Data Analysis on Complicated Construction Data Sources: Vision, Research, and Recent Developments

Lucio Soibelman[1], Jianfeng Wu[1], Carlos Caldas[2], Ioannis Brilakis[3], and Ken-Yu Lin[4]

[1] Department of Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA
{lucio, jianfen1}@andrew.cmu.edu
[2] Department of Civil, Architecture and Environmental Engineering, University of Texas at Austin, Austin, TX 78712, USA
caldas@mail.utexas.edu
[3] Department of Civil and Environmental Engineering, University of Michigan Ann Arbor, MI 48109, USA
brilakis@engin.umich.edu
[4] Ming-Jian Power Corporation, Taipei, Taiwan
kengyuli@hotmail.com

Abstract. Compared with construction data sources that are usually stored and analyzed in spreadsheets and single data tables, data sources with more complicated structures, such as text documents, site images, web pages, and project schedules have been less intensively studied due to additional challenges in data preparation, representation, and analysis. In this paper, our definition and vision for advanced data analysis addressing such challenges are presented, together with related research results from previous work, as well as our recent developments of data analysis on text-based, image-based, web-based, and network-based construction sources. It is shown in this paper that particular data preparation, representation, and analysis operations should be identified, and integrated with careful problem investigations and scientific validation measures in order to provide general frameworks in support of information search and knowledge discovery from such information-abundant data sources.

1 Introduction

With the ever-increasing application of new information technologies in the construction industry, we have seen an explosive increment of data in both volume and complexity. Some research efforts have been focused on applying data mining tools on construction data to discover novel and useful knowledge, in support of decision making by project managers. However, a majority of such efforts were focused on analyzing electronic databases, such as productivity records and cost estimates available in spreadsheets or single data tables. Other construction data sources, such as text documents, digital images, web pages, and project schedules were less intensively studied due to their complicated data structures, even though they also carry important and abundant information from current and previous projects.

Such a situation may thwart further efforts in improving information search or lessons learning in support of construction decision making. For example, a productivity

I.F.C. Smith (Ed.): EG-ICE 2006, LNAI 4200, pp. 637 – 652, 2006.

database could help project planners obtain an approximate range of productivity values for a specific activity, say, building a concrete wall based on historical records. However, other contextual information influencing the individual values of previous activities may not be available in the database for various reasons: specifications for different walls might be available only in text formats; correlations between wall-building activities and other preceding/succeeding/parallel jobs are stored in schedules; and pictures verifying such correlations, e.g., space conflicts, could be buried in hundreds of, even thousands of digital images taken in previous projects. This demonstrates the need for the development of advanced tools for effective and efficient information search and knowledge discovery from these texts, schedules, images, so that construction practitioners could make better informed decisions on more comprehensive data sources.

In this paper, our vision for data analysis on complicated construction data sources is first introduced, including its definition, importance, and related issues. Previous research efforts relevant to data analysis on various construction data sources are reviewed. Our recent work in graphical analysis on network-based project schedules is also shown, and initial results from this ongoing research are demonstrated for further discussions.

2 Definition and Vision

Traditionally, research in construction data analysis has been focused on transactional databases, in which each object/instance is represented by a list of values for a given set of features in a data table or a spreadsheet. Comprehensive Knowledge Discovery in Databases (KDD) processes, as well as many statistical and machine learning algorithms, have been introduced and applied in previous research [1,2].

In reality, there are a variety of construction data sources besides those available in spreadsheets or data tables, including semi-structured or unstructured text documents such as contracts, specifications, change orders, requests for information, and meeting minutes; unstructured multimedia data such as 2D or 3D drawings, binary pictures, as well as audio and video files; and structured data for specific applications that need more complicated data structures for storage and analysis than single data tables, e.g., network-based schedules. Currently a large percentage of project information is still kept in these sources for two major reasons. First, the large numbers of features, or high dimensionalities of such information make it inefficient to store and manage related data, such as text files with hundreds of words or images with thousand of pixels, in transactional databases. Second, information contained in the original data formats, such as graphical architectures of schedules and texture information of images could be lost if they are transferred into a single data table or spreadsheet. This situation requires that KDD processes, originally developed for transactional databases, be adapted to be applicable for more complicated construction data sources.

Compared with transactional construction databases that have been analyzed in a number of existing research efforts, construction data sources with complicated data structures are relatively less studied for information retrieval or lessons learning, even though they contain important and abundant historical information from ongoing and previous projects. As a result, useful data could be buried in large volumes of records

due to difficulties in extracting relevant files for various purposes in a timely manner, while many information-intensive historical data could be left intact after projects are finished, without being further analyzed to improve practices in future projects.

Three major challenges could be attributed to such a relative scarce of related research. First, in addition to data preparation operations for general KDD processes [3], special attention should be taken to filtering out incorrect or irrelevant data from available construction data sources, while ensuring integrity and completeness of original information and further integrating it with other contextual information from other data sources if necessary. Second, extra efforts are needed to reorganize project data, originally recorded in application-oriented formats, into analysis-friendly representations that support efficient and effective KDD implementation. Third, data analysis techniques, which were initially developed to identify patterns and trends from single tables or spreadsheets, should be adapted in order to learn useful knowledge from text-based, image-based, web-based, network-based, and other construction data sources.

