AUTOMATED CLASSIFICATION METHODS: SUPPORTING THE IMPLEMENTATION OF PULL TECHNIQUES FOR INFORMATION FLOW MANAGEMENT

Carlos H. Caldas[1] and Lucio Soibelman[2]

ABSTRACT

One of the requisites for implementing lean construction processes is the management of information flows through the networks of cooperating project organizations. Information flows about directives, criteria, prerequisites, availability, commitments, and resources are essential to production control and work structuring. Since a large percentage of these project information is generated in text format, methods for managing the information contained in these types of documents becomes essential to improve work flow reliability. Information management systems have been used for this purpose. One limitation of the text-based information management aspects in current systems is the reliance on push methods. Push systems schedule the release of information based on demand. On the other hand, pull systems release information based on system status. For that reason, the implementation of pull information systems is an essential requirement of lean construction delivery systems. This paper describes a methodology to support the implementation of pull techniques in construction management information systems based on automated text classification methods.

KEY WORDS

Construction management, information flows, information management, machine learning, pull systems, text/data mining.

[1] Ph.D. Candidate. Department of Civil and Environmental Engineering. University of Illinois at Urbana-Champaign. 3142 Newmark CE Lab. Urbana, IL 61801. caldas@uiuc.edu.

[2] Assistant Professor. Department of Civil and Environmental Engineering. University of Illinois at Urbana-Champaign. 3129c Newmark CE Lab. Urbana, IL 61801. soibelma@uiuc.edu.

INTRODUCTION

Managing the flow of information through networks of cooperating project members is an essential requirement of Lean Construction Delivery Systems (Ballard 2000). According to Koskela (2000), the use of communications and information technologies in the construction industry can support the delivery process by helping eliminate non-value-adding activities through enhanced controls and also by making non-value-adding activities more efficient. Inter-organizational information systems are being increasingly 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 systems. Furthermore, the data collected provide a valuable source for data mining (Soibelman and Kim 2002; Han and Kamber 2001). Discovered knowledge can be used to improve the performance of future activities and projects.

Since 1999, the Construction Knowledge Discovery and Dissemination Group (CKDD) in the Department of Civil and Environmental Engineering at the University of Illinois at Urbana-Champaign has been working on the analysis of data from these inter-organizational information systems. Some of the observations in our research were: (i) current systems rely on push mechanisms for information management; (ii) a large percentage of the data is stored on semi structured and unstructured files; (iii) it is very hard to find the information needed for decision making; (iv) the information is not integrated with other systems; (v) there is not clear association between information and their related project product and process components.

Given that a large percentage of the project documents is generated in text format, methods for organizing and improving the access to the information contained in these types of documents becomes essential to work flow reliability. Inter-organizational construction management information systems have been used for this purpose. Current systems that provide features for information organization and access include project websites, project management software, and document management systems.

As previously mentioned, one limitation of the existing systems is the dependence on push techniques. Push information systems release information based on manual classification and retrieval methods controlled by human experts. On the other hand, pull systems release information based on system status and should be implemented using automated processes. For that reason, the development of pull information systems is an essential requirement of lean construction delivery systems (Ballard 2000). One example of the limitations of manual classification and retrieval is the time and effort that would be required in order to access all documents that contains the directives, criteria, prerequisites, availability, commitments, and resources for each of the components of an activity definition model (ADM).

Another limitation of available 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 project component or object. This limitation can be illustrated by the case in which a project manager wants to access information contained in meeting minutes regarding a specific project component 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 current 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 is 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. Similar case occurs when a designer assigns a description of a particular object in a project model. If a cost estimator wants to match this description to an item in the cost database, he/she will have to manually decide on what cost item to assign.

The previously mentioned limitations and the drive towards fully integrated and automated project processes justify the need for the development of pull techniques for information flow management based on automated classification methods for construction project documents. Some benefits and applications of such methodologies include, but are not limited to:

- Information Organization and Access: Development of intelligent search engines.
- Procurement: Identification of materials that meet the project specifications.
- Project Control Systems: Enable automated access to project specifications.
- Data Analysis: Identify problem areas and potential causes of delays, cost overruns, or quality deviations.
- Knowledge Generation: Extract lessons that could be applied in future activities and projects.

