A Relativity Evaluation Approach to Unstructured Document Integration and Retrieval for Building Information Modeling

Yunyi Zhang[1], Jianping Zhang[2], Qiang Liu[3], Jiarui Lin[4]

1) Ph.D. Student, Department of Civil Engineering, Tsinghua University, Beijing, China. Email: yunyi2525@qq.com
2) Ph.D., Prof., Department of Civil Engineering, Tsinghua University, Beijing, China. Email: zhangjp@tsinghua.edu.cn
3) Ph.D. Candidate, Department of Civil Engineering, Tsinghua University, Beijing, China. Email: liuqiang10@mails.tsinghua.edu.cn
4) Ph.D. Candidate, Department of Civil Engineering, Tsinghua University, Beijing, China. Email: jiarui_lin@foxmail.com

Abstract:

A large amount of construction project data is stored in unstructured documents, which are difficult to handle with highly structured Building Information Models (BIMs). This paper aims at proposing a method for unstructured information integration and retrieval based on Industry Foundation Classes (IFC) standard, thus increasing the efficiency of information utilization.

This research firstly concentrated on mechanisms for evaluating relativity between text and BIM objects by their features. Document features can be represented by key terms extracted from its content/title based on text mining. Similar features of BIM objects can be extracted from their name, type, description, and properties. The correlation indexes are calculated and ranked by establishing a vector space based on the TF-IDF method with features extracted above.

When integrating documents and model entities, correlation methods based on title and content of documents are provided. Weights of the two indices are determined by fuzzy comprehensive evaluation method, capable of searching a recommended list of correlated entities according to selected documents. After users' confirmation, the relation between documents and model entities can be established based on IFC standard.

Furthermore, to improve document retrieval, related documents are analyzed and separated into subsections by subtitles, forming a semi-structured document. Once a model entity is selected, correlation indices of related subsections are calculated by the above-mentioned method, providing a ranked list of related subsections. It is flexible for the users to obtain required information effectively. Additional query information can be specified if needed.

Finally, an illustrative application shows that the correlation between documents and model entities can be established with the above-mentioned approach, and well-named BIM entities and statements of the documents will improve the accuracy of the method. Hence, unstructured information can be well integrated and retrieved based on the proposed method, thus increasing the efficiency of project information utilization.

Keywords: Relativity evaluation, BIM, Unstructured Information, Project Documents

1. INTRODUCTION

Information management is one of the core aims of Building Information Modeling (BIM) technology. BIM-based project management methods can be utilized to support information exchange between different stages, enable the collaborative work between participants in a project, and realize the predictive analyze for the whole life-cycle management of the project.

BIM utilizes a highly structured data set to present various features and properties of project data. It brings a lot of benefits, e.g. it improves the interoperability between different BIM-based software, simplifies the model viewing and data integration process.

However, one of the difficulties to handle project data is that they can be stored in various formats. Specifically, project data can be classified into: structured data, semi-structured data, unstructured text data, unstructured graphic data and unstructured multi-media data (Simoff & Maher, 1998).

Unstructured data may be partially substituted by structured ones with the trend of informatization, however, they are somewhat irreplaceable in a period of time. The reasons are that project data refer to extremely broad extensions, so it is impractical to represent all the data by enormous dimensions. Moreover, text and image processing techniques still need further study to handle various project data generated in the life-cycle of a construction project.

Information such as contracts, change orders, field reports, and requests for information, etc., is highly valuable for planning, implementation, control and analysis (Soibelman & Caldas, 2000). However, a large amount of such

information is stored in unstructured documents, which are difficult to handle with highly structured Building Information Models.

It has been realized that it cannot satisfy the need to manage complicated project only by manual management of documents. Corporations are actively establishing or developing information management systems in order to improve the efficiency of searching and retrieving data. However, most of the management systems are not based on models, they simply manage the model and documents separately, ignoring the relativity between them.

Model-based information management systems are among the frontier of study. Mao and Zhu (2007) proposed a metadata model for information integration with BIM model, focusing mainly on the specific RFI data format. Caldas et al (2008) proposed a methodology for the integration of project documents, focusing on classifying and relating text documents.

Most of the studies emphasize the process of recognizing the relativity between documents and model and retrieving related information while viewing model. This paper aims at providing a relativity evaluation approach to unstructured document integration and retrieval, solving the problem of separation between BIM models and unstructured information in project management, hence improving the efficiency of searching and retrieving unstructured documents.

