impossible to be used for a large collection of drawings.

Berchtold and Kriegel [2] proposed an S3 system to support the management and retrieval of industrial CAD parts, described using polygons and thematic attributes. It retrieves parts using bi-dimensional contours drawn using a graphical editor or sample parts stored in a database. Although the preliminary results were found good for S3, it relied heavily on matching contours and ignores the spatial relationships and shape information. As a result, it is unsuitable for retrieving complex multi-shape drawings.

The third approach for content-based CAD retrieval was proposed by Park and Um [21] based on the dominant shape, where the objects were described by recursively decomposing its shape into a dominant shape, auxiliary components and their spatial relationships. The drawback of Park and Um's method is that it contains only a small set of geometric primitives and the less efficient matching algorithm. As a result, it is hard to work with large databases of vast drawings.

Fonseca and Jorge [12] proposed a content-based CAD retrieval system based on the spatial relationships and dominant shapes. Their system aims at achieving automatic simplification, classification and indexing of existing drawings, to make retrieval process more effective and accurate. Although the content-based CAD retrieval approach performs well for CAD retrieval, the users are required to know the spatial relationships and dominant shapes contained in the files. Such requirements may not be useful for engineers or architects in construction or civil engineering, since they are more familiar with the domain terminology or engineering specifications that annotate the drawing objects, even though the concept of “content-based” of the abovementioned methods provides a promising direction to solve the problem with the traditional naming system for indexing the CAD documents.

3. Text mining techniques revisited

In this section, the core text mining techniques adopted in this paper are revisited to provide required theoretical backgrounds for the proposed system described in the next section.

3.1. Text mining

Text mining (TM), also known as Knowledge Discovery from Text (KDT) or Document Information Mining, is a process to discover the implicit and useful information and knowledge stored in the documents [11,27]. The KDT process usually employs techniques such as Information Retrieval (IR), Information Extraction (IE), Computational Linguistics, Natural Language Processing (NLP), Data Mining (DM), and Knowledge Representation. Each of the above techniques has formed a specific and quite matured domain of research. The main difference between traditional DM and KDT is that the former focuses on the structured data in the databases; while the latter tackles semi- or non-structured texts. Dörre et al. [9] addressed two difficulties in KDT: (1) the manual approach for characteristic analysis of mass documents was inefficient; and (2) the key attributes were uneasy to define as the dimensionality of text data is large. In effect, successful application of KDT requires an additional data preparation process compared with the traditional DM, which will be described in the following.

3.2. Vector Space Model (VSM)

A special technique of IE named Vector Space Model (VSM) is used to extract the characteristic information of CAD documents in this paper. VSM is an algebraic model for representing text documents as vectors of identifiers. The original VSM was first proposed by Salton et al. [25] and described as follows:

For a document d_i and all words $w_l \in W$ (W is the corpus), the frequency $f(w_l|d_i)$ or *probability* $p(w_l|d_i)$, which characterizes the probability of the considered keyword (w_l) in the given document (d_i), are calculated using Eq. (1).

$$p(w_l|d_i) = \frac{f(w_l|d_i)}{\sum_m f(w_m|d_i),} \tag{1}$$

where the numerator in the left-hand side counts the frequency of keyword w_l occurring in document d_i, while the denominator calculates all keywords in d_i.

Using a vector space of L ($=||W||$) dimensions, a document d_i is given as a vector $\vec{x}_i$ of word *probabilities* as the following Eq. (2):

$$\vec{x}_i = [p(w_1|d_i), \ldots, p(w_L|d_i)]^T. \tag{2}$$

Due to the large amount of distinct words in any non-trivial corpus ($L \approx 10^5$–10^7) the vector space is extremely high dimensional but sparsely occupied [23]. Such highly sparse vectors need further scheme to improve the processing efficiency.

3.3. Corpus-based VSM

Using word stems to represent documents usually results in the inappropriate fragmentation of multi-word concepts [20]. As a result, using pre-stored phrases instead of single keywords or word stems as the terms may produce a VSM that better represents the human recalling process. It also results in a more effective document retrieval. As a result, key terms instead of keywords are used in VSM for information extraction in this research. Such an approach is called Corpus-based approach.

Kupiec et al. [18] proposed a primitive corpus-based approach based on Bayesian classifiers to enhance the document retrieval in VSM. Such an approach can significantly reduce the dimension of VSM and improve the efficiency of informational retrieval. The critical concept of Kupiec et al.'s corpus-based VSM is revisited in the following to provide background of the proposed method.

Assume that a query sentence s is considered to a test document S and F_1~F_k are critical features (i.e., key terms) used to characterize the test document, then the *probability* of s belonging to S can be calculated using Baye's rule by Eq. (3) [18].

$$\mathrm{P}(s \in S|F_1, F_2, \ldots, F_k) = \frac{\prod_{j=1}^{k} P\left(F_j|s \in S\right) P(s \in S)}{\prod_{j=1}^{k} P\left(F_j\right)}, \tag{3}$$

where $P(s \in S)$ stands for the probability of the sentence s belonging to S; $P(F_j)$ is the occurrence probability of key phrase F_j; and $P(F_j|s \in S)$ is the conditional probability of F_j occurring in S, given s belonging to S.

It is noted that Kupiec et al. had assumed the independence among the critical features, thus the occurrence probability of multiple key terms in the same document equals to the product of the occurrence probability of individual key terms. For a training corpus, $P(s\in S)$ is a constant, while $P(F_j|s\in S)$ and $P(F_j)$ can be estimated directly from the training set by counting the occurrences as shown in Eqs. (4)–(6).

$$P\left(F_j|s\in S\right)=\frac{\#\left(\text{sentence in document } S \text{ and has key term } F_j\right)}{\#(\text{sentence in document } S)}, \qquad (4)$$

where the numerator counts all sentences with key term F_j in the training corpus; the denominator counts all sentences in the training corpus.

$$P\left(F_j\right)=\frac{\#\left(\text{sentence in traing corpus and has key term } F_j\right)}{\#(\text{sentence in traing corpus})}, \qquad (5)$$

where the numerator counts all sentences with key phrase F_j in the training corpus; the denominator counts all sentences in the training corpus.

$$P(s\in S)=\frac{\#(\text{sentence in document } S)}{\#(\text{sentence in traing corpus})}, \qquad (6)$$

where the numerator counts all sentences in document S; the denominator counts all sentences in the training corpus.

In practical implementation, Eq. (4)–(6) are adopted to replace Eq. (3). As a result, numbers of word frequency in the test document S and that in the training corpus are calculated respectively to determine the *probability* of query s belonging to S.