This paper presents a unique approach to support the development of pull techniques for information flow management, based on methods for automated classification of construction project documents. In order to accomplish this goal, a combination of techniques from the areas of Information Retrieval and Text Mining were explored. Pattern classification techniques form the basis for the proposed classification methods. Algorithms like Support Vector Machines, Rocchio, Naïve Bayes, and K-Nearest Neighbors were explored. As a result, a framework for automated 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. Case studies based on data obtained from past and current projects were conducted to validate the results and demonstrate the applicability of the implemented techniques.

INFORMATION FLOW MANAGEMENT

In the distributed and dynamic construction environment, the ability to manage information flows from different sources and in different data formats becomes crucial to the improvement of the construction processes supported by inter-organizational

information systems. Maher and Simoff (1998) argue that a major hurdle in managing construction information is related to its variety of data types, including:

- Structured data files, stored in specific applications or database management systems (e.g.: data warehouse, enterprise resource planning, cost estimating, scheduling, payroll, finance, accounting).
- Semi structured data files (e.g.: HTML, XML or SGML files).
- Unstructured text data files (e.g.: contracts, specifications, catalogs, change orders, requests for information , field reports, meeting minutes).
- Unstructured graphic files, stored in binary format (e.g.: 2D and 3D drawings).
- Unstructured multimedia files (e.g.: pictures, audio, video).

For instance, let's consider a typical construction situation in that a construction manager wants to find all available information about one construction activity, say, building a wall. He/she will probably find the drawings in CAD files, the specifications in text documents, the cost estimates in spreadsheets, the schedule in particular application formats, the contracts in text documents, communication among project members in e-mail files, and price quotes in different websites. A major task is how to retrieve, classify, and integrate information in these different file formats, especially considering that the files can also be stored in different organizations, computers, or file systems.

Information integration methodologies have been investigated worldwide in order to improve information organization and access in inter-organizational management information systems. Teicholz (1999) argues that project information should be integrated in three dimensions: *(1) horizontal integration of multiple disciplines that take part in a construction project; (2) vertical integration of multiple stages in the life cycle of a facility and (3) longitudinal integration overtime, which is also related with the capture of knowledge that allows improved performance or better decisions in the future*.

Fisher and Kunz (1995) argue that technical and managerial strategies have been used to improve information integration. On the technical side, there are four approaches to achieve integration (Rezgui et al. 1996; Zhu and Issa 2001): *(i) communication between applications; (ii) knowledge-based interfaces linking multiple applications and multiple databases; (iii) integration through geometry, and (iv) integration through a shared project model holding all the information relating to a project according to a common infrastructure model*.

The technical integration through a shared data model can be based on the creation of a centralized project model using 3D/4D CAD (Aalami et al., 1998) or on the development of infrastructure to facilitate the integration of decentralized project information using distributed software architectures (ToCEE 2000; Soibelman and Peña-Mora 2000). The adoption of data standards can support these integration approaches. Examples of initiatives in this area are presented by Eastman (1999), and include the ISO-STEP, the Industry Foundation Classes (IFC) created by the International Alliance for Interoperability (IAI 1996), and the aecXML specification that is being developed by the AEC Working Group (aecXML, 1999).

One limitation of the current AEC/FM information integration approach is the focus on structured data types. Some recent research work addressed some of the issues related with unstructured data integration. Fruchter (1999) describes tools to capture, share and

reuse project information. Wood (2000) describes an approach to extracting concepts from textual design documentation. Brüggemann et al. (2000) proposed the use of arbitrarily structured metadata to markup documents. Scherer and Reul (2000) uses text clustering techniques to classify documents and retrieve project knowledge from heterogeneous AEC/FM documents. Yang et al. (1998) and Kosovac et al. (2000) proposed the use of controlled vocabularies (thesauri) to integrate heterogeneous data representations.