2. EXTRACTING FEATURES OF DOCUMENTS AND MODEL ELEMENTS

2.1 Recognizing Features of Unstructured Documents

Extracting features in advance can avoid repetition of document processing and increasing the efficiency of document integration and retrieval, otherwise, all the documents have to be full-text scanned whenever a new query is conducted. As is mentioned above, unstructured documents can be classified in to unstructured text data, unstructured graphic data and unstructured multi-media data. Therefore different method should be implemented for various types of documents.

The title or name is often the general description of the document, and some information system even regulated the standard for naming files. Therefore, the title/name can be recognized as an important clue to represent the feature of the document. Graphic data and multi-media data are tedious to be processed, but their filename can be a good description of the information if they are well-named.

As for unstructured text documents, more information can be mined by utilizing NLP (natural language processing) techniques. Several steps have to be taken are as follows. First, documents are tokenized into single words, which is simple for Latinized languages such as English because there are spaces between words, tokens are then lemmatized into the stem of the word. However, for languages like Chinese or Japanese, tokenizing maybe a challenging job, but quite a few researches have turn out to be satisfying. Punctuations and stop-words which express no specific meaning are ought to be eliminated. Hence, terms and their frequency in each document are collected. As a matter of fact, a vector space model is established by features of documents. In the model, each term is a dimension of the vector space, and each document is a vector compose of term frequency in each dimension.

2.2 Organizing Unstructured Documents

This research focuses on a BIM-based information management system, so documents uploaded are given a GlobalID as its unique identification. In the file-database, a file table stores the records of all the files, and the records are consist of the basic properties of the documents. The title or name of the document is one of the most important field in the record because it is one of the document feature for calculation relativity with model elements.

As for the vector space model for unstructured text documents, a dictionary table and a document-term index table are needed for efficient query. In the dictionary table, each term is given a unique code and its total frequency in all documents are recorded. A term that exists in every documents with high frequency may be useless to characterize a document. Whenever a new document is uploaded, new terms should be added to the table, and the frequency of existed terms also needs to be updated. In the document-term index table, each record consists of a term, a document and the frequency of the term appearing in the document.

As can be noticed, all the features of unstructured documents are stored in tables, thus transforming into structured information, and can be queried efficiently by SQL or other structured query languages.

2.3 Recognizing Features of BIM Elements

BIM data are highly structured, and the IFC provides a data representing standard for BIM (buildingSMART, 2013), so retrieving features of BIM elements is much easier than unstructured documents. After selecting a specific model element, all its properties such as name, type, and description can be extracted, which can all be seen as the features of the element.

It should be noticed that many properties of the element are express in property sets related to the element. Therefore, the name, type, description and value of the elements related to it should be considered as well. Experiment shows that they are necessary, but may not be considered to have the same importance with the original properties. A deduction coefficient of 0.5 applied to related element features can achieve a relatively satisfying result.

3 RELATIVITY EVALUATION APPROACH

Based on different document formats and different document features, a title/name-based and a content based relativity evaluation approaches are provided. The correlation between document and model elements can be established after user's confirmation.

3.1 Title/Name-Based Relativity Evaluation Approach

As is analyzed above, if documents are well-named, the title of the document can be a fairly good clue for document integration and retrieval. Especially for those graphic or multi-media documents, the title maybe the most reliable clue because their contents are difficult to process.

The title/name of a document is often short because of the limitation of operation system, so they can be regarded as a short string. Two string matching algorithms are applied to judge the correlation between title/name and the feature of a model element, i.e. LD (Levenshtein Distance) and LCS (Longest Common Substring).

LD is the minimum number of operations required to transform one string into the other. LCS is just as what the name suggests, but the substring has to be continuous. These two algorithms are both quite commonly used in natural language processing, and they are both efficient in calculation. In order to combine their advantages, an integrated formula is established to describe the similarity between two strings.

$$sim_1 = \frac{LCS}{LD + LCS} \tag{1}$$

where *LCS*: Longest Common Substring.
LD: Levenshtein Distance.

It should be noticed that the similarity calculated by the formula above does not show the probability that a document is related to a model element, because the description in a document does not consist with that in the model element. However, a ranking list can be returned indicating the most likely relating model elements of a documents.

