Study on the Automatic Correlation Mechanism of Unstructured Information under BIM

Kunya Guo[1,*], Fengqiang Li[2], Zhenjiang Lei[1], Ran Ran[3], Yan Sun[4], Liang Yi[5]

[1]ICT Department, State Grid Liaoning Electric Power Supply Co., Ltd: Shenyang, 110006, China

[2]Office (Party Committee Office), State Grid Liaoning Electric Power Supply Co., Ltd: Shenyang, 110006, China

[3]Operation & Maintenance Center, State Grid Liaoning Information and Communication Company, Shenyang 110006, China

[4]Planning Review Center, State Grid Liaoning Electric Power Supply Co., Ltd: Shenyang, 110015, China

[5]Fujian Yirong Information Technology Co., Ltd: Fuzhou, 350100, China

*Corresponding Author email: gky@ln.sgcc.com.cn

Keywords: BIM; unstructured; automatic correlation mechanism

Abstract: Contributed by the rapid development of information technology, the research field of unstructured information has been paid a host spot in recent years. In view of that, on the basis of document name and document content, under BIM, this paper carried out detailed discussion on correlation degree calculation, calculation method selection, document and entity association method, and application effect evaluation.

1. Introduction

In recent years, China's architectural field has achieved a considerable development, although, the problems of “Information Gap" and "Information Island" still exist widely in this field. BIM technology can solve these problems to some extent; however, unstructured graphic documents, text documents and multimedia documents have put forward great challenges to the application of BIM technology. Dealing with this challenge is precisely the reason why this paper studies the automatic association mechanism of unstructured information under BIM.

2. Correlation degree calculation based on document name and document content

2.1 Correlation degree of document name

As a concise description of project document, document name can provide strong support for the identification of the model object, which has a fairly good performance in the identification of drawings, photos, videos and so on. If the naming standard of engineering documents and the development of professional coding rules can be realized at the same time, the use effect of document name correlation will be further enhanced. Taking into account that the name, description, and encoding are generally short, this paper uses Levenshtein Distance (LD) and Longest Common String (LCS) these two strings matching method, thereby carrying out higher correlation calculation of better quality. The concrete calculation revolves around the two strings, which are document information strings, model entity name, type, property, and other information strings respectively [1].

The core of the Levenshtein Distance (LD) string matching method is to convert one string to the required operand. The operation can be simply described as character changes, characters adding, and characters delete; the matching method of longest common string (LCS) is to take advantage of the length of the longest same string with the same subalphabets. The string involved in this method must have a contiguous attribute. It is worth noting that, due to the names and the descriptions of project

  DOI: 10.25236/systca.18.067

documents as well as other factors, the number of the degree of similarity generated by file name correlation degree calculation is often not high, which is resulted from that many engineering documents are not fully in line with IFC format and custom. This needs to be paid attention to.

2.2 Correlation calculation of document content

Due to the text document's content processing is relatively mature, the vector space model and probabilistic information retrieval model in this field can better serve the research of this paper. And based on the relevant research results, this paper has imitated the working mechanism of search engine, thus it can automatically match the model entity for the document with the expression in line with the specification. Table 1 directly shows the correlation degree's calculation method based on document content; the content of the three links referring document preprocessing, document information representation, and correlation degree calculation is shown as follows: (1) Document preprocessing. This link is to make full use of the document information, which will bring about relatively positive impact for the efficiency and quality enhancement of the follow-up links, so the study selected ShootSearch to be responsible for Chinese word processing. In addition, the removal of deactivated words is to avoid adding burdens to the document retrieval that are not of any use. (2) Document information representation. Combining search engine thinking, this paper chooses TF method to carry out correlation degree calculation, the core idea of which is to understand the character of the text through a lot of appearing words. Therefore, vector space model is used in this link, and the calculation formula $tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}$ is applied, of which the n_{ij} refers to the times that the term t_j appears in the document d_i. In addition, formula $idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$ needs to be applied, so that the weights w_{ij} can be obtained by multiplying TF and IDF. (3) Correlation degree calculation. After obtaining text vector space, the correlation degree calculation can be carried out centering sim_{ij}, which is the correlation degree of the document d_i and each entity e_i; the common measurement method is $sim_{ij} = \cos\theta_{ij} = \frac{d_i \bullet e_j}{\|d_i\| \bullet \|e_j\|}$, wherein we can see that the more similar two vectors and the higher the similarity, the higher the consistency of document and entity terminology.

Table 1 method of correlation degree calculation based on document content

Link	Document preprocessing	Document Information representation	Correlation degree calculation
Process	Chinese participle → remove the stop words;	Count the word frequency →Count reverse document frequencies → Form vector space model;	Select documents → Traverse each entity to compute correlation → Return to correlation ranking

3. Selection of calculation methods

3.1 Multimedia documents

Engineering documents often involve engineering drawing, conference recording, video and other multimedia content, in order to reduce the cost of the analysis and improve the efficiency of the analysis, this paper chooses the calculation method based on document name, which can be used to improve the matching effect by guaranteeing the naming convention of multimedia documents.

3.2 Text documents

The existing technology can achieve higher quality and speed of text document name and content analysis; on the ground of that, this paper chooses two methods based on name and content, so that similarity sim_1 and sim_2 can be obtained; thereby the similarity can be finally obtained by weighted average. In order to ensure the quality of similarity calculation, the fuzzy comprehensive evaluation method of multifactor decision analysis is selected, which can guarantee the maximum utilization of the name and content information of the project text and document [2].