CONSTRUCTION DOCUMENT CLASSIFICATION SYSTEM

From the observations and problems presented in the previous sections, we can infer that information flow management plays an important role in lean project delivery systems. Since a great percentage of the information exchanged among construction organizations is stored in unstructured text data files, the management of the information contained in these types of documents becomes crucial. In order to improve the management of text-based information, an automated document classification system was devised and implemented. The importance of this study is that automated document classification methods can be used as a foundation for the development of pull techniques for information flow management. For instance, it can be used to implement automated information routing mechanisms and to develop system status report tools.

The Construction Document Classification System (CDCS) was implemented in order to test the feasibility of the proposed approach. The system automates the steps involved in the document classification process and is currently composed of six main modules named data collection, data conversion, data preparation, dimensionality reduction, learning, and classification. These modules are described and detailed in the next sections. Case studies conducted to assess the feasibility of the proposed methodology are also presented.

Data Collection

The first module provides support for the collection of the text-based documents that should be classified. Initially, we only considered the case in which the data is stored in a central location, but the module will also be adapted in order to collect data from distributed databases.

Data Conversion

The documents to be classified are usually stored in different data formats, including: word processor, spreadsheet, e-mail, HTML, XML, PS, and PDF files. In order to apply the classification algorithms, these files needs to be converted to text file format. This module implements interfaces to file converter systems in order to create a text version of each document, while keeping the original documents in their native formats and locations. The text versions were then used in the remaining steps of the classification process.

Data Preparation

Classification algorithms cannot directly interpret text documents. For this reason, a preparation and indexing procedure that maps a text document into a compact representation of its content needs to be uniformly applied to training and test documents.

This module supports the transformation of text documents, which typically are strings of characters, into a representation suitable for the classification task.

The data preparation module applies the vector space model for document representation. In the vector space model, vectors of words represent documents. The collection of documents is represented by an *m* x *n* term-by-document weighted frequency matrix $A = \{a_{ij}\}$, where a_{ij} was defined as the weight of a word *i* in document *j*. Each of the *m* unique terms in the document collection is assigned a row in the matrix, while each of the *n* documents in the collection is assigned a column in the matrix. A non-zero element a_{ij} , indicates not only that term *i* occured in document *j*, but also the number of times the term appears in that document or its relative weight. Since the number of terms in a given document was typically far less than the number of terms in the entire document collection, the matrix A is usually very sparse.

Several ways of determining the weights a_{ij} are suported by the system, including: Boolean weighting, absolute frequency, tfxidf-weighting, and tfc-weighting, (Salton and Buckley 1988). These approaches were developed based on two empirical observations regarding text documents: (i) the more times a word occurs in a document, the more relevant it is to the topic of the document, and (ii) the more times the word occurs throughout all documents in the collection, the more poorly it discriminates between documents. It is important to mention that different weighting schemes can conduct to different effectiveness and this is one of the reasons why they all should be considered.

Dimensionality Reduction

According to Sebastiani (1999), a major characteristic, or difficulty of text classification problems is the high dimensionality of the feature space. Standard classification techniques cannot deal with such a large feature set, since processing is extremely costly in computational terms. Hence, there was a need to reduce the original feature set, which is commonly known as dimensionality reduction (DR) in the pattern recognition literature.

Various DR functions have been proposed, either from the information theory or from the linear algebra literature (Yang and Pedersen 1997). While each of the dimensionality reduction methods have its own rationale, the ultimate word on its value was given by the effectivenes results from applying it to the collection of documents and then training a classifier on the reduced representation. In CDCS, the information gain method was selected as the dimensionality reduction method.

Learning

After reducing the dimensions of the representation of the document set, a number of statistical classification and machine learning techniques can be applied to classify text-based documents. In construction industry, the classes can be represented by construction project components. Hence, document classification is defined as the task of assigning a Boolean value to each pair $\{ d_j , o_i \} \in D \times O$, where D is a domain of documents and $O = \{ o_1 , ... , o_n \}$ is a set of project components (classes). A value of T (true) assigned to $C (d_j , o_i)$ indicates a decision that document d_j is related with component o_i, while a value of F (false) indicates that d_j is not related with the component o_i.