3.2 Content-Based Relativity Evaluation Approach

The TF-IDF method (Croft & Harper, 1979) which is commonly used in search engines is applied to analyze the correlation between a document and a model element based on document content. The main idea is quite simple, that a term should be given a higher weight if it appears more frequently in the document selected, but less frequently in other documents.

The vector space model established in previous steps is a preparation for this approach. Note the document set as D, for each document $d_i \in D$, an n-dimensional vector $\{w_{i1}, w_{i2}, ..., w_{in}\}$ can be its representative, where w_{ij} stands for the weight of term t_j in document d_i. In previous step, the weight are simply the frequency of the term in the document, but here we need to modify them into a term weight considering the length of the selected document and other documents.

The TF-IDF method consists of two parts. The first one is TF (Term Frequency), but we need to standardize the frequency considering the length of different documents. The formula is as follows:

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{2}$$

where n_{ij} : The frequency of term t_j in document d_i.

The second part of TF-IDF method is IDF (Inverse Document Frequency) representing the rareness of a term. If a term appears frequently in the document selected, but less frequently in others, it should be consider a character of the document. The formula is as follows:

$$idf_i = \log \frac{|D|}{\left|\{ j : t_i \in d_j \}\right|} \tag{3}$$

where $|D|$: the number of all documents

$\left|\{ j : t_i \in d_j \}\right|$: the number of documents that contains the term t_i

As can be seen, all variables used in the calculation can be easily queried in the tables, therefore, the organization of the documents is effective. The overall term weight is the product of TF and IDF parts.

Similarly, a vector representing a model element can be constructed by the same method. The similarity between a document and a model element can be represent by the cosine of the two vectors' angle.

$$sim_2 = \cos\theta_{ij} = \frac{d_i \cdot e_j}{\|d_i\| \times \|e_j\|} \tag{4}$$

where d_i : the vector representing document i.

e_j: the vector representing model element j.

3.3 Integrating the Two Relativity Evaluation Approach

As for unstructured text documents, both the two relativity evaluation approach can be implemented, so an overall relativity has to be calculated in order to find correlated model entities. This research utilizes fuzzy comprehensive evaluation method (Sadiq et al., 2004) to integrate the two approaches.

$$sim = \alpha \cdot sim_1 + \beta \cdot sim_2 \tag{5}$$

where α, β : the relative weights, $\alpha + \beta = 1, \alpha \geq 0, \beta \geq 0$

According to the fuzzy comprehensive evaluation method, the relative weights can be chosen as the coefficient of variance of the two approaches, i.e. $\alpha / \beta = V_1 / V_2$.

4. APPLICATION OF RELATIVITY EVALUATION APPROACH

The relativity evaluation approach proposed in this paper can be utilized both in the process of unstructured document integration and retrieval. An overview of the method is as Figure 1.

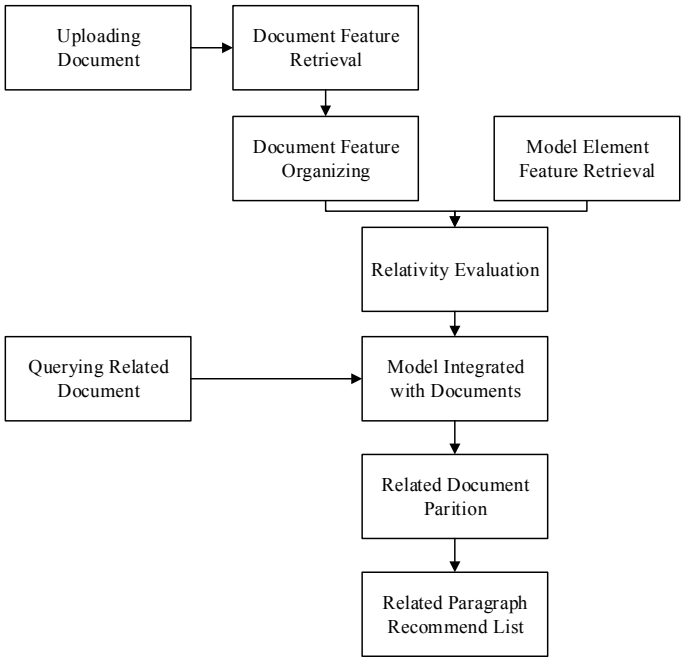

Figure 1. Overview of the relativity evaluation approach for integration and retrieval

4.1 Correlating Unstructured Documents with Model

By implementing the approaches above, the relativity between each unstructured documents and model elements can be calculated. After establishing the correlation between documents and model elements, the integration of unstructured documents can be realized.

