A COMBINED TEXT MINING METHOD TO IMPROVE DOCUMENT MANAGEMENT IN CONSTRUCTION PROJECTS

Caldas, Carlos H. [1] **and Soibelman, L.** [2]

ABSTRACT

Information is an important element of project delivery processes. Throughout the project life-cycle, information is needed for the planning, implementation, control, and analysis of project activities. Moreover, the use of communications and information technologies can support the project delivery process by helping to make value adding activities more efficient as well as to eliminate non-value adding activities through enhanced project controls.

In the last few decades, the architecture, engineering, and construction (AEC) industry has experienced an increased availability of electronic project data due to new information technologies. Evolving project requirements also led to more complex projects that generate large amounts of data. These data are usually generated and stored in different data types such as relational databases, electronic text documents, CAD drawings, pictures, video, and audio, among others. These factors justify the importance of efficient and effective project data and information management techniques.

A large percentage of AEC information is stored on text documents, including contracts, specifications, meeting minutes, change orders, field reports, and requests for information. This emphasizes the need for the development of methods for improving the organization, search, and retrieval of this type of data in project information management systems. This paper describes a study that aimed to formalize a method to address these issues. The proposed method combines the vector space model for information retrieval and text mining algorithms. Experiments conducted using document collections from several construction projects demonstrated the efficiency of the proposed approach.

KEYWORDS

text mining, document management, information systems, project management, information retrieval.

[1] Assistant Professor. Department of Civil Engineering. University of Texas at Austin. 1 University Station C1752. Austin, TX 78712-0273. caldas@mail.utexas.edu

[2] Associate Professor. Department of Civil and Environmental Engineering. Carnegie Mellon University. Porter Hall 118N, Pittsburgh, PA 15213. lucio@andrew.cmu.edu

INTRODUCTION

Several studies emphasized the importance of information integration methodologies to improve organization and access in inter-organizational construction information management systems (Sanvido and Medeiros, 1990; Reinschmidt et al., 1991; Turk et al., 1994; Fisher and Kunz, 1995; Rezgui et al., 1996; Eastman, 1999; Teicholz, 1999; Zhu et al., 2001; Froese, 2003)

The majority of the architecture, engineering, construction, and facilities management (AEC/FM) information integration initiatives focus on structured data. Structured data is defined here as data that have a database like structure, usually in information systems that use some form of database in the background. However, a large percentage of the construction data is exchanged using text-based documents such as contracts, change orders, requests for information, among others. These documents contain valuable information for project planning, implementation, control, and analysis. This large number of documents creates challenges for project information management. Therefore, the management of the information contained in these types of documents becomes crucial to construction management.

This paper presents a research study that investigated methodologies to provide semi-automatic support for project document integration in model-based information systems. The main goals of this study were to improve the organization and access to large document collections in project management information systems and to promote the integration of text documents in model-based information systems. A text information integration methodology was proposed, implemented, verified, and validated. The methodology is composed of three main steps: classification, retrieval and ranking, and association. This paper summarizes these steps, introduces a prototype software system that implements the proposed methodology, and presents an overview of the research experiments and results. Details of the proposed approach are included in Caldas and Soibelman (2005).

TEXT INFORMATION INTEGRATION METHODOLOGY

The formalization of the proposed methodology required the identification of the challenges associated with text information integration. For this purpose, construction databases were analyzed and feedback from industry practitioners was obtained regarding current approaches and technologies used for project document management, including project websites, document management systems, and project contract management systems. A text information integration methodology (TIIM) was then formalized and implemented. The methodology was then verified and validated using data from existing projects. The steps of the TIIM methodology will be briefly presented in the following sections.

Classification

The first step of the TIIM model is automated project document classification. This step is based on pattern classification algorithms. However, available classification algorithms cannot be applied directly. Several steps need to be accomplished beforehand. Decisions made in these preprocessing steps affect the results. These factors justified and motivated the development of an automated document classification process to implement the first step of the TIIM method (Caldas and Soibelman, 2003a). Experiments using the proposed process demonstrated that factors and parameters such as the choice of the classification algorithms and index weighting methods, as well as the use of dimensionality reduction techniques, affect the final classification results.

Since each construction document can belong to more than one class, the classification process was designed to handle multiple binary classifications. In this case, each document is compared with each class. For each class, a binary decision is made in order to define whether or not the document is related with that particular class. This process is repeated to all classes defined for the project being considered. Pattern classification algorithms are then used to create a classification model for each particular class by using a set of training documents that have previously been classified manually by a domain expert. This approach relies on the existence of previously classified documents.

The main components of the proposed document classification process are detailed in Caldas and Soibelman (2003a) and include: data collection, data conversion, dimensionality reduction, data preparation, data transformation, learning, and document classification.