In our vision, a complete framework should be designed, developed, and validated for data analysis on particular types of complicated construction data sources. Three essential parts should be included in such a framework. First, careful problem investigations should be done in order to understand data analysis issues and related work in previous research. Second, detailed guidelines for special data preparation, representation, and analysis operations are also needed to ensure generality and applicability of the developed framework. And eventually, scientific criteria and techniques should be employed to evaluate validity of data analysis results and feasibility of extending such a framework to same or similar types of construction data sources.

3 Related Research Areas and Previous Work

In this section, previous research for analysis of text-based, web-based, and image-based construction data are presented, reviewing work done by other researchers, together with frameworks developed in our research group to support advanced data analysis on complicated construction data sources.

3.1 Text Mining for Management of Project Documents

In construction projects, a high percentage of project information is exchanged using text documents such as contracts, change orders, field reports, requests for information, and meeting minutes, among many others [4]. Management of these documents in model-based project management information systems is a challenging task due to difficulties in establishing relations between such documents and objects. Manually building the desired connections is impractical since these information systems typically store thousands of text documents and hundreds of model objects. Current technologies used for project document management, such as project websites, document management systems, and project contract management systems do not provide direct support for this integration. There are some critical issues involved, most of them due to the large number of documents and project model objects and differences in vocabulary. Search engines based on term match are available in many information systems used in construction. However, the use of these tools also has some limitations in cases

where multiple words share the same meaning, where words have multiple meanings, and where relevant documents don't contain the user-defined search terms [4].

Recent research addressed some issues related to text data management. Information systems and algorithms were designed to improve document management [5,6,7]; controlled vocabularies (thesauri) were used to integrate heterogeneous data representations including text documents [8]; various data analysis tools were also applied on text data to create thesauri, extract hierarchical concepts, and group similar files for reusing past design information and construction knowledge [9,10,11]. In one of the previous studies conducted by our research group, a framework was devised to explore the linguistic features of text documents in order to automatically classify, rank, and associate them with objects in project models [4,12]. This framework involved several essential steps as discussed in our vision for data analysis on complicated construction data sources:

- Special **Data Preparation** operations for text documents were identified, such as transferring text-based information into flat text files from their original formats, including word processors, spreadsheets, emails, and PDF files; removing irrelevant tags and punctuations in original documents; removing stop words that are too frequently used to carry useful information for text analysis like articles, conjunctions, pronouns, and prepositions; and finally, performing word stemming to get rid of prefixes and/or suffixes and group words that have the same conceptual meanings.
- The preprocessed text data were transformed into a specific **Data Representation** using a weighted frequency matrix $A=\{a_{ij}\}$, where a_{ij} was defined as the weight of a word i in document j. Various weighting functions were investigated based on two empirical observations regarding text documents: 1) the more frequent a word is in a document, the more relevant it is to topic of the document; and 2) the more frequent a word is throughout all documents in the collection, the more poorly it differentiates between documents. By selecting and applying appropriate weighting functions, project documents were represented as vectors in a multi-dimensional space. Query vectors could then be constructed to identify similar documents based on similarity measures such as Euclidian distance and the cosine between vectors.
- **Data Analysis** tools were applied to build document classifiers, as well as document retrieval, ranking, and association mechanisms. The proposed methods require the definition of classes in the project model. These classes are usually represented as items of a construction information classification system (CICS) or components of a work breakdown structure (WBS). Classification models are created for each of these classes and project documents are automatically classified. Project document classification creates a higher semantic level that facilitates the identification of documents that are associated to selected objects. This is one of the major characteristics of the proposed methodology and represents a major distinction from existing project document management approaches. In order to retrieve documents that are related to selected objects, data that characterize the object is extracted from the model and used by the retrieval and ranking mechanisms. In the retrieval and ranking phase, all project documents are analyzed, but a higher weight is given to documents belonging to the object's class. In the Association step, documents are linked to the model objects.

The proposed methodology uses both the object's description terms and the object's class as input for the data analysis process. In methods based just on text search, only documents that contain the search terms can be located. Documents that use synonyms would not be found. If only the object's class is considered, some documents that belong to the object's class but are not related to the object would be mistakenly retrieved. By analyzing all model classes, but giving a higher weight to the object's class, related documents in other classes can also be identified.

A prototype system, Unstructured Data Integration System (UDIS), was developed to classify, retrieve, rank, and associate documents according to their relevance to project model objects [12]. Figure 1 in the next page illustrates UDIS.

The validation of the proposed methodology was based on the analysis of more than 25 construction databases and 30,000 electronic documents. The following two measures were used for comparison: recall and precision, which represent the percentage of the retrieved documents that were related to the selected model object, and the percentage of the related documents that were actually retrieved, respectively. Figure 2 in the following page shows how recall and precision are defined in information retrieval problems. The UDIS prototype achieved better performance in both recall and precision than typical project contract management systems, project websites, and commercial information retrieval systems [12].

Some of the major advantages of this framework include: 1) it did not involve manual assignments of metadata to any text documents; 2) it did not require a controlled vocabulary that would only be effective if it is adopted by all users of an information system; and 3) it could provide automated mechanisms to map text documents to project objects using their internal characteristics instead of user-defined search terms.