Once a new document is uploaded, it would be analyzed and the features would be extracted. Features of model elements are easy to be extracted because they are in the form of structured data, so it is unnecessary to pre-process the model. Each model element would be traversed and a correlation index between the element and uploaded document would be calculated. Then a ranking list can be returned to the user, where the most related model elements are at the top. Users can easily select the related elements from the list or relate all elements whose correlation index is above a threshold.

After user's confirmation, the correlation can be recorded within the IFC framework. An associative IFC class – *IfcRelAssociatesClassification* can establish the correlation between documents and objects. This class provides an external link to the location of the related document.

4.2 Retrieving Related Document

After establishing the correlation between documents and model elements, the retrieval processing should be simple, just return the list of documents that are related to the element selected. However, sometimes the documents are quite long and may consist of several parts, in which each describes one kind of elements. Under this circumstance, one may want to locate the paragraph directly related to the model elements.

Documents are often separated into paragraphs. Some well-formed documents have sub-titles indicating the content of each part. The whole long document can be divided into several parts and each part can be regarded as a new document. Because the number of paragraphs in a document is limited, TF-IDF method can be implemented again to search the most correlated paragraphs.

5. RESULTS AND DISCUSSIONS

In order to test the feasibility of the approaches, an illustrative example is provided. The model is an office building provided as a free model for uncommercial use by BIM Vision (BIM Vision, 2015). This IFC file consists of 4906 rooted entities and 264450 resource entities. The model consists of slabs, beams, columns, walls and other architectural elements, which is a typical project model.

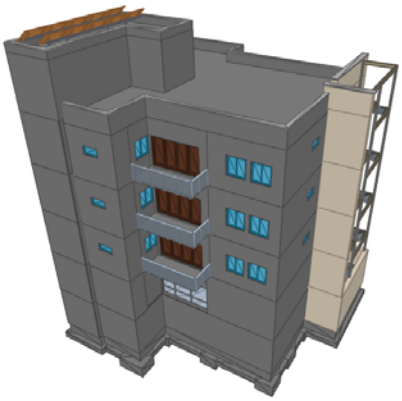

Figure 2. Sample BIM model

The document set consists of 10 different kinds of unstructured text information, and taking a RFI about the roof slab as an example to illustrate the result of correlation calculation. This document is apparently related to a specific model element, so the most ideal result is to return the roof slab element. Below is the result of this method.

GlobalID	Name	Type	Correlation
1Wz6OBb8v3zxlqKXFcScod	Roof Slab	IfcSlab	0.24394445772068812
0x5oiMr9H7GADVRP92cJqH	Interior Floor Slab	IfcSlab	0.22214670077634782
181iT06rz7dAlhfkhDUQxM	Interior Base Slab	IfcSlab	0.21943465379911947
2WZwJgS0D7LB1adieW9lXZ	Typical Ceiling Assembly	IfcSlab	0.21649286587621172

Figure 3. Recommend list of correlated model element of a RFI about the roof slab

This result shows that the related roof slab is at the top of the recommend list, thus this element can be selected from the list and the document can be integrated with the model element by establishing a correlation.

Further, we take a calculation report as an example to illustrate the retrieval procedure. The report is in fact related to almost all structural elements in the model, and the report is fairly long. Therefore, as an information manager, when he selects the roof slab, the most ideal result is that the specific paragraph of slab calculation can be located. Figure 4 is the result of this method, and as we can see, the paragraph describing the calculation parameters and processes of the load on the roof slab is correctly extracted.