By using machine learning algorithms, an inductive process automatically built a classifier (classification model) for each class by observing the characteristics of a set of documents that have previously been classified manually by a domain expert. The

classification problem was an activity of *supervised learning*, since the learning process was driven by the previous knowledge of the categories in some of the documents that were used to build the model. Hence, this approach relies on the existence of an initial corpus of documents previously classified according to their relevance to a set of project components. A document d_j is called a positive example of o_i if $C(d_j, o_i) = T$ and a negative example of o_i if $C(d_j, o_i) = F$.

After generating the classification model, its effectiveness is evaluated. The alternative adopted for this evaluation is to split the initial collection of documents into two sets:

- Training Set: set of documents that were used to create the classification model;
- Test Set: set of documents that were used for testing the effectiveness of the classifier.

The documents in the Test Set should not participate in the inductive construction of the classifier. If this condition were not satisfied then the experimental results obtained would probably be unrealistically good. The definition of the size of the Training Set is also crucial to avoid overfitting. This happens when the classifier performs with few errors on the Training Set and does not generalize to the new test cases.

The Machine Learning techniques that have being used for text classification and are being implemented in this module include: Rocchio Algorithm, Naïve Bayes, Decision Trees, K-Nearest Neighbors, Support Vector Machines, and Boosting Algorithms.

As previously mentioned, there are several methods for data preparation, dimensionality reduction, and learning. Their choice also affects the classification results. Therefore it is important to test different combinations of these methods in order to improve the accuracy of the classifier. CDCS provides an environment for the use and simulation of different preprocessing and learning methods, giving more flexibility and power for the classification task, and also simplifying the classification process for the users.

Classification

Since each construction document can belong to more than one class (one individual document can be related to more than one project component), the system 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 if the document is related or not with that particular class (project component). The large number of classes that usually needs to be defined in order to classify construction documents imposes another challenge to the classification task. For multiple binary classifications, a classification model has to be created for each of the existing classes.

In CDCS, the classification structure can be defined according to a hierarchy of classes. For instance, considering the CSI MasterFormat (MasterFormat 1995) as the classification structure, the document is initially classified according to each element of the first level (CSI MasterFormat Divisions). For the elements in the first level in which the classification decision was true (meaning that the document was related with that particular CSI MasterFormat Division), the binary classification can then be conducted for the second hierarchical level (CSI MasterFormat Level 2). Following the same process, for the elements in the second level in which the classification decision was true

(meaning that the document was related with that particular CSI MasterFormat Level 2 Element), the binary classification can then be conducted for the third hierarchical level (CSI MasterFormat Level 3). Figure 1 presents the results from the hierarchical classification of a construction document based on 3 of the CSI MasterFormat Divisions. In this particular example, the document was classified as *Woods and Plastics* in the first level, *Architectural Woodwork* in the second level, and *Wood Stairs and Railing* in the third level.