3.2 Text Mining for Ontology-Based Online A/E/C Product Information Search

As apposed to managing a set of in-house document collection in Section 3.1, online documents, especially those containing Architecture/Engineering/Construction (AEC) product information, have also received increasing attention while more and more manufacturers/suppliers host such information on the Internet. This is because the web provides boundary-less information sources, which could be used by industry practitioners to survey the A/E/C product market for product specification, selection, and procurement applications. Challenges similar to those encountered when dealing with in-house project document collection include the ambiguity of natural languages (e.g., polysemy, synonym, and structural ambiguity), and the lack of consistent product descriptions (e.g., definitions of product properties, domain lexicons for describing the products, and product information formatting) within A/E/C. Besides those, the enormous search space of the Internet, together with industry practitioners' varied ability to conduct effective queries, presents additional difficulties to the search of online product information. Concerns about the effectiveness of general search engines for retrieving domain specific information were also raised [13].

Domain knowledge represented in reusable formats can be of great help to tackle the above mentioned challenges. In knowledge engineering, ontology is used to represent the knowledge-level characterization of a domain. *Taxonomy* serves as the simplest ontology, which organizes a set of control vocabularies into a hierarchical

structure. The employment of machine learning methods on training data sets to generate a classification system represented by hierarchical clusters of grouped project documents [4] is one noble transformed example of domain knowledge utilization. Other applications of domain knowledge represented in the form of taxonomy have also been reported [14,15,16].

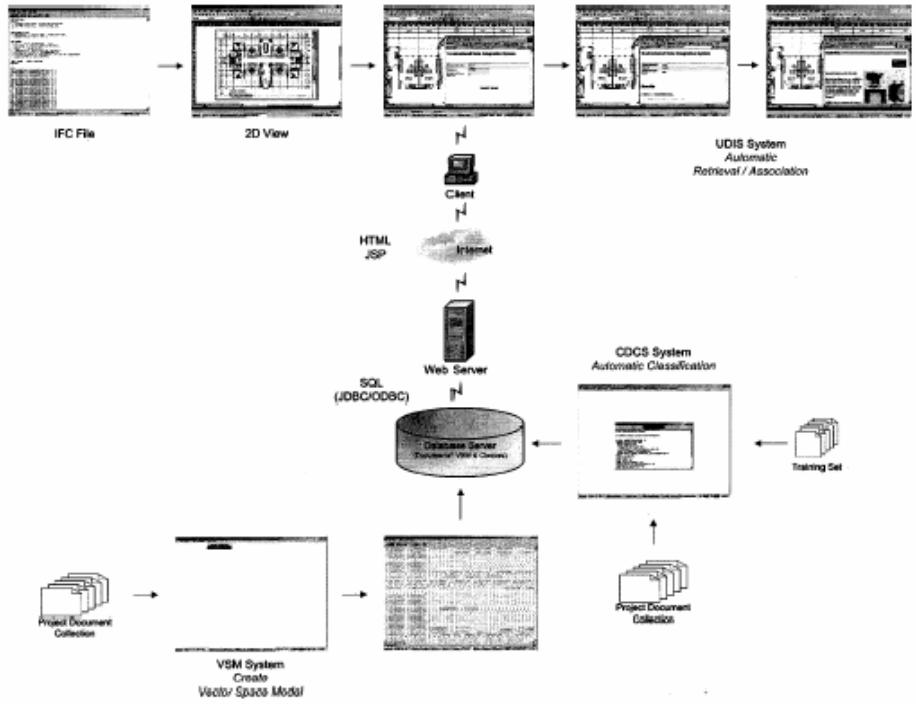

Fig. 1. Unstructured Data Integration System [12]

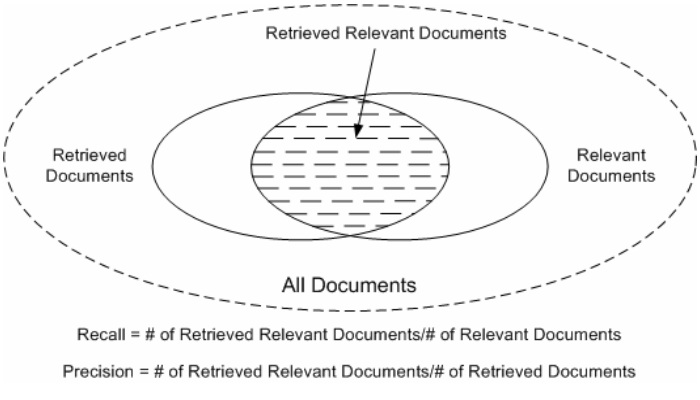

Fig. 2. Definitions of recall and precision in Information Retrieval (IR)

Our research in [17] used *thesaurus* as a more elaborate form of ontology to help organize domain concepts related to a given A/E/C product. This application demonstrated how domain knowledge in the form of a thesaurus was utilized to support data preparation, representation, and analysis to identify highly relevant online product documents for a given product domain. Main steps as defined by our vision for analyzing complicated construction data sources in [17] are described below.

- **Data Preparation** operations were conducted automatically at two levels. Level one leveraged the represented thesaurus to identify a high quality pool of possibly relevant online documents. This was achieved by first expanding a single query into multiple queries (i.e., the growing strategy) and then pooling (i.e., the trimming strategy) the top-k documents returned by a general search engine after the expanded queries were entered as inputs. Level two performed word operations using the linguistic techniques such as removing stop words and stemming, similar to that introduced in Section 3.1.
- The prepared textual data were converted into document vectors for **Data Representation** purposes. The document vectors were assumed to exist in a multi dimensional space whose dimensions were defined by the concepts derived from the product thesaurus. The values of the vectors were defined by a weighted term-frequency matrix $A=\{a_{ij}\}$ where a_{ij} was the weight of a thesaurus term i in document j.
- The **Data Analysis** was achieved by two stages. In the first stage, each document vector was evaluated based on each expanded query, considering the thesaurus terms as well as the Boolean connectors resided within the query. In the second stage, a synthesis evaluation was calculated for a given document vector, considering all the possibly expanded queries with the use of domain thesaurus.