4. Proposed Content-based CAD Retrieval System (CCRS)

In this section, the proposed system for the retrieval of CAD documents, namely Content-based CAD Retrieval System (CCRS) is described in details. The system architecture of CCRS is described first. It is followed by the computational procedure of the proposed CCRS. Then, the prototype implementation is demonstrated.

4.1. System architecture

The system architecture of the proposed CCRS is depicted in Fig. 1. Major components of CCRS are described as follows: (1) document base—an indexing database that stores the existing CAD documents; (2) Corpus—a database that stores the known key terms of the relevant problem domain, including the Chinese Knowledge and Information Processing (CKIP) corpus (containing 143,705 key terms) provided by the Academia Sinica of Taiwan [8] and a Problem Domain Keyword Base (namely PDKB) (containing 8222 key terms) provided by the construction organization generating the CAD documents; (3) textual content extraction module—to extract the textual information of the CAD drawing in form of a text file, namely characteristic document (CD); (4) characteristic extraction module—to extract the characteristic information from the CD in term of VSM; (5) VSM base—a database storing the VSMs of all existing CAD drawings stored in the Document base; (6) document matching module—perform matching calculations between the VSMs of the query description and that of the CAD drawings; (7) document query—an entry interface for user to input query description of the desired CAD document either in terms of key terms or a natural language sentence; and (8) retrieved CAD documents—output interface of the retrieved CAD documents.

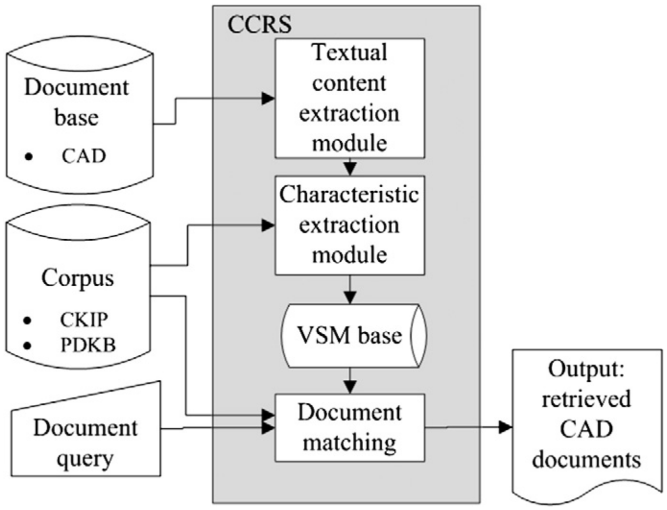

CCRS
Document base
• CAD
Textual content extraction module
Characteristic extraction module
Corpus
• CKIP
• PDKB
VSM base
Document query
Document matching
Output: retrieved CAD documents

Fig. 1. System architecture of CCRS.

4.2. Computational procedure

The computational procedure of the proposed CCRS consists of the following seven steps:

(1) The historical CAD drawing files are retrieved from CAD document base of the firm, see Fig. 5 for example.

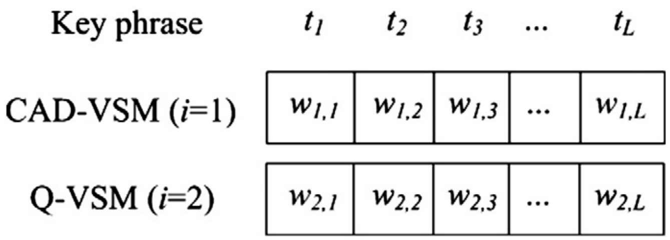
(2) The textual information of the CAD documents is extracted from the CAD database in the form of text files. In this step, each textual annotation in the CAD file is extracted as a single line in the text file (see Fig. 6 for example). The text file is named characteristic document (CD) for the CAD document, see Fig. 7 for example. Every CAD document is associated with a CD.

(3) The CDs are processed with the text mining techniques described in Section 3. First, each line in the CD is segmented into key terms stored in corpus—the maximum matching algorithm (MM) is adopted to pick the longest possible terms as suggested by Wu et al. [30]. Then, each CD is converted into a CAD-VSM. Assume that there are L key terms in the corpus, the elements of the CAD-VSM are the weights calculated by Eq. (7) and represented as shown in Fig. 2. The converted CAD-VSMs are stored in the VSM base.

$$w_{i,j}=\frac{L_j}{L_{i,\max}}\left(0.5+0.5\frac{tf_{i,j}}{tf_{i,\max}}\right)\times\log\left(\frac{N}{df_j}\right) \qquad (7)$$

where, $w_{i,j}$ is the weight of the jth key term associated with the ith document; L_j is the length of the jth key term; $L_{i,max}$ is the longest key phrase of the ith document; $tf_{i,j}$ is the term frequency of jth key phrase in the ith document; $tf_{i,max}$ is the highest term frequency of all key phrases in the ith document; df_j is the total number of documents with jth key phrase; and N is total number of documents in the database.

Assume that there are K documents in the database, a term-document weighting matrix (denoted as $\overrightarrow{W}=\left[w_{i,j}\right]$) can be



Key phrase	t_1	t_2	t_3	…	t_L
CAD-VSM (i=1)	$w_{1,1}$	$w_{1,2}$	$w_{1,3}$	…	$w_{1,L}$
Q-VSM (i=2)	$w_{2,1}$	$w_{2,2}$	$w_{2,3}$	…	$w_{2,L}$

Fig. 2. VSM representation.

constructed as shown in Fig. 3. In Fig. 3, every row represents the CAD-VSM for a specific CD.

(4) The query input provided by the user of CCRS is converted into a Q-VSM similar to the CAD-VSM as shown in Fig. 2, and is denoted as $\vec{q} = [\bar{w}_k] = [\bar{w}_1, \bar{w}_2, \cdots, \bar{w}_L]^T$.

(5) Document matching is performed by matching Q-VSM with all pre-stored CAD-VSM's in VSM base based on the inner product of $\vec{q}$ and W, as shown Eq. (8).

$$S_i = \sum_{j=1}^{L} \left(\bar{w}_j \times w_{ij} \right), \tag{8}$$

where S_i is the relative similarity between the query description and the ith document; $\bar{w}_j$ is the jth element in $\vec{q}$; $w_{i,j}$ is the jth element in the ith row of W.

Eq. (8) provides a similarity measure between the query and the pre-stored CAD documents. The higher value of S_i indicates the higher relevance between the query input and the ith document.

(6) The CAD-VSMs with the highest similarities are selected as the retrieved CAD documents and reported to the user with the output interface. Should the user be unsatisfied with the retrieved CAD files, he/she can revise the query description (or phrases) and conduct query again; otherwise, the retrieved CAD files are reported to the user as the retrieval result.