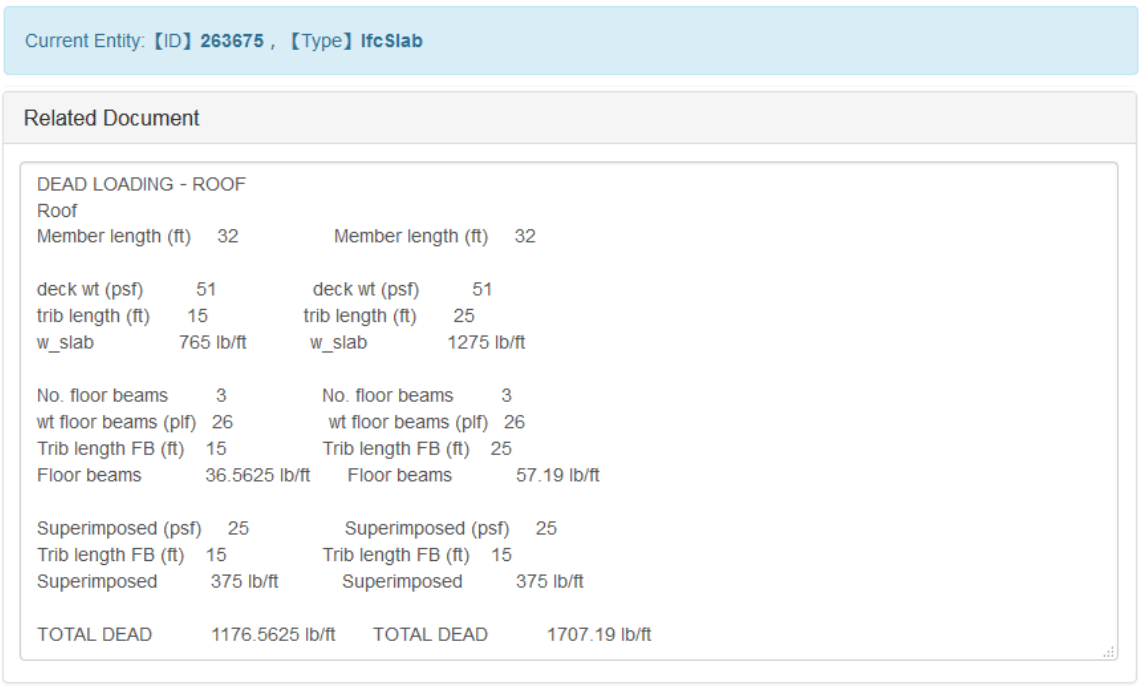

Figure 4. Paragraph retrieval result of a roof slab

6. CONCLUSIONS

This research proposed a relativity evaluation approach to unstructured document integration and retrieval for building information model. Results show that this approach is capable of relating documents to model elements and query correlated paragraphs if model elements are well named and the statements in the documents are consistent with models. Applying this approach to information management systems can improve the organization and retrieval efficiency of unstructured documents. Hence, unstructured information can be well integrated and retrieved based on the proposed method, thus increasing the efficiency of project information utilization.

Several improvements are still expected in further studies. For example, a threshold maybe preferable to reduce the length of recommend list, thus the integration process can be accomplished automatically, without users'

confirmation.

ACKNOWLEDGEMENT

This research was supported by the National High Technology Research and Development Program (863 Program) of China (No. 2013AA041307), the National Natural Science Foundation of China (No. 51278274), and the Tsinghua University-Glodon Joint Research Center for Building Information Model (RCBIM).

REFERENCES

BIM Vision. (2015). Retrieved from BIM Vision website: http://www.bimvision.eu/download/, accessed on November 24, 2015.

buildingSMART. (2013). *IFC4 Release Candidate 4*. Retrieved from buildingSMART website: http://www.buildingsmart-tech.org/ifc/IFC2x4/rc4/html/index.htm, accessed on November 24, 2015.

Croft, W. B., & Harper, D. J. (1979). Using probabilistic models of document retrieval without relevance information. Journal of documentation, 35(4), 285-295.

Mao, W., Zhu, Y., & Ahmad, I. (2007). Applying metadata models to unstructured content of construction documents: A view-based approach. Automation in construction, 16(2), 242-252.

Sadiq, R., Husain, T., Veitch, B., & Bose, N. (2004). Risk-based decision-making for drilling waste discharges using a fuzzy synthetic evaluation technique. Ocean Engineering, 31(16), 1929-1953.

Simoff, S. J., & Maher, M. L. (1998, October). Ontology-based multimedia data mining for design information retrieval. In Proceedings of ACSE Computing Congress (Vol. 320). Cambridge, MA: ACSE.

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

Soibelman, L., and Caldas, C. (2000). "Project extranets for construction management: The American experience." Proc., Entac-2000, Salvador, Brazil.