A semi-automated thesaurus generation method was developed in this research in order to supply domain knowledge for retrieving relevant A/E/C online product information. With the aid of domain knowledge in the generated thesaurus, the enormous Web space was successfully reduced into a manageable set of candidate online documents during data preparation. The use of thesauri in data representation also helped solve the language ambiguity problems so that synonyms like “lift” and “elevator” were treated indifferently. In this research for the purpose of validation, five data sets originating from different A/E/C product domains were processed by a developed prototype to return a list of search results that contained highly relevant online product information. To assess the true relevance of the returned search results, each web document was examined and labeled manually as ‘relevant’ or ‘non-relevant’. Then retrieval performance was evaluated accordingly. The performance evaluation focused particularly on the number of distinct product manufacturers identified because the goal of this research is to help A/E/C industry practitioners obtain more product information from the available marketplace. Hence for each data set, the number of distinct product manufacturers derived from 1) the search results generated from the established prototype; 2) the searching results returned by Google; and 3) an existing catalog compiled by SWEETS were compared as displayed in Table 1. It was observed that in most cases, the developed research prototype consistently retrieved more distinct product manufactures than the other two approaches.

Table. 1. Prototype performance comparisons for each of the five data sets [17]

Num of Distinctive Product Manufacturers Discovered	1	2	3	4	5
This research	8	12	28	26	30
Searching via Google	3	7	19	17	36
SWEETS	5	1	2	6	1

Compared with approaches that enforce standardized data representation, the developed framework in [17] has more flexibility in dealing with text-based and heterogeneous A/E/C product specifications on the Internet. The automatic query expansion process in the framework also relieved searchers' burden to structure complex queries using highly related terms. Therefore, an industry practitioner can take advantage of the developed A/E/C domain-specific search engine to survey the virtual product market for making more informed decisions.

3.3 Image Reasoning for Search of Jobsite Pictures

Pictures are valuable sources of accurate and compact project information [18]. It is becoming a common practice for jobsite images to be gathered periodically, stored in central databases, and utilized in project management tasks. However, the volume of images stored in an average project is increasing rapidly, making it difficult to manually browse such images for their utilization. Moreover, the labeling systems used by digital equipment are tweaked to provide distinct labels for all images acquired to avoid overlaps, but they do not record any information regarding the visual content. Consequently, images stored in central databases cannot be queried using conventional data and text search methods and thus, retrievals of jobsite images have so far been dependent on manual labeling and indexing by engineers.

Research efforts have been developed to address the image retrieval concerns described above. A prototype system, for example, was developed by Abudayyeh [19] based on a MS Access database that allowed engineers to manually link construction multimedia data including images with other items in an integrated database. The use of thesauri was also proposed by Kosovak et al. [8] to assist image indexing by enabling users to label images with specific standards. However, both solutions are difficult to use in practice due to the large number of manual operations that is needed to classify images in meaningful ways. Also, besides manual operations that are still involved in both solutions, the issue of how to index images in different ways was not addressed. This issue is important since it is not unusual that an image could be used for multiple purposes, even if it had been taken for just one use. For example, an engineer may have taken a picture of protective measures to prove their existence, but the same picture might contain other useful content needed in the future, such as materials or structural components included in its background [18].

A novel construction site image retrieval methodology based on material recognition was developed to address these issues [18,20]. A similar framework was developed based on our vision for data analysis on complicated construction data sources:

- **Data Preparation:** images from construction site were first preprocessed by extracting basic graphical features, such as intensity, color histograms, and texture presentations using various filtering methods. Features unrelated to image content such as intensity of lights and shades were then removed automatically.
- **Data Representation:** preprocessed images were first divided into separate regions representing different construction objects using appropriate clustering methods; graphical features of each region were further compressed into quantifiable, compact, and accurate descriptors, or signatures. As a result, a binary image was transformed into several clusters, each with a list of values for a given set of feature signatures.
- **Data Analysis:** the feature signatures of each image region were compared with signatures in an image collection of material samples. Using proper threshold and/or distance functions, the actual material of each region was recognized. An example is given in the next page (Figure 3) to show how a site image is decomposed into clusters, represented with signatures, and identified with actual materials in each separated region.

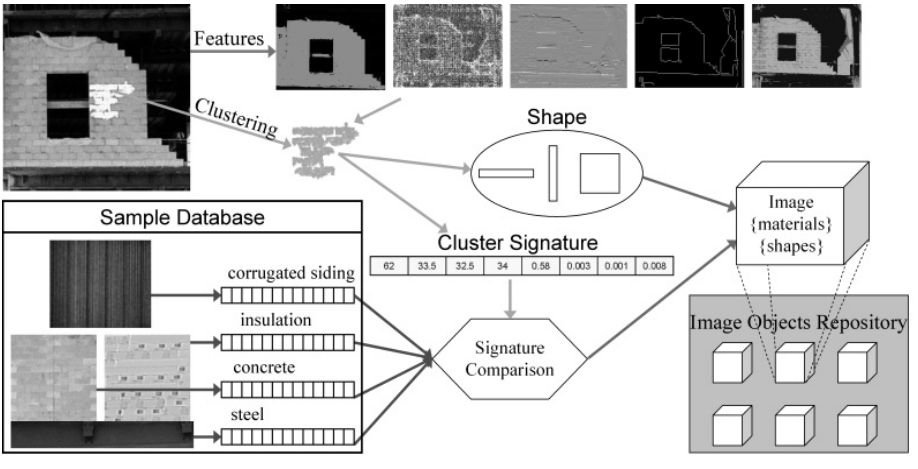

Fig. 3. An example for image data preparation, representation, and analysis [20]

The material information was then combined with other spatial, temporal, and design data to enable automated, fast, and accurate retrieval of relevant site images for various management tasks in a developed system as illustrated below (Figure 4).