The flowchart of computational procedure of CCRS is depicted in Fig. 4. In order to improve the computation efficiency, the CAD-VSMs are preprocessed and stored in the VSM base before query is conducted. As a result, for each query only the query description (in form of natural language) or keywords need to be converted into Q-VSM. Then, the similarity matching is performed by vector computation that is done instantly within seconds.

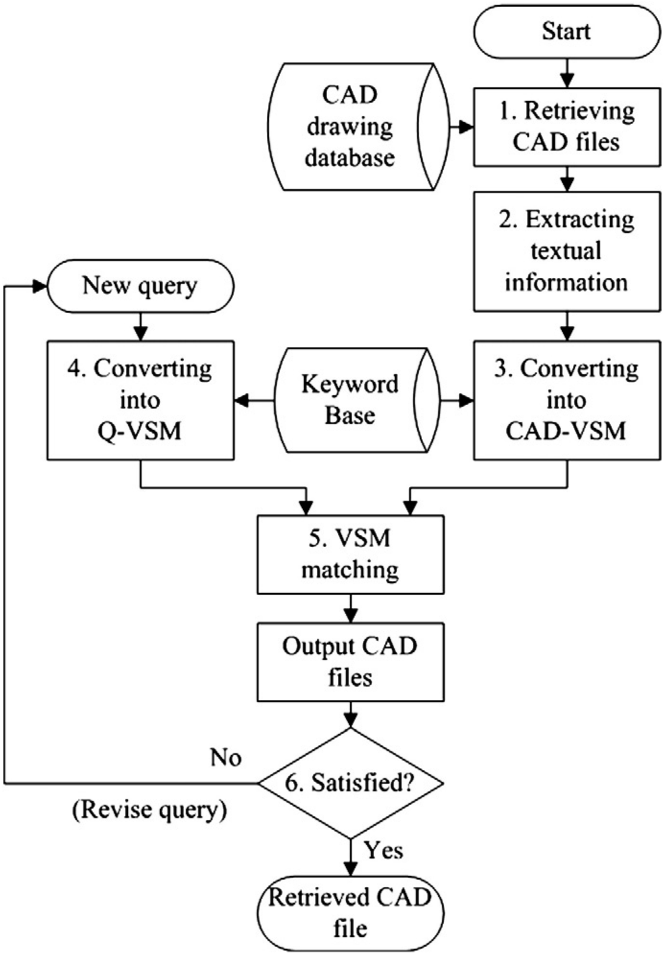


Fig. 4. Computational flowchart of CCRS.

4.3. Prototype implementation

The proposed CCRS and its computational algorithm have been implemented with a Visual Basic program. Fig. 5 shows the original CAD drawing annotated in Chinese language, where two annotations are indicated for examples: (1) "Length of shear wall" at the top; and (2) "Shear wall reinforcement profile" at the bottom. Fig. 6 shows the extracted characteristic document (CD) of the CAD drawing in Fig. 5, where the annotated examples are also indicated as individual lines in the extracted text. Fig. 7 shows the stored CD in Structured Query Language (SQL) database, where the CD of Fig. 6 is indicated. Figs. 8 and 9 show the query ("Shear wall" in Chinese) and the output interface of CCRS, respectively.

5. System testing

The proposed CCRS has been tested with a sample CAD database comprising 2094 CAD documents from three sources: (1) Source I—1331 CAD drawings collected from the "M2+M8" project, a design–build residential construction project conducted by the Department of Defense of Taiwan; and (2) Source II—629 CAD drawings collected from the "Villages 17, 18 & 19", a design–build residential construction project conducted by the Hsinchu City Government, Taiwan; and (3) Source III—134 civil work CAD drawings collected from the public infrastructure engineering drawings published by the Public Construction Council (PCC) of Taiwan [22]. Sources I & II are rehabilitation projects of two groups of old residential buildings for the households of Taiwan Army officers. A total 23 high-rise buildings of 10–14 floors were constructed. Total floor area was 169,096 m^2. The total design and construction budget for the two projects was nearly TWD$ 5,000,000,000 (USD$167 million). Source III is the public engineering drawing database provided by the Public Construction Council [22] of Taiwan, which consists of a set of standard drawings for the public infrastructure construction works.

$$\begin{array}{c|cccccc|}
 & Term_1 & Term_2 & \cdots & \cdots & \cdots & Term_N \\
Doc_1 & w_{11} & w_{12} & \cdots & \cdots & \cdots & w_{1N} \\
Doc_2 & w_{21} & w_{22} & \cdots & \cdots & \cdots & w_{2N} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
Doc_K & w_{K1} & w_{K2} & \cdots & \cdots & \cdots & w_{KN}
\end{array}$$

Fig. 3. Term-document weighting matrix.

5.1. Data classification

The CAD drawings collected from three sources were classified into three main drawing types, each type is comprised of several drawing categories: (1) Civil work drawing type (Civil)—consisting of river construction work drawings (RV), drainage drawings (DR), road work drawings (RD), road lighting drawings (RL), bridge drawings (BR), slope protection work drawings (SP), topographic site layout drawings (TS); (2) Architectural drawing type (Arch)—consisting of architectural design drawings (AD), construction drawings (CD); structural drawings (SD); landscape drawings (LS); and (3) Mechanical, electrical and plumbing drawing type (MEP)—consisting of fire protection drawings (FD), Plumbing drawings (PD), Data communication drawings (DC), Electrical drawings (ED), and other MEP drawings

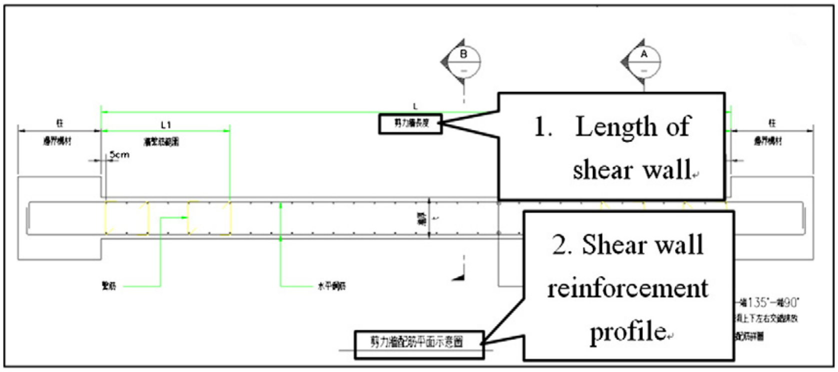


Fig. 5. Original CAD drawing.

(others). The numbers of drawings for each category according to the abovementioned classifications are shown in Table 1.

5.2. Experiment design

In order to verify the proposed CCRS, two performance indexes were measured for each sample: (1) *Recall*—measurement of the capability of a retrieval system to retrieve all relevant answers; and (2) *Precision*—measurement of the efficiency of a retrieval system to retrieve only the relevant answers.