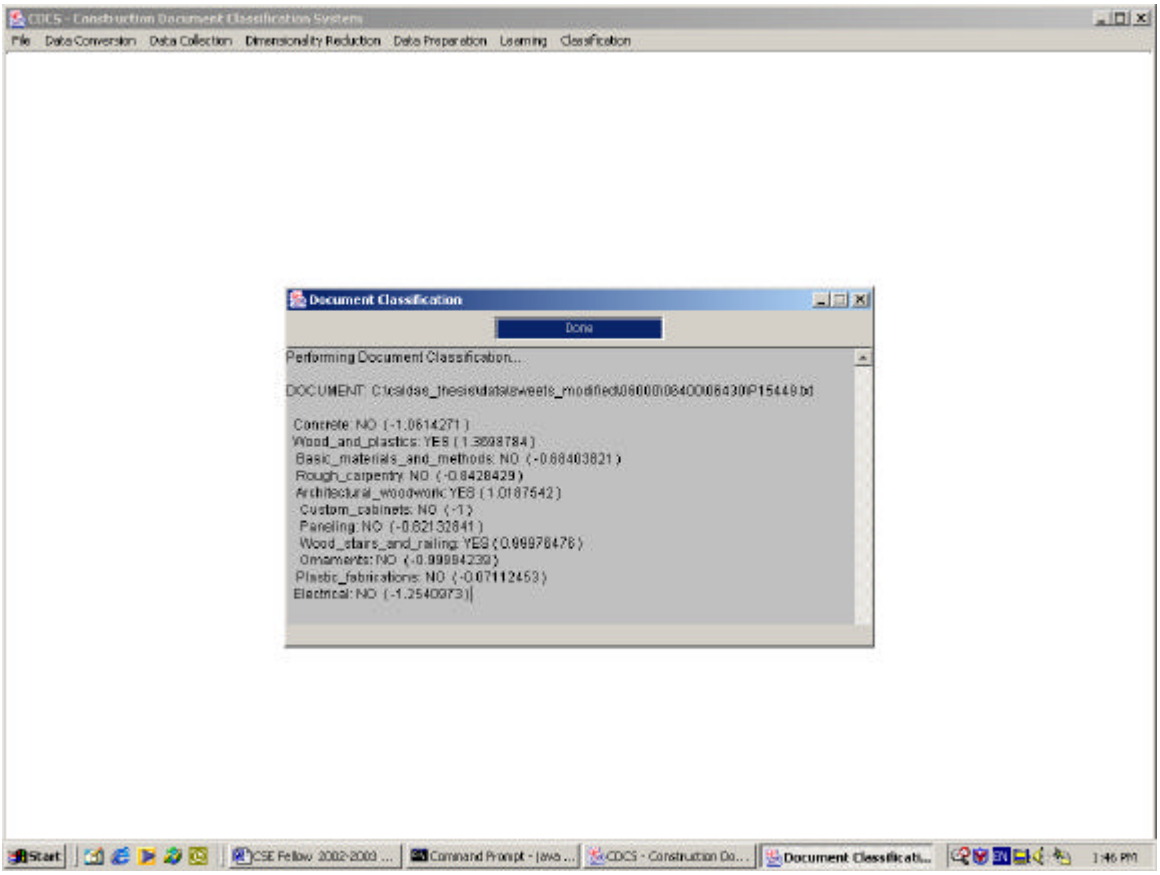

Figure 1: Hierarchical Document Classification Using CDCS

CASE STUDIES

The first case study used data from construction inter-organizational information systems (project extranet) to evaluate the accuracy of text classification algorithms. A database from one building project in Boston, MA was selected for this case study. The database was used by 16 project team organizations and contained more than 5,000 document files (1.5 GB). Several types of construction documents were available in this database, including specifications, meeting minutes, requests for information (RFI), architect supplemental instructions (ASI), change orders, and field reports among others. Meeting minutes were selected for this evaluation. They store information about weekly progress meetings among project participants. Each topic discussed during the meeting is recorded in separate items. These items are grouped by topics according to specific project divisions defined by the project team.

Originally there were 92 meeting minutes. Each item for all of these meeting minutes was automatically extracted from the original document and stored in separate document files. A total of 845 documents were then used in the document classification process. In this analysis, we tested four text classification algorithms namely Rocchio, Naïve Bayes, K-Nearest Neighbors, and Support Vector Machines. The choice of these algorithms was based on results from previous classification algorithms evaluations reported in the literature (Yang 1999; Sebastiani 1999). The Bow Toolkit (McCallum 1996) was used in these experiments. A commercial text mining tool, IBM Miner for Text, was also tested.

Preliminary results from this database revealed an excellent performance of the proposed classification methods. For instance, Support Vector Machine (SVM) achieved 91.12% accuracy. This accuracy result is comparable to human performance in text classification tasks (Weiss et al. 1997). This algorithm outperformed IBM Miner for Text (60.95% accuracy). Preliminary classification results for the other methods tested were: Rocchio (64.71%), Naïve Bayes (58.82%), and K-Nearest Neighbors (49.11%).