As shown in Figure 4, in addition to image attributes (e.g., materials) that can be automatically extracted using the above method, time data can be obtained from digital cameras when image are taken on construction sites; 2D/3D locations can be obtained from positioning technologies; and as-planned and as-built information for construction products can be retrieved from a model-based system or a construction database. By combining all such information together, images in the database can be compared with the query's criteria in order to filter out irrelevant images and rank the selected images according to their relevance. The results are then displayed on the

screen, and the user could easily browse through a very limited amount of sorted images to identify and choose the desired ones for specific tasks. Figure 4 provides a detailed example of retrieving images for a brick wall using the developed prototype. This method was tested on a collection of more than a thousand images from several projects and evaluated with similar criteria like recall and precision in Section 3.1. The results showed that images can be successfully classified according to the construction materials visible within the image content [20].

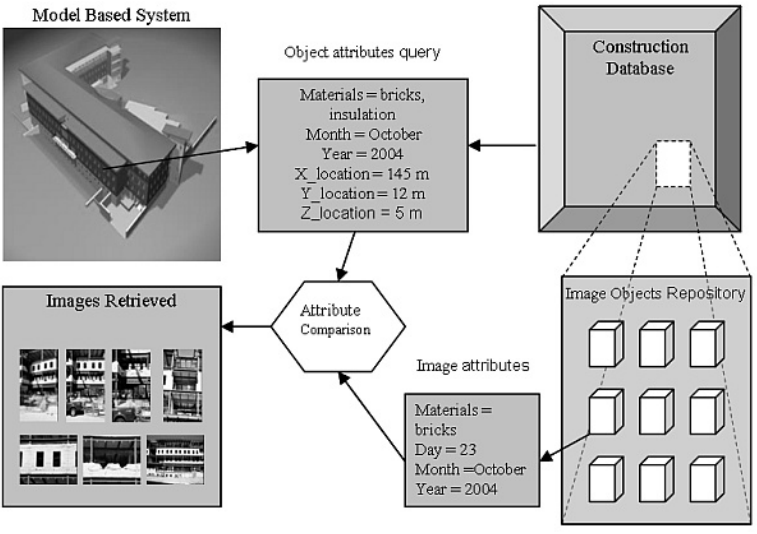

Fig. 4. Image retrieval based on image contents and other attributes [20]

The original contribution of this research work on construction site image indexing and retrieval could be extended to solve other related problems. For example, our research group is currently investigating a new application of image processing, pattern recognition, and computer vision technologies for fully automated detection and classification of defects found in acquired images and video clips of underground pipelines.

A major motivation of our new research work resides on the rapid aging of pipeline infrastructure systems. It is desirable to develop and use more effective methods for assessing the condition of pipelines, evaluating the level of deterioration, and determining the type and probability of defects to facilitate decisions making for maintenance/repair/re-habilitation strategies that will be least disruptive, most cost-effective and safest [21]. Currently, trained human operators are required to examine acquired visual data and identify the detailed defect class/subclass according to the standard pipeline condition grading system, such as Pipeline Assessment and Certification Program (PACP) [22], which makes the process error-prone and labor-intensive. Building on our previous research efforts in image analysis [18] and former research efforts in defect detection and classification of pipe defects [23,24,25], we are studying more advanced data preparation, representation, analysis solutions to boost

automated detection and classification of pipe defects based on improved problem investigation, feature selection, data representation, and multi-level classifications of pipeline images.

4 A Recent Development in Scheduling Data Analysis

This section presents our recent work applying the same vision, which has been successfully applied for analysis of unstructured text and image data, for analysis of network-based project schedules. A preliminary study including problem investigation, data preparation, data representation, data analysis, and pattern evaluations are detailed below.

4.1 Problem Investigation

Since computer software for Critical Path Method, or CPM-based scheduling started to be used in the construction industry decades ago, planning and schedule control history of previous projects has been increasingly available in computerized systems in many companies. However, there is a current lack of appropriate data analysis tools for scheduling networks with thousands of activities and complicated dependencies among them. As a result, schedule historical data are left intact without being further analyzed to improve future practices. At the same time, companies are still relying on human planners to make critical decisions in various scheduling tasks manually and in a case-by-case manner. Such practices, however, are human-dependent and error-prone in many cases due to subjective, implicit, and incomplete scheduling knowledge that planners have to use for decision making within a limited timeframe.