The testing procedure consists of the following four steps:

Step 1 Key terms selection
A set of key terms are randomly selected from the CDs of the sampled CAD documents shown in Table 1. Since the CDs are extracted and stored in a SQL database as shown in Fig. 7, key terms from each CD are selected randomly as the query input of CCRS (as shown in Fig. 8). Finally, 114 CDs were sampled and tested.

Step 2 Document retrieval
CCRS retrieves a list of candidate CAD documents after query input, as shown in Fig. 9.

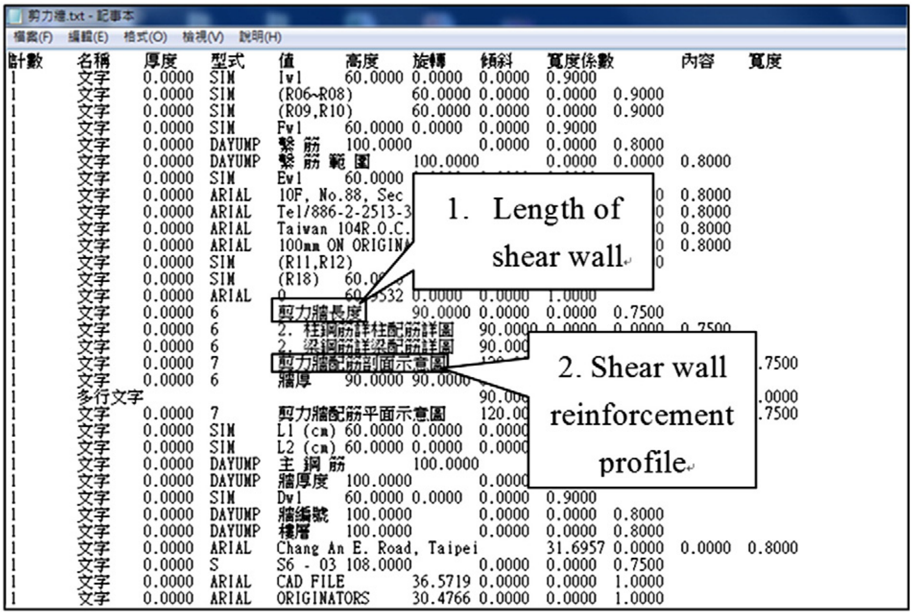
Step 3 Document review
The CD of each candidate CAD document of Step 2 is reviewed to determine the relevance with query input. In this research, this step is performed by converting the CD into a Microsoft (MS) Word file. Then, the key terms are matched by the built-in function of MS Word called by a VB program. If the match result is TRUE, it returns with the value “1”, otherwise it returns “0”.

Step 4 Relevance determination
The CAD document is determined to be “relevant” if the associated CD contains all key terms of query input; otherwise it is determined to be “irrelevant”.

Step 5 Frequency counting and index calculation
The relevant and irrelevant counts of the sampled documents are accumulated, and then the two performance indexes are calculated.

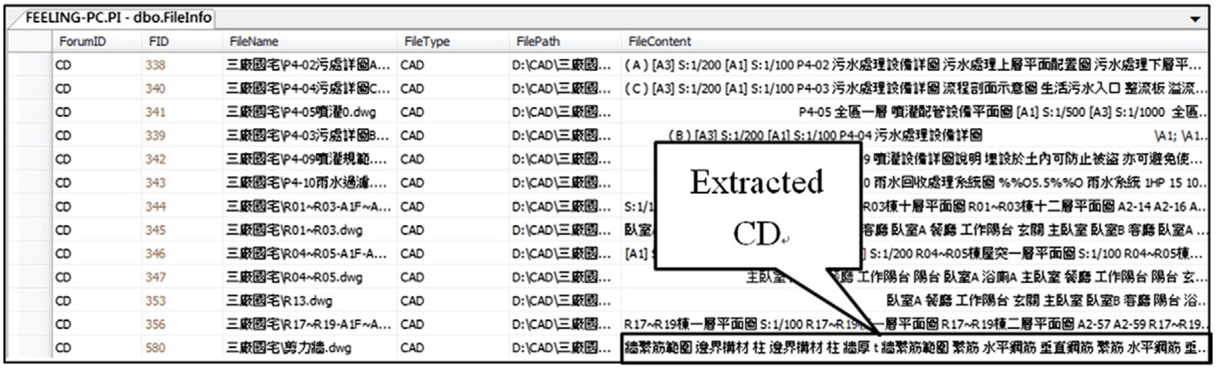
The indexes of *Recall* and *Precision* are calculated and defined in the following:

(1) Index of *Recall*
The index of *Recall*, denoted as $R(\%)$, is used to measure the capability of CCRS to retrieve all relevant files stored in the


剪力牆.txt - 記事本
檔案(F) 編輯(E) 格式(O) 檢視(V) 說明(H)
1. Length of shear wall
2. Shear wall reinforcement profile

```
計數   名稱    厚度    型式    值           高度     旋轉     傾斜     寬度係數          內容     寬度
1      文字    0.0000  SIM     Iv1          60.0000  0.0000   0.0000   0.9000
1      文字    0.0000  SIM     (R06~R08)             60.0000  0.0000   0.0000   0.9000
1      文字    0.0000  SIM     (R09,R10)             60.0000  0.0000   0.0000   0.9000
1      文字    0.0000  SIM     Fv1          60.0000  0.0000   0.0000   0.9000
1      文字    0.0000  DAYUMP  繫 筋        100.0000          0.0000   0.0000   0.8000
1      文字    0.0000  DAYUMP  繫 筋 範 圍           100.0000          0.0000   0.0000   0.8000
1      文字    0.0000  SIM     Ev1          60.0000
1      文字    0.0000  ARIAL   10F, No.88, Sec                                  0        0.8000
1      文字    0.0000  ARIAL   Tel/886-2-2513-3                                 0        0.8000
1      文字    0.0000  ARIAL   Taiwan 104R.O.C.                                 0        0.8000
1      文字    0.0000  ARIAL   100mm ON ORIGIN                                  0        0.8000
1      文字    0.0000  SIM     (R11,R12)                                        0
1      文字    0.0000  SIM     (R18)        60.0
1      文字    0.0000  ARIAL   0            60.9532  0.0000   0.0000   1.0000
1      文字    0.0000  6       剪力牆長度            90.0000  0.0000   0.0000   0.7500
1      文字    0.0000  6       2. 柱鋼筋詳柱配筋詳圖          90.0000  0.0000   0.0000   0.7500
1      文字    0.0000  6       2. 梁鋼筋詳梁配筋詳圖          90.000
1      文字    0.0000  7       剪力牆配筋剖面示意圖                                       .7500
1      文字    0.0000  6       牆厚         90.0000  90.0000
1      多行文字                                      90.000                               .0000
1      文字    0.0000  7       剪力牆配筋平面示意圖           120.00                       .7500
1      文字    0.0000  SIM     L1 (cm) 60.0000  0.0000   0.0000
1      文字    0.0000  SIM     L2 (cm) 60.0000  0.0000   0.0000
1      文字    0.0000  DAYUMP  主 鋼 筋              100.0000
1      文字    0.0000  DAYUMP  牆厚度       100.0000          0.0000
1      文字    0.0000  SIM     Dv1          60.0000  0.0000   0.0000   0.9000
1      文字    0.0000  DAYUMP  牆編號       100.0000          0.0000   0.0000   0.8000
1      文字    0.0000  DAYUMP  樓層         100.0000          0.0000   0.0000   0.8000
1      文字    0.0000  ARIAL   Chang An E. Road, Taipei           31.6957  0.0000   0.0000   0.8000
1      文字    0.0000  S       S6 - 03 108.0000                   0.0000   0.0000   0.7500
1      文字    0.0000  ARIAL   CAD FILE              36.5719  0.0000   0.0000   1.0000
1      文字    0.0000  ARIAL   ORIGINATORS           30.4766  0.0000   0.0000   1.0000
```