4. Correlation method of documents and entities

After clarifying the relationship between the project document and the model entity, this paper obtains the process of adding the correlation relationship between documents and entities shown in table 2. When the BIM server uploads documents, the database can automatically add document names, uploaders, version numbers and other information records. At the same time, follow the IFC standard, so that you can add a pairing record between an entity and a document Globalid in a database; the record can establish an associated entity while exporting IFC files; Fig 1 is a reference diagram of the related IFC entity to the document, and the information displayed in the diagram plays an important role in the document's association with the entity.

Table 2 add procedure of the correlated relationship between documents and entities

Stage	Content	Details
1	Correlation degree calculation;	Automatically calculate the correlation and return to the list of related entities;
2	Users;	Users confirm the relevant entity;
3	BIM server;	Add correlated records to the correlated database;
4	IFC documents;	Establish correlation entity, and store correlated relation;

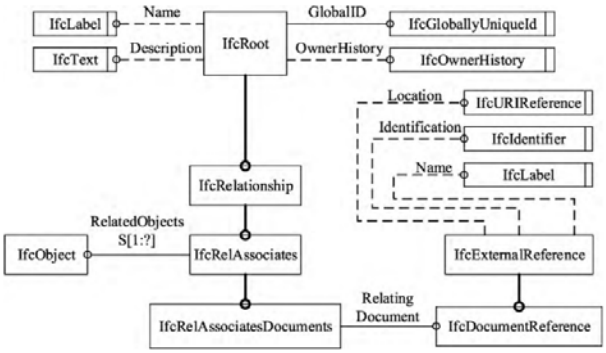

Figure 1 Reference diagram of the related IFC entity correlating with the document

5. Application effect evaluation

In order to verify the practical value of this research, this paper has carried out practical verification based on the service platform BIMDISP of the distributed BIM data integration and application of Tsinghua University, and thereby realizing the development of engineering document related service. Fig. 2 is the example selected in this paper, which is an 8~10 layers model of a building project. Fig 3 gives a simple display of the model's file information, in which each entity naming convention has a certain representative significance.

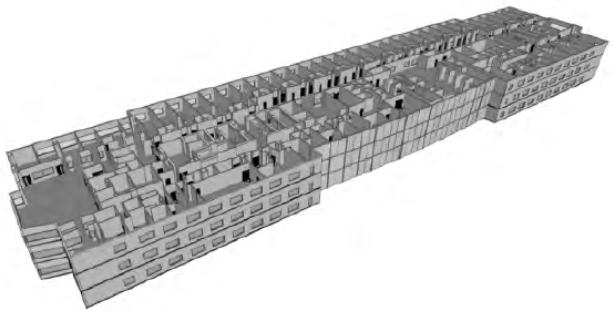

Fig. 2 Schematic diagram of the model example

Table 3 Information of the example

Information	IFC Entities	Interchangeable entities	Direct expression of building component entities
Number	325430	57782	2689

In order to measure the query effect, using the accuracy rate and recall rate as the index, all the objects of the query can be divided into four categories; and table 4 shows the classification visually. In the case of the project document of "flat window purchase Picture", the user uploads the document, the list of related entities returned by the system as shown in Figure 3, the associated relationship can be put into the database, and the expression result can be queried when the model is browsed. Combined with the above results, it is better to prove that this paper studies the document Association mechanism has the higher practical value.

Table 4 Evaluation method of accuracy and recall rate

	Retrieved	Not retrieved
Related entities	A	B
Unrelated entities	C	D

GlobalID	Name	Edition	ID	Correlation degree
176t9ucL4fhbqidRcdlGl	Dan Shanping window 1- with veneer：600×1200 mm：198436	ifcWindow	3181	0.4491342455377054
176t9ucL4fhbqidRcdlGz	Dan Shanping window 1- with veneer：600×1200 mm：198411	ifcWindow	3139	0.4491342455377054
Ellipsis of the same entity				
176t9ucL4fhbqidRcdlN4	Dan Shanping window 1- with veneer：600×1200 mm：198436：1	ifcOpeningElement	3097	0.348029052711994
176t9ucL4fhbqidRcdlNf	Dan Shanping window 1- with veneer：600×1200 mm：198367:1	ifcOpeningElement	3055	0.348029052711994

Figure 3 List of related entities returned by the system

6. Conclusion

To sum up, the automatic correlate mechanism of unstructured information under BIM can provide more powerful support for the development of architecture field. On the basis of this, the author has proved the practical significance of the research, such as the calculation of document name correlation degree, the calculation of document content relevancy, the add procedure of document and entity correlating relationship, and other specific content. Therefore, in the relevant field of theoretical research and practical exploration, the contents of this article can play certain reference

significance.

References

[1] Kokkalis Z T, Iliopoulos I D, Pantazis C, et al. What's new in the management of complex tibial plateau fractures?[J]. Injury-international Journal of the Care of the Injured, 2016, 47(6):1162-1169.

[2] Huang K T, Hung Y W, Fang R X, et al. P-147: Study on the Correlation of Flicker Shift Phenomenon and Ion Accumulation Mechanism in FFS Mode LCD Panel [J]. Sid Symposium Digest of Technical Papers, 2017, 48(1):1834-1837.