Many research efforts have been focused on improving current scheduling practices. Delay analysis tools were applied to identify fluctuations in project implementation, responsible parties for the delays, and predict possible consequences [26,27]; machine learning tools like genetic algorithms (GA) were applied to explore optimal or near-optimal solutions for resource leveling and/or time-cost tradeoff problems, by effectively searching only a small part of large sample space for construction alternatives [28,29,30]; simulation and visualization methods were developed on domain-specific process models to identify and prevent potential problems during implementation [31,32]; and knowledge-based systems were also introduced to collect, process, and represent knowledge from experts and other sources for automated generation and reviewing of project schedules [33,34]. Such research efforts helped improve efficiency and quality of scheduling work in many aspects, but none of them addressed the above "data rich but knowledge scarce" issue, i.e., no research described here has worked on learning explicit and objective knowledge from abundant planning and schedule control history of previous projects.

Our research is intended to address this issue related to scheduling knowledge discovery from historical network-based schedules, so that more explicit and objective patterns could be identified from previous schedules in support of project planners' decision making. A case study was implemented on a project control database to identify special problems in preparing, representing, and analyzing project schedules for this purpose.

4.2 Data Preparation

The project control database for this case study was collected from a large capital facility project. Three data tables were used in the scheduling analysis: one data table with detailed information about individual activities; one with a complete list of predecessors and successors of dependencies among the activities; and a third table with historical records about activities that were not completed as planned during implementation, and reasons for their non-completions.

In this step, we first separated nearly 21,000 activities into about 2,600 networks based on their connectivity. Another special data preparation task identified was the removal of redundant dependencies in networks [35]. As shown in Figure 5, a dependency from activity X to Y is redundant if there is another activity Z, such that Z is succeeding to X, and Y is succeeding to Z, directly or indirectly. Obviously, the dependency from X to Y is unnecessary because Y can not be started immediately after X anyway.

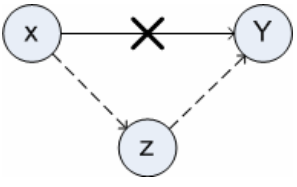

Fig. 5. An example for redundant dependencies in scheduling networks

4.3 Data Representation

Architectures of scheduling networks may influence their risks and performance in implementation to a great extent. A challenge here was to describe such networks in a concise and precise way, so that networks having similar architectures and performance could be abstracted into same or close descriptions. On the other hand, it would also be difficult, if not impossible, to identify common patterns among all 2,600 scheduling networks with varied size and complicated graphical architectures by just reviewing them visually or in their original formats. Like the developed solutions in data analysis for text-based and image-based construction data sources, a general data representation is necessary to transform scheduling information into appropriate formats for desirable data analysis.

Intuitively, how a network is split into branches or converged from branches could be a good indicator of its reliability during implement. A possible explanation for such an intuition is that such splits or converges generate or eliminate parallelisms, if we assume that parallel implementation is a major cause for resource conflicts, and thus variability during actual implementation. We used such an intuition as our major hypothesis in this case study, and developed a data representation to describe scheduling networks in an abstract way focusing on how a work flow goes through it by splitting/converging, and finally arriving at its end. In this study, we developed a novel abstraction process comprising of two major stages:

– Recursive Divisions in a top-down manner: the network in Figure 6 could be divided into two sequential sub-networks, N1/N2, so that N1 is preceding to N2 only through activity B, and N2 could be further divided into two parallel sub-networks, N21/N22, which only have common starting task B and ending tasks C; such divisions could continue until a sub-network eventually consists of only 'atomic' sequences of tasks without any branches (e.g., all the three sequences from A to B).

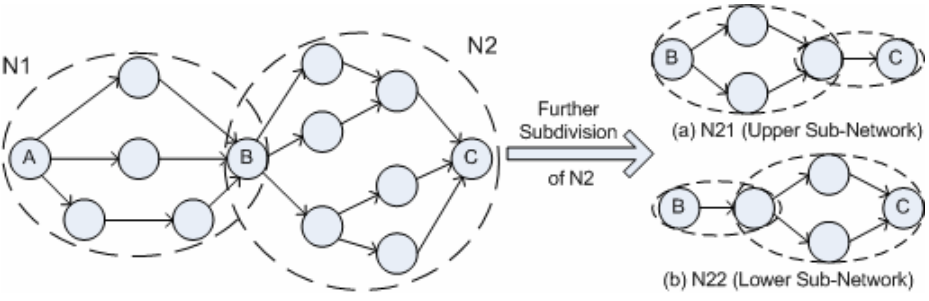

Fig. 6. An example for recursive division of a scheduling network

– Type description in a bottom-up manner: in this stage, task identifications were removed since we only concerned the abstract description of the network architecture. In a reverse direction to recursive divisions, the type description for a network/sub-network could be obtained by following the rules illustrated in Figure 7.

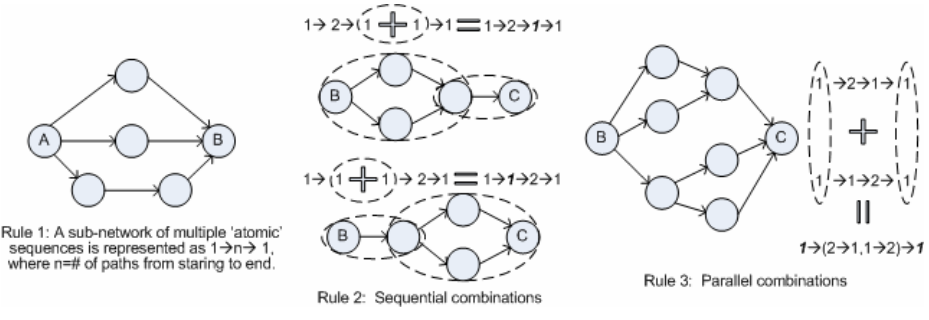

Fig. 7. Examples for type descriptions rules

4.4 Data Analysis and Primary Pattern Evaluations

Type descriptions for all 2,600 networks in this study were generated using a computer program developed on a Matlab® 7.0 platform. By comparing type descriptions of networks with their probabilities of non-conformances, i.e., having non-completion tasks during implementation, some interesting patterns were found. For example, in this project, the non-conformance rate of networks that started with a single task (type description: "1→n→…") was significantly lower than that of networks started with