Fig. 6. Extracted characteristic document (CD).



FEELING-PC.PI - dbo.FileInfo

	ForumID	FID	FileName	FileType	FilePath	FileContent
	CD	338	三廠國宅\P4-02污處詳圖A...	CAD	D:\CAD\三廠國...	(A) [A3] S:1/200 [A1] S:1/100 P4-02 污水處理設備詳圖 污水處理上層平面配置圖 污水處理下層平...
	CD	340	三廠國宅\P4-04污處詳圖C...	CAD	D:\CAD\三廠國...	(C) [A3] S:1/200 [A1] S:1/100 P4-03 污水處理設備詳圖 流程剖面示意圖 生活污水入口 整流板 溢流...
	CD	341	三廠國宅\P4-05噴灌0.dwg	CAD	D:\CAD\三廠國...	P4-05 全區一層 噴灌配管設備平面圖 [A1] S:1/500 [A3] S:1/1000 全區..
	CD	339	三廠國宅\P4-03污處詳圖B...	CAD	D:\CAD\三廠國...	(B) [A3] S:1/200 [A1] S:1/100 P4-04 污水處理設備詳圖 \A1; \A1..
	CD	342	三廠國宅\P4-09噴灌規範....	CAD	D:\CAD\三廠國...	9 噴灌設備詳圖說明 埋設於土內可防止被盜 亦可避免使...
	CD	343	三廠國宅\P4-10雨水過濾....	CAD	D:\CAD\三廠國...	0 雨水回收處理系統圖 %%O5.5%%O 雨水系統 1HP 15 10..
	CD	344	三廠國宅\R01~R03-A1F~A...	CAD	D:\CAD\三廠國...	S:1/1 R03棟十層平面圖 R01~R03棟十二層平面圖 A2-14 A2-16 A..
	CD	345	三廠國宅\R01~R03.dwg	CAD	D:\CAD\三廠國...	臥室 客廳 臥室A 餐廳 工作陽台 玄關 主臥室 臥室B 客廳 臥室A ..
	CD	346	三廠國宅\R04~R05-A1F-A...	CAD	D:\CAD\三廠國...	[A1]] S:1/200 R04~R05棟屋突一層平面圖 S:1/100 R04~R05棟...
	CD	347	三廠國宅\R04~R05.dwg	CAD	D:\CAD\三廠國...	主臥室 廳 工作陽台 陽台 臥室A 浴廁A 主臥室 餐廳 工作陽台 陽台 玄...
	CD	353	三廠國宅\R13.dwg	CAD	D:\CAD\三廠國...	臥室A 餐廳 工作陽台 玄關 主臥室 臥室B 客廳 陽台 浴..
	CD	356	三廠國宅\R17~R19-A1F~A...	CAD	D:\CAD\三廠國...	R17~R19棟一層平面圖 S:1/100 R17~R19 一層平面圖 R17~R19棟二層平面圖 A2-57 A2-59 R17~R19..
	CD	580	三廠國宅\剪力牆.dwg	CAD	D:\CAD\三廠國...	牆繫筋範圍 邊界構材 柱 邊界構材 柱 牆厚 t 牆繫筋範圍 繫筋 水平鋼筋 垂直鋼筋 繫筋 水平鋼筋 垂..

Extracted CD

Fig. 7. CDs of the CAD files stored in SQL database.

database. It can be defined as the ratio of “the No. of retrieved relevant drawings” to “the No. of all relevant drawings in the database”. Such a measure is related to the searching completeness of the information retrieval system and is more important for the CCRS. In this paper, the index of *Recall* is calculated using Eq. (9).

$$R\,(\%) = \frac{\sum_{j=1}^{N} f_j}{\sum_{i=1}^{N} F_i} \tag{9}$$

where N is the total number of testing queries; $R(\%)$ is the average *Recall* of CCRS for the N testing queries; F_i is the total number of relevant CAD drawings stored in the database for the ith testing query; and f_j is the number of relevant CAD documents (containing all query terms) retrieved by CCRS for the jth testing query.

(2) Index of *Precision*

The *Precision*, $P(\%)$, of an information retrieval system is used to measure the percentage of the relevant documents in all retrieved documents. Such a measure is related to the search efficiency of an information retrieval system. While an information retrieval system is able to retrieve all relevant documents in the database (*i.e.*, *Recall* is high) and also includes many irrelevant files (*i.e.*, *Precision* is low), the user has to waste tremendous efforts in screening out irrelevant files out from the retrieval result. Such a system is considered inefficient and is of course not very useful for engineering applications. In this paper, the *Precision* is defined in Eq. (10).

$$P(\%) = \frac{\sum_{j=1}^{N} d_j}{\sum_{i=1}^{N} D_i}, \tag{10}$$

where N is the total number of testing queries; $P(\%)$ is the average *Precision* of CCRS for the N testing queries; D_j is the number of CAD documents retrieved by CCRS for the jth testing query; and d_j is the number of relevant CAD files retrieved by CCRS in the jth testing query.