A second case study was used to evaluate the performance of automated hierarchical classification methods. Hierarchical classification is even more challenging than flat classification because the accuracy tends to reduce in the lower hierarchical levels. This usually happens because it is more difficult to differentiate the classes at the lower levels since they contain fewer training documents and the documents are more similar. The database selected for this case study was the Sweet's Product Marketplace (Sweets, 2002). This database stores data from over 10,700 manufacturers and 61,300 products for the construction industry. In this database, construction products are classified using the hierarchical structure of CSI MasterFormat (MasterFormat 1995). This study was conducted using some documents extracted from 3 CSI MasterFormat Divisions of this database. Support Vector Machines were used as the classification algorithm. Preliminary results indicated an average accuracy of 94.60% for the first hierarchical level, 89.55% for the second level, and 87.45% for the third. Figure 2 presents the accuracy results for this case study.

CSI MASTERFORMAT CODE	CLASS NAME	HIERARCHICAL LEVEL	TEST CASES			CLASSIFICATION ACCURACY (%)
			TOTAL	RIGHT	WRONG	
3000	Concrete	1	284	268	16	94.37
6000	Wood_and_plastics	1	348	327	21	93.97
16000	Electrical	1	239	229	10	95.82
	LEVEL 1		871	824	47	94.60
3050	Basic_materials_and_methods	2	105	78	27	74.29
3100	Forms_and_accessories	2	36	32	4	88.89
3200	Concrete_reinforcing	2	11	11	0	100.00
3400	Precast_concrete	2	31	28	3	90.32
3500	Cementitious_decks_and_underlayment	2	37	31	6	83.78
3600	Grout	2	11	11	0	100.00
3900	Restoration_and_cleaning	2	47	39	8	82.98
6050	Basic_materials_and_methods	2	50	45	5	90.00
6100	Rough_carpentry	2	56	52	4	92.86
6400	Architectural_woodwork	2	151	136	15	90.07
6600	Plastic_fabrications	2	67	63	4	94.03
16100	Wiring_methods	2	65	64	1	98.46
16500	Lighting	2	99	96	3	96.67
16700	Communications	2	57	51	6	89.47
16800	Sound_and_video	2	19	17	2	89.47
	LEVEL 2		842	754	88	89.55
3410	Plant_precast_structural_concrete	3	18	15	3	62.50
3450	Plant_precast_architectural_concrete	3	13	9	4	69.23
3910	Concrete_cleaning	3	10	9	1	90.00
3920	Resurfacing	3	21	15	6	71.43
3930	Rehabilitation	3	17	13	4	76.47
6060	Wood	3	14	13	1	92.86
6080	Factory_applied_coatings	3	20	20	0	100.00
6090	Fastenings	3	8	8	0	100.00
6130	Heavy_timber_construction	3	17	16	1	94.12
6150	Wood_decking	3	12	11	1	91.67
6410	Custom_cabinets	3	60	58	2	96.67
6420	Paneling	3	18	16	2	88.89
6430	Wood_stairs_and_railing	3	34	34	0	100.00
6440	Ornaments	3	38	35	3	92.11
16120	Conductors_and_cables	3	28	28	0	100.00
16130	Raceway_and_boxes	3	32	32	0	100.00
16510	Interior_luminaires	3	20	19	1	95.00
16520	Exterior_luminaires	3	24	16	8	66.67
16550	Special_purpose_lighting	3	28	16	12	57.14
16710	Communications_circuits	3	26	19	7	73.08
16720	Telephone_and_intercommunication_equipment	3	12	9	3	75.00
	LEVEL 3		470	411	59	87.45

Figure 2: Hierarchical Classification Results

CONCLUSIONS

A large percentage of the communications exchanged and documents stored in inter-organizational construction management information systems are based on textual information. Automatic classification of these documents can be used to improve the management of information flows among project organizations that use these systems. These automated classification methods can also be used to support the implementation of pull techniques for information flow management.

In this paper, the prototype of a construction document classification system was presented. The system automates the steps involved in the document classification process. Issues of class scalability and support for hierarchical classification were considered during its implementation. The system supports the generation of classification models for construction projects. After creating these models, construction documents can be easily and quickly classified according to the user-defined project components.