multiple tasks (type description: "n→1→…"). This pattern is simple in its statement, but it could be hard to discover if the complicated network structures had not been abstracted into precise and concise descriptions. And it makes sense considering that a schedule that starts with multiple parallel tasks at the beginning could be more risky due to lack of sufficient preparations by managerial team at an early stage. Also, comparisons between non-completion rates of activities at different positions turned out some meaningful results. The non-completion rate of tasks connecting sequential or parallel sub-networks, i.e., those with more than one preceding and/or succeeding tasks, was significantly larger than that of tasks among 'atomic' sequences, i.e., those with at most one preceding and succeeding activities. It also verify our previous hypothesis made during data representation that how a network is split into branches or converged from branches could be a good indicator of reliability during implement for specific networks and activities.

These initial results also demonstrated the potentials of extending our current work in many directions: 1) the recursive division method makes it possible to decompose large and complicated scheduling networks into multiple levels of components for further representation of project planning and schedule control history; 2) the abstract descriptions could be viewed as extracted graphical features of scheduling networks, which could be integrated with other features of activities and dependencies for identifying other general and useful patterns in scheduling data; and 3) the complete process of scheduling data preparation, representation, and analysis provided primary insights and experiences for the following development of this research on a larger scale.

5 Conclusion and Future Work

Complicated construction data sources are different from transactional databases in their information contents, data structures, and storage sizes. In this paper, text-based, image-based, web based, and network-based construction data were used as examples to present our vision for possible advanced data analysis on these types of data sources. Several required steps essential for achieving this vision were presented, including problem investigation to define relevant issues; data preparation to remove features not carrying useful information; data representation to transform data sources into new precise and concise descriptions; and proper data analysis for information search or for extraction of lessons learned. The success obtained by the implementation of the prototypes developed to test our vision suggests that further research should be developed to extend the results obtained and to allow the development of the required tools to deal with other types of construction data sources, such as audio records of construction meetings (e.g., for automatically extracting and classifying useful information based on voice recognition) or video data from construction sites (e.g., automated estimations of overall and/or specific productivity values of construction workers).

Further research efforts on advanced data analysis on complicated construction data sources could be integrated with a broad range of related work. For example, retrieved data or identified patterns from such data sources could be fed into existing research in construction decision support systems (DSS) to enhance current solutions

with improved access to historical data and/or knowledge; such research efforts may also be combined with current data modeling and information management research to integrate construction data sources with various formats and broad contexts for data retrieval, queries, and analysis; and it is expected that new challenges and pattern recognitions tasks could be identified and inspire innovative research methods in data representation and data mining communities as well.

Acknowledgements

The authors would like to thank US National Science Foundation for its support under Grants No. 0201299 (CAREER) and No. 0093841, and to all industrial collaborators who provided the data sources for our previous and current studies.

References

1. Buchheit, R.B., Garret J.H. Jr., Lee, S.R. and Brahme, R.: A Knowledge Discovery Case Study for the Intelligent Workplace. *Proc. of the 8th Int. Conf. on Comp. in Civil and Building Eng.*, Stanford, CA (2000), 914-921
2. Soibelman, L. and Kim, H.: Data Preparation Process for Construction Knowledge Generation through Knowledge Discovery in Databases. *J. of Comp. in Civil Eng.*, ASCE, 1(2002), 39-48.
3. Soibelman, L, Kim, H., and Wu, J.: Knowledge Discovery for Project Delay Analysis. *Bauingenieur*, Springer VDI Verlag, Feburary 2005
4. Caldas, H.C., Soibelman, L., and Han, J.: Automated Classification of Construction Project Documents. *J. of Comp. in Civil Eng.*, ASCE, 4(2002), 234-243
5. Hajjar, D. and AbouRizk, M.S.: Integrating Document Management with Project and Company Data. *J. of Comp. in Civil Eng.*, ASCE, 1(2000), 70-77.
6. Zhu, Y., Issa, R.R.A., and Cox, R.F.: Web-Based Construction Document Processing via Malleable Frame. *J. of Comp. in Civil Eng.*, ASCE, 3(2001), 157-169.
7. Fruchter, R., and Reiner, K.: Model-Centered World Wide Web Coach. *Proc. of 3rd Congress on Comp. in Civil Engineering*, ASCE, Anaheim, CA (1996), 1-7.
8. Kosovac, B., Froese, T.M. and Vanier, D.J.: Integration Heterogeneous Data Representations in Model-Based AEC/FM Systems. *Proc. of CIT 2000*, Reykjavik, Iceland (2000), 556-567
9. Yang, M.C., Wood, W.H., and Cutkosky, M.R.: Data Mining for Thesaurus Generation in Informal Design Information Retrieval. *Proc. of Int. Congress on Comp. in Civil Eng.*, ASCE, Boston, MA (1998), 189-200
10. Scherer, R.J. and Reul, S.: Retrieval of Project Knowledge from Heterogeneous AEC Documents. *Proc. of the 8th Int. Conf. on Comp. in Civil and Building Eng.*, Stanford, CA (2000), 812-819
11. Wood, W.H.: The Development of Modes in Textual Design Data. *Proc. of the 8th Int. Conf. on Comp. in Civil and Building Eng.*, Stanford, CA (2000), 882-889
12. Caldas, H.C., Soibelman, L., and Gasser, L.: Methodology for the Integration of Project Documents in Model-Based Information Systems. *J. of Comp. in Civil Eng.*, ASCE, 1(2005), 25-33
13. Dewan, R., Freimer, M., and Seimann, A.: Portal Kombat: The Battle between Web Pages to become the Point of Entry to the World Wide Web. *Proc. of the 32nd Hawaii Intl. Conf. System Science*, IEEE, Los Alamitos, CA, USA (1999)