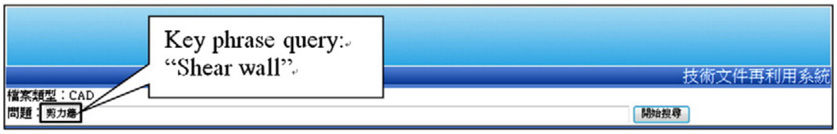
5.3. Testing results

The query tests were conducted following the experiment procedure described previously. A sample test is selected to explain the testing result of CCRS. Assume that an engineer would like to find a design drawing of “shear wall”. He/she specifies the key term for the query as “shear wall (剪力牆)”. The search results from CCRS for the query are shown in Fig. 9 and Table 2.

The testing results for 114 sampled CAD documents with randomly selected key terms are shown in Table 3. It is noted that no upper limit is set for the number of retrievals.

The search results show that the proposed CCRS is able to retrieve all relevant CAD documents stored in the database in terms of the query input. However, some irrelevant documents are found in the retrieval list, too. This is due to the limitation of text mining technique that segment the query description into terms by matching the pre-stored key phrases in the corpus. For example, the key term “shear wall (剪力牆)” can be segmented into three terms: “shear wall (剪力牆)”, “shear (剪力)”, and “wall (牆)”. Should an irrelevant document contain “shear (剪力)” and “wall (牆)” but no “shear wall (剪力牆)”, it will be still retrieved, too. Such a drawback can be eliminated by restricting the segmentation of CCRS by “maximum matching only algorithm” as described in Step 3 of the *Computational Procedure* of CCRS in Fig. 4. However, the proposed CCRS was developed for keyword query but for nature language query. As a result, the maximum matching algorithm [30] is adopted instead of maximum matching only algorithm.

The testing results of *Recall* and *Precision* reveal that the proposed CCRS is able to retrieve the relevant CAD drawings if correct key terms are queried. However, it is also found that some retrieved documents contain all the query key terms but may not meet the expectation of the user. Such a problem is due to two essential reasons addressed by Caldas et al. [5]: (1) multiple words share the same meaning; and (2) words have multiple meanings. Such problems are essential for all text mining methods. In order to improve such a problem, the searching strategy of CCRS is planned and discussed in the next section.



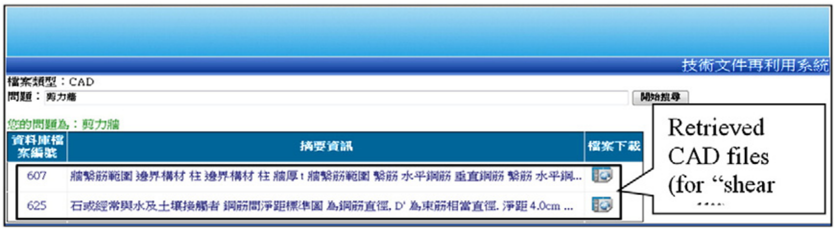
技術文件再利用系統

檔案類型：CAD

問題：剪力牆

Key phrase query:
“Shear wall”.

Fig. 8. CCRS-document query interface.



檔案類型：CAD
問題：剪力牆

您的問題為：剪力牆

資料庫檔案編號	摘要資訊	檔案下載
607	牆繫筋範圍 邊界構材 柱 邊界構材 柱 牆厚t 牆繫筋範圍 繫筋 水平鋼筋 垂直鋼筋 繫筋 水平鋼...	
625	石或經常與水及土壤接觸者 鋼筋間淨距標準圖 為鋼筋直徑, D' 為束筋相當直徑, 淨距 4.0cm ...	

Fig. 9. CCRS-retrieved CAD documents output interface.

6. Searching strategy planning

The objective of search strategy planning is to narrow down the scope of search result, so that the desired CAD document is ranked higher in the retrieval list. As addressed previously, the retrieved result may include many irrelevant CAD documents that contain the specified keywords but are not really desired by the user of CCRS. A better strategy is to define query descriptions more specifically to improve the *Precision*.

To measure the performance of query specification, a scenario is considered to assume that the user would like to retrieve a specific CAD drawing (namely “target document”) from the database. Since the retrieval result of CCRS is a list of candidate CAD documents ranked with similarity index values defined in Eq. (8), the search objective can be defined as “decrease the rank number of the target document in the retrieval list”.

6.1. Strategy planning

In order to achieve the objective, five searching strategies were planned and tested including: (1) Strategy I—query with two content key terms (terms that describe the content of the CAD drawing); (2) Strategy II—query with three content key terms; (3) Strategy III—query with one content key term and one drawing type term (the term classifying the type of drawing); (4) Strategy IV—query with two content key terms and one drawing type term; and (5) Strategy V—query with three content key terms and one drawing type term. The drawing type term is defined by the drawing type and the word “figure (圖)” in Chinese. For example, the drawing type of drainage structure will be a phrase comprised of “drainage structure (排水結構)” and “figure (圖)” in Chinese.

6.2. Searching results

The testing process of the five strategies refers to the following: (1) a specific sampled CAD document is selected; (2) content and drawing type key terms are recorded by reviewing the content of the document; and (3) query inputs for the five strategies are generated, and query tests are conducted. The search results of the five strategies for the 114 sample documents are shown in Table 4, where the average ranking of the target document in the retrieval list are shown numerical order; i.e., Ranking “1” means that the target document is ranked the first place in the retrieval list, and so forth. The testing result shows that the average ranks for the five different strategies are: (1) Strategy I—rank = 11.0; (2) Strategy II—rank = 6.8; (3) Strategy III—rank = 14.8; (4) Strategy IV—rank = 5.8; and (5) Strategy V—rank = 4.7. It is found that Strategy II (three key terms), Strategy IV (two content keywords and one drawing type keyword), and Strategy V (three content keywords and one drawing type keyword) are able to achieve relatively good results (with average ranking of less than 7). As a result, Strategies II, IV, and V are recommended for users of CCRS.

7. Discussions

The proposed content-based text mining technique and the associated CCRS have been verified as an effective tool for the retrieval of CAD documents in the previous sections. Some issues regarding practical application of CCRS are discussed in this section.

7.1. Requirements of textual content for CAD drawings

As described in Section 4, the proposed content-based text mining technique requires the textual information (stored as a CD in the database) to characterize the CAD document. Due to this requirement, the objects in the CAD document must be annotated with textual labels. Such textual labels are used in CAD documents to specify a component or to indicate the dimensional or material information of the drawing objects. Without such textual information, the proposed CCRS is unable to characterize the document. In fact, the more detailed annotation will help CCRS to characterize the CAD document more specifically. It should be noted that the requirement of textual content for CAD drawings also assumes correct spelling of the annotations; i.e., if the annotated is misspelled, it may not be retrieved correctly with the proposed method.

Table 1
Data classification for testing experiment.