Retrieval and Ranking

The proposed approach the retrieval and ranking of project documents combines the vector space model for information retrieval and pattern classification techniques to develop a specialized information retrieval and ranking system for the AEC/FM domain.

In this step, the concept of an AEC/FM domain-specific information retrieval and ranking system was formalized and a prototype system was implemented to prove this concept. The vector space model (Salton and Lesk, 1968) was selected for document representation because the resulting model could be uniformly applied both to the *classification* and the *retrieval and ranking* steps of the proposed TIIM model.

As a starting point, a query vector is constructed using terms extracted from the project model related to a selected model object. This vector, as well as the classification notations of the classifications associations from the selected object, is used as input for information retrieval.

A classification-based retrieval and ranking algorithm was developed to implement this step of the model. In order to apply the classification-based retrieval and ranking algorithm, all project documents must be classified using the process described previously and all document vectors need to be normalized.

The documents that match the model object's class are identified and selected. Then, the class centroid vector of all document vectors belonging to the object's class is calculated. The coordinates of the class centroid vector are calculated as the average of the coordinates of each document vector that belongs to the class under consideration. A term vector is created based on terms extracted from the object's description. This vector has coordinate values equal to one for the coordinates that correspond to the object's terms and coordinate values equal to zero for all other coordinates.

A query vector is then calculated based on the term vector and the class centroid vector. This query vector is the representation of the project model object in the multidimensional space composed of project document vectors. Since the class centroid vector, the term vector, and the query vector are in the same multidimensional space of the project document collection vector space model, these vectors have the same number of coordinates.

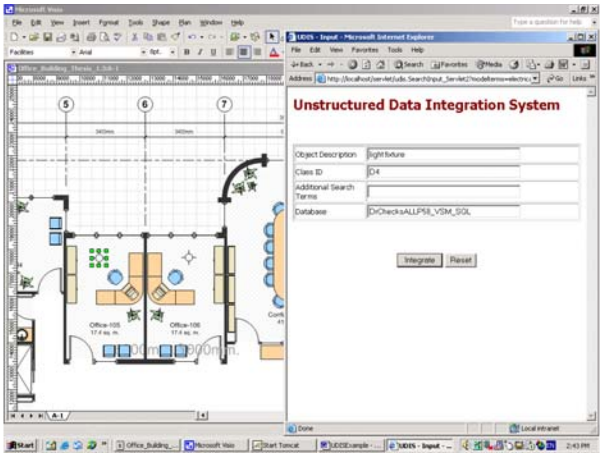
The identification of documents that are related to the selected object is based on the similarity between the query vector and each of the document vectors from the project document collection. This similarity is calculated using the cosine between two vectors. The similarity measure is calculated for all documents in the collection. Results are ranked and the documents whose similarities measurements are above a defined threshold are retrieved.

Association

Association of text documents to project model objects is the last step on the text information integration model. After project documents have been ranked and retrieved, a reference to the relevant documents is added to the project model. It is important to emphasize that the actual document is not added to the model, just a reference to it. This facilitates future updates, saves storage space and computer memory, and helps the integration with existing information systems. This association was implemented in the context of IFC specification (IFC 2x2, 2003). The document association relationship *IfcRelAssociateDocument* is used to establish links between a document reference *IfcDocumentReference* or document and any type of project object *IfcObject*.

TIIM MODEL: IMPLEMENTATION

A model-based information system prototype based on the TIIM methodology was developed for proof of concept and to conduct the validation experiments. This prototype system, called Unstructured Data Integration System (UDIS), is composed of smaller prototype systems that were developed to implement each of the steps of the TIIM model, including the Vector Space Modeler System (VSMS) and the Construction Document Classification System (CDCS) (Caldas and Soibelman, 2003a). VSMS is used to generate the vector space model representation of the project document collection. CDCS classifies project documents according to items of a construction information classification system. Figure 1 shows the UDIS


UDIS - Input - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Unstructured Data Integration System
Integrate
Reset

Object Description	light fixture
Class ID	D4
Additional Search Terms	
Database	DrChecksALLP58_VSM_SQL

Figure 1: Unstructured Data Integration System

TIIM MODEL: RESULTS AND VALIDATION

The study involved the analysis of more than 25 construction databases and 30,000 electronic documents. The document types included requests for information (RFI), design reviews, specifications, and change orders, among others. The typical file formats were PDF and MS Word. These databases were from real projects obtained from private contractors and the US Army Corps of Engineers. They encompassed different type of projects, such as commercial building, hotels, dormitories, and military bases, among others. Twenty of these databases were specifically used in the experiments conducted to validate the text information integration methodology. These experiments were performed using the prototype software systems previously described. The main objectives were to compare the proposed integration method with existing solutions for project document management. The technologies used for comparison were project contract management systems, project websites, and information retrieval systems. Popular commercial systems currently being used by the AEC/FM industry were used to conduct this study. Originally these systems do not provide support for the integration of project documents in model-based information systems. Therefore, the comparison was based on the existing retrieval and ranking capabilities of these systems.

