AUTOMATION IN CONSTRUCTION

Automating hierarchical document classification for construction management information systems

Carlos H. Caldas, Lucio Soibelman*

*Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign,
Newmark CE Lab. MC 250, 205 North Mathews Avenue, Urbana, IL 61801, USA*

Abstract

The widespread use of information technologies for construction is considerably increasing the number of electronic text documents stored in construction management information systems. Consequently, automated methods for organizing and improving the access to the information contained in these types of documents become essential to construction information management. This paper describes a methodology developed to improve information organization and access in construction management information systems based on automatic hierarchical classification of construction project documents according to project components. A prototype system for document classification is presented, as well as the experiments conducted to verify the feasibility of the proposed approach.
© 2003 Elsevier Science B.V. All rights reserved.

1. Introduction

The use of communications and information technologies in the construction industry is creating new opportunities for collaboration, coordination, and information exchange among organizations that work on a construction project. Inter-organizational construction management information systems are increasingly being used for this purpose. They comprise a set of interrelated components that collect, retrieve, process, store, and distribute data to support planning, control, and decision-making among project organizations. In the distributed and dynamic construction environment, the ability to exchange and integrate data from different sources and in different formats becomes crucial to the development of the construction processes supported by these management information systems. Furthermore, the data collected in these systems provide a valuable source for data mining [11,28]. Discovered knowledge can be used to increase the performance of future activities and projects.

Given that a large percentage of the project documents is generated in text format, methods for organizing and improving access to the information contained in these types of documents become essential to construction information management. Construction information classification systems (CICSs) can be used to support this information management process. The classification structure in a construction information

* Corresponding author. Tel.: +1-217-333-4759; fax: +1-217-333-9464.
E-mail addresses: caldas@uiuc.edu (C.H. Caldas),
soibelma@uiuc.edu (L. Soibelman).

classification system (CICS) defines concept hierarchies that can be used for document classification, providing a common framework for document organization and management among project organizations. These classification frameworks can be embedded in inter-organizational information systems, like project websites, project management software, and document management systems. Examples of CICSs include: the CSI MasterFormat [17], CSI UniFormat [33], CI/SfB, Uniclass, and the Overall Construction Classification System [20].

One limitation of the existing inter-organizational information systems is the reliance on manual classification methods conducted by human experts. With the growth in the use of information technologies by construction companies, the increasing availability of electronic documents, and the development of model-based systems, manual classification becomes impractical. One example of the limitations of manual classification is the time and effort that would be required to classify all documents created in a construction project (contracts, specifications, meeting minutes, change orders, field reports, and requests for information, among others), according to all components of a CICS.

Another limitation of the current systems is the consideration of documents as single units for the purpose of classification and retrieval. Many construction documents, including specifications and meeting minutes, should clearly be divided and then assigned to more than one item of a CICS. This limitation can be illustrated by the case in which a project manager wants to access information contained in meeting minutes regarding a specific CSI MasterFormat item in order to solve an issue. Using current technologies, the project manager would need to manually search and analyze each document individually in order to obtain the desired information.

A third problem that exists in available systems is the lack of support for differences in vocabularies and naming conventions. This problem can be illustrated by the case in which an architect gives a name for a particular object in a project model. Since there is usually no standard vocabulary among organizations that participate in a construction project, references to that particular object in project documents are often done using different names. Using current technologies, project managers would need to map the model object's name to the terms being used in the different construction documents.

The previously mentioned limitations and the push towards fully integrated and automated project processes justify the need for the development of automated classification methods for construction project documents that can explore the internal characteristics of these documents and adapt to different classification frameworks.

This paper presents a unique way to improve information organization and access in inter-organizational construction management systems based on methods for automated hierarchical classification of construction project documents according to CICSs items. In order to accomplish this goal, a combination of techniques from the areas of information retrieval and text mining was explored. As a result, a methodology for automated hierarchical document classification was devised and implemented. A prototype of a construction document classification system was also developed to provide easy deployment and scalability to the classification process. The developed prototype automated all steps of the text classification process. Experiments were conducted to validate the results and demonstrate the applicability of the implemented techniques.

2. Construction management information systems

The escalating globalization and complexity of construction projects have increased the participation of companies from diverse locations in project teams [3]. In this environment, effective inter-organizational construction management information systems able to minimize time and distance constraints are necessary. Examples of such systems are described extensively in literature [18,19,22,27,32,34,39]. In the distributed and dynamic construction environment, the ability to exchange and integrate information from different sources and in different data formats becomes crucial to the improvement of the construction processes supported by these systems. Simoff and Maher [26] argue that a key issue in managing construction information is the diversity of data types, including:

- structured data files, stored in database management systems or specific applications, such as data