Type[a]	Category[b]	No. of drawings	No. of sampled drawings	% of sampling
Civil	RV	23	2	8.7
	DR	23	2	7.4
	RD	27	2	7.4
	RL	7	1	14.3
	BR	30	3	10.0
	SP	24	2	8.3
	TS	14	1	7.1
Arch	AD	960	50	5.2
	CD	45	5	11.1
	SD	263	15	5.7
	LS	12	1	8.3
FEM	MD	316	16	6.0
	WE	48	4	4.4
	FD	39	3	7.7
	PD	36	3	8.3
	DC	35	3	8.6
	ED	189	10	5.3
	Others	3	1	33.3
Total	–	2094	114	5.4

[a] Civil—civil work drawings; Arch—architectural drawings; MEP—mechanical, electrical, and plumbing drawings.

[b] RV—river construction work; DR—drainage; RD—road work; RL—road lighting; BR—bridge; SP—slope protection; TS—topographic site layout; AD—arch/design; CD—construction; SD—structural; LS—landscape; MD—mechanical; WE—weak electric; FD—fire protection; PD—plumbing; DC—data communication; ED—electrical; others—other MEP drawings.

Table 2
Search results of sample query.

Key terms	No. of retrievals			Performance index	
	Total	Relevant	Irrelevant	*R*(%)	*P*(%)
“Steel structure (鋼結構)”	2	2	0	100	100

Table 3
Overall performance of 114 sample documents.

Average No. of retrievals			Performance index	
Total	Relevant	Irrelevant	*R(%)*	*P(%)*
10.8	3.7	7.1	100	57.2%

Moreover, the user is required to input relevant textual information (key terms) that may be contained in the CAD drawings. That is, the user (engineer or architect) of CCRS should be familiar with the language of CAD annotations. This usually does not cause a problem for engineers/architects of the same firm, since they are familiar with the annotation convention and the domain terminology of their work. It may, however, cause problems when the CAD database is imported from outside of the firm or from the other professional fields.

7.2. Requirement of domain corpus

The second requirement for the proposed CCRS is that there should be an appropriate domain corpus that contains the frequently used key terms (terminology) for Vector Space Model (VSM) transformation of the characteristic contents (i.e., CDs) of both the query input and the pre-stored CAD drawings. The lack of such corpus will result in poor document matching (Step 5 in Fig. 4) of the CCRS. Such a problem is especially true for the newly generated terms. There are two approaches for building such kind of domain corpus: (1) built by the domain experts—asking the domain experts to provide domain keywords directly; (2) extracting terms from the existing CAD documents—either automatic or manual process can be adopted to generate the domain key terms from existing CAD documents. Both of the abovementioned approaches can be employed to build the domain corpus.

7.3. Limitations with text mining techniques

The proposed CCRS adopts text mining techniques. As a result, limitations on the text mining techniques are also applicable to CCRS. Caldas et al. [5] addressed three limitations for a text-mining based construction document classification system: (1) multiple words share the same meaning; (2) words have multiple meanings; and (3) relevant documents do not contain the user-defined search terms. Such limitations are also found in the proposed CCRS. The establishment of domain corpus may moderate all of the above three problems. The frequently used terminology provides a "common language" for the architects and engineers, so that the ambiguity existing in the meanings of the frequently used terms can be improved. By updating the corpus, newly generated terms can be included to enrich the corpus, so that user-defined search terms have a better chance to be contained in the CAD documents.

Table 4
Search results of different strategies.

Type[a]	Category[b]	No. of files in database	No. of files sampled	Average ranking by different strategies[c]				
				I	II	III	IV	V
Civil	RV	23	2	3	3	2	1	1
	DR	23	2	1	1	1	1	1
	RD	27	2	21	12	22	13	3
	RL	7	1	1	1	1	1	1
	BR	30	3	5	4	22	2	1
	SP	24	2	3	1	6	2	1
	TS	14	1	4	4	19	5	4
Arch	AD	960	50	15	8	23	6	5
	CD	45	5	4	3	1	2	2
	SD	263	15	8	7	13	7	5
	LS	12	1	10	1	1	1	1
F&M	MD	316	16	75	42	51	31	36
	WE	48	4	12	7	9	3	2
	FD	39	3	5	4	8	4	4
	PD	36	3	4	3	12	3	2
	DC	35	3	17	15	63	13	10
	ED	189	10	7	6	12	8	4
	Others	3	1	3	1	1	1	1
Average	–	–	–	11.0	6.8	14.8	5.8	4.7

[a] Annotated as in Table 1.
[b] Annotated as in Table 1.
[c] I—2 content keywords; II—3 content keywords; III—1 content keyword + drawing type; IV—2 content keywords + drawing type; V—3 content keywords + drawing type.

The synonymy and misspelling problems described in Section 7.1 will also affect the performance of CCRS since the proposed CCRS relies on correct textual information both for the stored CAD drawings and the query input. The Problem Domain Keyword Base (PDKB) approach proposed in Section 7.2 will also improve the abovementioned synonymy and misspelling problems. In this research, a PDKB with 8222 key terms of problem domain are provided to enrich the corpus of CCR. When the users are more familiar with the common language used by the organization, they will find it easier to retrieve the desired CAD drawings.

8. Conclusions and recommendations

This paper presents a novel approach, namely Content-based CAD drawing Retrieval System (CCRS), for retrieval of CAD documents from vast CAD databases that could not be retrieved efficiently with the traditional textual naming and indexing systems. The proposed CCRS is based on the textual characteristic content of the CAD document and a text mining technique called Corpus-based VSM. A prototype CCRS has been developed with Visual Basic to implement the proposed method.

A real world drawing database with 2094 Chinese annotated CAD documents collected from two design–build residential construction projects and a public infrastructure engineering drawing database was selected for case study to verify the feasibility of the proposed CCRS. From the preliminary testing result, it is found that the proposed CCRS is able to retrieve all relevant CAD documents with relatively high precision when appropriate query is specified.

In order to improve the retrieval accuracy, five searching strategies are planned and tested for retrieval of a target CAD document from the database. Considering the rank of the target document in the retrieval list, it is found that Strategy II (query with three content key terms), Strategy IV (query with two content key terms and one drawing type term), and Strategy V (query with three content key terms and one drawing type term) are able to achieve relatively good results (with an average rank less than 7). That is, the users are assured to obtain the desired CAD document by screening the first 7 retrieved documents. Such a result has significantly improved the CAD retrieval practice in current CAD document management systems. It is thus concluded that the proposed content-based CAD retrieval method provides a promising solution to the current difficulty for the effective and efficient retrievals of CAD documents. With such a tool, the engineers/architects are equipped with a more powerful tool to retrieve reusable design case in their designing process; otherwise it was almost impossible to reuse the thousands of hundred CAD drawings stored in the database of the firm.