14. El-Diraby, T.A., Lima, C., and Feis, S.: Domain Taxonomy for Construction Concepts: Toward a Formal Ontology for Construction Knowledge. *J. of Comp. in Civil Eng*., ASCE, 4 (2005), 394–406.
15. Lima, C., Stephens, J., and Böhms, M.: The bcXML: Supporting Ecommerce and Knowledge Management in the Construction Industry. *ITCON*, (2003), 293–308.
16. Staub-French, S., Fischer, M., Kunz, J., Paulson, B., and Ishii, K.: An Ontology for Relating Features of Building Product Models with Construction Activities to Support Cost Estimating", CIFE Working Paper #70, July 2002, Stanford University.
17. Lin, K. and Soibelman, L.: Promoting Transactions for A/E/C Product Information. *Automation in Construction*, Elsevier Science, in print
18. Brilakis, I. and Soibelman, L.: Material-Based Construction Site Image Retrieval. *J. of Comp. in Civil Eng.*, ASCE, 4(2005), 341-355
19. Abudayyeh, O.: Audio/Visual Information in Construction Project Control. *J. of Adv. Eng. Software*, Elsevier Science, 2(1997), 97-101
20. Brilakis, I. and Soibelman, L.: Multi-Modal Image Retrieval from Construction Databases and Model-Based Systems, *J. of Constr. Eng. and Mgmt.*, ASCE, July 2006, in print
21. Water Environment Federation (WEF): Existing Sewer Evaluation and Rehabilitation. *ASCE Manuals and References on Eng. Practices, Ref. No. 62*, Alexandria, VA, 1994.
22. Pipeline Assessment and Certification Program (PACP) Manual, Second Edition Reference Manual, NASSCO, 2001
23. Chae, M.J. and Abraham, D.M.: Neuro-Fuzzy Approaches for Sanitary Sewer Pipeline Condition Assessment. *J. Comp. in Civil. Eng,* ASCE, 1(2001), 4–14
24. Sinha, S.K. and Fieguth, P.W.: Automated Detection of Cracks in Buried Concrete Pipe Images. *Automation in Construction*, Elsevier Science, 1(2005), 58-72
25. Sinha, S.K. & Fieguth, P.W.: Neuro-Fuzzy Network for the Classification of Buried Pipe Defects. *Automation in Construction*, Elsevier Science, 1(2005), 73-83
26. Hegazy, T. and Zhang, K.: Daily Windows Delay Analysis. *J. of Constr. Eng. and Mgmt.*, ASCE, 5(2005), 505-512
27. Shi, J.J., Cheung, S.O. and Arditi, D.: Construction Delay Computation Method. *J. of Comp. in Civil Eng.*, ASCE, 1(2001), 60-65
28. Hegazy, T. and Kassab, M.: Resource Optimization Using Combined Simulation and Genetic Algorithms. *J. of Constr. Eng. and Mgmt.*, ASCE, 6(2003), 698-705
29. Feng, C., Liu, L. and Burns, S.A.: Using Genetic Algorithms to Solve Construction Time-Cost Trade-Off Problems. *J. of Comp. in Civil Eng.*, ASCE, 3(1997), 184-189
30. El-Rayes, K. and Kandil, A.: Time-Cost-Quality Trade-off Analysis for Highway Construction. *J. of Constr. Eng. and Mgmt.*, ASCE, 4(2005), 477-486
31. Akinci, B., Fischer, M. and Kunz, J.: Automated Generation of Work Spaces Required by Construction Activities. *J. of Constr. Eng. and Mgmt.*, ASCE, 4(2002), 306-315
32. Kamat, V.R. and Martinez, J.C.: Large-Scale Dynamic Terrain in Three-Dimensional Construction Process Visualizations. *J. of Comp. in Civil Eng.*, ASCE, 2(2005), 160-171
33. Dzeng, R. and Tommelain, I.D.: Boiler Erection Scheduling Using Product Models and Case-Based Reasoning. *J. of Constr. Eng. and Mgmt.*, ASCE, 3(1997), 338-347
34. Dzeng, R. and Lee, H.: Critiquing Contractors' Scheduling by Integrating Rule-Based and Case-Based Reasoning. *Automation in Construction*, Elsevier Science, 5(2004), 665-678
35. Kolisch R., Sprecher A. and Drexl A.: Characterization and Generation of a General Class of Resource-Constrained Project Scheduling Problems. *Mgmt. Science*, (1995), 1693-1703