Two case studies were conducted to verify the feasibility of the proposed approach. The first case study tested the performance of text classification algorithms using a

building project database as testbed. The results for different classification methods were presented. Support Vector Machine was the algorithm that gave the best results. Its accuracy performance was comparable to human-based document classification. The second case study analyzed the classification accuracy for hierarchical classification structures. A construction products' database, originally classified according to a hierarchical structure, was used in this analysis. The preliminary results presented were very promising, demonstrating the potential and applicability of automated document classification methods for the development of pull information systems.

REFERENCES

Aalami, F.B., Fischer, M. and Kunz, J.C. (1998) "AEC 4D-CAD production model: definition and automated generation." *CIFE WP 052*.

aecXML (1999). aecXML < http://www.aecxml.org> (Out 12, 2000).

Ballard, G. (2000). *Lean Construction Institute: Research Agenda*. Lean Construction Institute. July, 2000.

Brüggemann, B. M., Holz, K. -P., and Molkenthin, F. (2000) "Semantic documentation in engineering". *Proceedings. of the ICCCBE-VIII*, Palo Alto, CA, August, 2000, 828-835.

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.

Fruchter, R. (1999) "A/E/C Teamwork: a collaborative design and learning space." *Journal of Computing in Civil Engineering*, 13(4), 261-269.

Han, J. and Kamber, M. (2001) *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

IAI (1996). International Alliance for Interoperability. < http://iaiweb.lbl.gov/> (Oct 12, 2000)

Koskela, L. (2000). An Exploration Towards a Production Theory and its Application to Construction. PhD Dissertation, VTT Building Technology, Espoo, Finland. 296 p., VTT Publications: 408.

Kosovac, B., Froese, T., and Vanier, D. (2000). "Integrating heterogeneous data representations in model-based AEC/FM systems." *Proceedings of CIT 2000*, Reykjavik, Iceland, 1, 556-566.

Maher, M.L. and Simoff, S.J. (1998) "Ontology-based multimedia data mining for design information retrieval" *Proc. of Computing in Civil Engineering*, ASCE, pp. 212-223.

MasterFormat (1995). *MasterFormat 1995 Edition*. Construction Specifications Institute.

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 Buckley C. (1988) "Term weighting approaches in automatic text retrieval.", *Information Processing and Management*, 2(5), 513-523.

Scherer, R. J. and Reul, S. (2000) "Retrieval of project knowledge from heterogeneous AEC documents". *Proceedings of the ICCCBE-VIII*, Palo Alto, Ca, August, 2000, 812-819.

Sebastiani, F. (1999) "Machine learning in automated text categorisation." *Technical Report IEI-B4-31-1999*, Istituto di Elaborazione dell'Informazione, CNR, Pisa, Italy.

Soibelman, L. and Kim, H. (2002) "Generating construction knowledge with Knowledge Discovery in Databases" *Journal of Computing in Civil Engineering*, ASCE, 16(1), 39-48.

Soibelman, L. and Peña-Mora, F. (2000). "A distributed multi-reasoning mechanism to support the conceptual phase of structural design" *J. Struct. Engineering*, ASCE, 126 (6), 733-742.

Sweet's (2002) Sweet's Product Marketplace <http://sweets.construction.com/default.jsp> (April 3, 2002)

Teicholz, P. (1999). "Vision of future practice." *Berkeley-Stanford Workshop on Defining a Research Agenda for AEC Process/Product Development in 2000 and Beyond*, Stanford, CA.

ToCEE - Towards a Concurrent Engineering Environment Project (2000). "The ToCEE client-server system for concurrent engineering." *Final Report - ESPRIT Project No. 20587*.

Yang, M. C., Wood, W. H. and Cutkosky, M. R. (1998) "Data mining for thesaurus generation in informal design information retrieval". *Proc. International Computing Congress*, 189-200.

Yang, Y. and Pedersen, J.O. (1997) "A comparative study on feature selection in text categorization". *Proceedings of ICML-97*, Nashville, TN, 412-420.

Weiss, S. A., Kasif, S., Brill, E. (1997) "Text classification in USENET newsgroups: A progress report." Department of Computer Science. The Johns Hopkins University. April 1997,

Wood, W.H. (2000) "The development of modes in textual design data". *Proceedings of the ICCCBE-VIII*, Palo Alto, CA, August, 2000, 882-889.

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