Although the primitive testing results show that the proposed content-based text mining technique and the associated CCRS can be a promising tool for the retrieval of Chinese annotated CAD drawings, two pre-requirements need to be fulfilled before practical application, including: (1) the CAD documents should contain textual information; and (2) there should be a domain corpus containing domain key terms. Moreover, the limitations of text mining techniques are also applicable to the proposed method, e.g., synonymy and misspelling problems. It is recommended that an appropriate and updated problem domain corpus can moderate the limitations of the adopted text mining techniques.

The proposed CCRS has been verified with the selected Chinese annotated CAD documents. Similar applications can be extended to other databases of CAD documents annotated in other languages (such as English), too.

Acknowledgments

The valuable case study information presented in this paper was provided by the Hsinchu City Government and the Public Construction Commission, Taiwan. The authors would like to express sincere appreciations to the contributors.

References

[1] Bechtel, Bechtel On Line Reference Manual, Bechtel Corporation, Gaithersburg, MD, USA, 1994.

[2] S. Berchtold, H.-P. Kriegel, S3: similarity in CAD database systems, in: Proceedings of the International Conference on Management of Data (SIGMOD'97), Tuscon, AZ, USA, 1997.

[3] I. Brilakis, L. Soibelman, Content-based search engines for construction image databases, Automation in Construction 14 (4) (2005) 537–550.

[4] C.H. Caldas, L. Soibelman, Automating hierarchical document classification for construction management information systems, Automation in Construction 12 (4) (2003) 395–406.

[5] C.H. Caldas, L. Soibelman, J. Han, Automated classification of construction project documents, Journal of Computing in Civil Engineering 16 (4) (2002) 234–243.

[6] Y. Cao, K.W. Chau, M. Anson, J.P. Zhang, An intelligent decision support system in construction management by data warehousing technique, Lecture Notes in Computer Science 2480 (2002) 360–369.

[7] S.K. Chang, B. Perry, A. Rosenfeld, Content-based Multimedia Information Access, Kluwer Academic Publishers, Norwell, MA, USA, 1999.

[8] CKIP, Chinese Knowledge and Information Processing. Webpage, Website: http://godel.iis.sinica.edu.tw/CKIP/engversion/index.htm Institute of Information Science, Academia Sinica, Taiwan, 2012. (visited 2012/2).

[9] J. Dörre, P. Gerstl, R. Seiffert, Text mining: finding nuggets in mountains of textual data, in: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 1999), Vol. 1, ACM, NY, USA, 1999, pp. 398–401.

[10] C.M. Eastman, P. Teicholz, R. Sacks, K. Liston, BIM Handbook: a Guide to Building Information Modeling for Owners, Managers, Architects, Engineers, Contractors, and Fabricators, in: 2nd ed., John Wiley and Sons, Hoboken, NJ, USA, 2011, p. 3.

[11] R. Feldman, I. Dagan, Knowledge Discovery in Textual Databases (KDT), in: Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD 1995), Vol. 1, 1995, pp. 112–117.

[12] M.J. Fonseca, J.A. Jorge, Towards content-based retrieval of technical drawings through high-dimensional indexing, Computers and Graphics 27 (1) (2003) 61–69.

[13] M. Gross, E. Do, Demonstrating the electronic cocktail napkin: a paper-like interface for early design, in: Proceedings of the Conference on Human Factors in Computing Systems (CHI'96), Vancouver, Canada, ACM, NY, USA, 1996, pp. 5–6.

[14] D. Hajjar, S.M. AbouRizk, Integrating document management with project and company data, Journal Computing in Civil Engineering, ASCE 14 (1) (2000) 70–77.

[15] C. Hendrickson, Project Management for Construction: Fundamental Concepts for Owners, Engineers, Architects and Builders, 2008. (Version 2.2, www.ce.cmu.edu/pmbook, visited 2011/11).

[16] N. Kartam, Knowledge-intensive database system for making effective use of construction lesson learned, in: Proceedings Computing in Civil Engineering (New York), Vol. 2, ASCE, New York, NY, USA, 1994, pp. 1139–1145.

[17] N. Kartam, I. Flood, Constructability feedback systems: issues and illustrative prototype, Journal of Performance of Constructed Facilities, ASCE 11 (4) (1997) 178–183.

[18] J. Kupiec, J. Pedersen, F. Chen, A trainable document summarizer, in: Proceedings of the 18th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, Vol. 1, 1995, pp. 68–73.

[19] J. Loss, AEPIC project: update, Journal of Performance of Constructed Facilities, ASCE 1 (1) (1987) 11–29.

[20] W. Mao, W.W. Chu, The phrase-based vector space model for automatic retrieval of free-text medical documents, Data & Knowledge Engineering 61 (1) (2006) 76–92.

[21] J. Park, B. Um, A new approach to similarity retrieval of 2D graphic objects based on dominant shapes, Pattern Recognition Letters 20 (6) (1999) 591–616.

[22] PCC, Basic Engineering Drawings for Public Construction Work. Website: http://pcces.archnowledge.com/csi/Default.aspx?FunID=Fun_10_52012, (visited 2012/06).

[23] D. Pullwitt, Integrating contextual information to enhance SOM-based text document clustering, Neural Networks 15 (8–9) (2002) 1099–1106.

[24] Y. Rui, T.S. Huang, S.-F. Chang, Image retrieval: past, present, and future, Journal of Visual Communication and Image Representation 10 (1) (1997) 1–23.

[25] G. Salton, A. Wang, C.S. Yang, A Vector Space Model for automatic indexing, Communications of the ACM 18 (11) (1975) 613–620.

[26] L. Soibelman, J.F. Wu, C. Caldas, I. Brilakis, K.Y. Lin, Management and analysis of unstructured construction data types, Advanced Engineering Informatics 22 (1) (2008) 15–27.

[27] D. Sullivan, Document Warehousing and Text Mining, Wiley Computer Publishing, New York, USA, 2001.

[28] L.M. Ting, Y.F. Hsiao, Technical document and knowledge management—a case study of Taipei MRT project, Web article, Public Construction Commission, Taiwan, 2012. (Website: www.pcc.gov.tw/epaper/9905/download/reader_1.doc, visited 2012/04, (in Chinese)).

[29] H.J. Wang, J.P. Zhang, K.W. Chau, M. Anson, 4D dynamic management for construction planning and resource utilization, Automation in Construction 13 (5) (2004) 575–589.

[30] J.W. Wu, Y.H. Tsou, C.K. Chiou, J.C.R. Tseng, Development of a ubiquitous virtual tutoring assistant system, in: Proceedings of the 2007 World Conference on Educational Multimedia, Hypermedia & Telecommunications (ED-MEDIA), Vancouver, Canada, June 25-June 29, 2007, 2007, 8 pp.