Documents related to selected IFC objects from each of the twenty construction databases were identified, retrieved, and ranked using the systems under comparison. The performance measures of precision and recall were calculated (Baeza-Yates and Ribeiro-Neto, 1999) and used for comparison. Precision is the percentage of the retrieved documents that were related to the selected IFC object. Recall is the percentage of the related documents that were actually retrieved. Statistical methods were used to analyze the results. Statistical paired tests were selected because the available experimental units were considerably different prior to their random assignment to the groups (systems under comparison) with respect to characteristics that may affect the experimental responses.

The average recall results were 66.86% for the UDIS system, 44.17% for the project contract management system, 31.06% for the project website, 41.64% for the information retrieval system 1, and 31.70% for the information retrieval system 2. The difference between the results of the system that implements the TIIM model and each of the 4 other approaches was considered significant as indicated by the low values for the level of significance (ρ-value) obtained from the statistical tests. The average precision results were 46.33% for the UDIS, 28.55% for the project contract management system, 29.20% for the project website, 31.45% for the information retrieval system 1, and 29.58% for the information retrieval system 2. Once again, this difference was considered significant according to the results of the tests. The results demonstrated a significant improvement in the capability to identify documents that are related to project model objects.

CONCLUSIONS

In summary, the contributions from this research project can be divided in two parts. First a methodology for integrating project documents in model-based information systems was developed. The methodology was verified and validated using documents from several existing projects. The research results demonstrated that the methodology promotes a significant improvement in the capability to identify documents that are related to project model objects. Access to project documents is improved because large collections of documents can be analyzed more effectively. Differences in vocabulary are minimized using the classification-based approach and process automation makes the results more consistent. Integrated documents can then be used to support the planning, implementation, control, and analysis of project activities. The second contribution was the development of semi-automatic methods for the classification, retrieval and ranking, and association. The way in which these methods were formalized facilitates their incorporation on existing project management information systems. This is an important requirement for the application of the proposed methodology in real-world projects.

The formalisms developed in this research to deal with text documents can be expanded to other types of data generated on AEC/FM projects. Interfaces with other research areas in construction engineering and management such as data collection, data mining, planning, optimization, and simulation can be established. Similarly, new research areas such as integrated project data analysis and proactive project information systems can be explored

REFERENCES

Baeza-Yates, R. and Ribeiro-Neto, B. Anumba, (1999). *Modern Information Retrieval*, ACM Press, New York, NY.

Caldas, C., Soibelman, L., and Gasser, L. (2005) "A methodology for the integration of project documents in model-based information systems." Journal of Computing in Civil Engineering, 19(1), 25-33.

Caldas, C. H. and Soibelman, L. (2003a) "Automating hierarchical document classification for construction management information systems." Automation in Construction, 12(4), 395-406.

Eastman, C.M. (1999). Building product models: computer environments supporting design and construction. CRC Press, Boca Raton, FL.

Fischer, M., and Kunz, J. (1995) "The circle: architecture for integrating software." Journal of Computing in Civil Engineering, 9(2), 122-133.

Froese, T. (2003) "Future directions for model-based interoperability". Proceedings of the 2003 ASCE Construction Research Congress, Honolulu, HI.

IFC 2x2 (2003). *IFC2x Edition 2*. International Alliance for Interoperability.

Reinschmidt, K. F., Griffis, F.H. and Bronner, P.L. (1991). "Integration of Engineering, Design, and Construction." Journal of Construction Engineering and Management, 117(4), 756-772.

Rezgui, Y., Brown, Y., Cooper, G., Yip, J., Brandon, P., and Kirkham, J. (1996) "An information management model for concurrent construction engineering." Journal of Automation in Construction, 5(4), 343-355.

Salton, G. and Lesk, M. E. (1968) "Computer evaluation of indexing and text processing." *Journal of the Association for Computing Machinery*, 15(1), 8-36.

Sanvido, V. E. and Medeiros, D.J. (1990). "Applying computer-integrated manufacturing concepts to construction." J. Constr.Engrg. and Mgmt., ASCE, 116(2), 365-379.

Turk, Ž., Björk, B-C., Johansson, C. and Svensson, K. (1994) "Document Management Systems as an Essential Step Towards CIC," Preproceedings CIB W78 Workshop on Computer Integrated Construction, VTT, Helsinki.

Zhu, Y., Issa, R. R. and Cox, R. F. (2001). "Web-based construction document processing via malleable frame" Journal of Computing in Civil Engineering, 15(3), 157-169.