Automation in Construction 51 (2015) 78–91

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

GA-based multi-level association rule mining approach for defect analysis in the construction industry

Yusi Cheng [a], Wen-der Yu [b,*], Qiming Li [a]

[a] Department of Construction and Real Estate, School of Civil Engineering, Southeast University, Nanjing, JS 210096, PR China
[b] Department of Construction Management, Chung Hua University, Hsinchu 300, Taiwan, ROC

ARTICLE INFO

Article history:
Received 17 May 2014
Received in revised form 29 October 2014
Accepted 19 December 2014
Available online 12 January 2015

Keywords:
Association rule
Data mining
Construction defects
GA
Concept hierarchy

ABSTRACT

In construction industry, work defects yield time and cost overruns of construction projects and also cause disputes between project participants during construction and operation phases. To date, there hasn't yet been an adequate analytical model to extract useful information from the database of construction defects. The information represented in the form of association rules could enhance quality management via defect prediction and causation analysis. This paper proposes a Genetic Algorithm (GA)-based approach that incorporates the concept hierarchy of construction defects to discover multi-level patterns of defects from the database of defects in the Chinese construction industry during 2000 to 2010. First, the domain knowledge of construction defect is incorporated into a concept hierarchy to adjust mining items at different levels according to the data sparseness and the interestingness of a rule. Second, a GA-based approach is proposed to generate interesting association rules without specified threshold of minimum confidence, taking advantage of the searching capability of GA. Finally, the redundant rules in the mining results are pruned by post-processing method. A test case is selected to demonstrate the feasibility and applicability of the proposed approach within the problem domain. It is concluded that the proposed method provided an effective tool to discover useful knowledge hidden in historical defect cases. The discovered knowledge indicating relationships between defects and defect causes enables project managers to make strategies for estimating and reducing defects.

1. Introduction

In construction industry, reduction of work defects has always been an important task during the construction and operation phases [1–3]. Yielding severe consequences such as reworks, schedule delays, cost overruns, and disputes, construction defects are considered as one of the most common problems that construction projects suffer from [1,4–9]. As many defect data are nominal, such as their characteristics, causes, and consequences, traditional quantitative analytical methods may not be useful in identifying the root causes resulting in the defects. An adequate analytical tool to extract useful information from previous construction defect database hasn't yet been developed, making the current practice of defect management ineffective and inefficient.

While developing an effective defect management model, it is important to identify the key factors resulting in defects so that construction engineers and managers are able to reduce defects by controlling the factors. Previous works on defect analysis focused on identifying factors from design, materials, workmanship, environment and maintenance by statistical methods [2,10–12]. Such methods of defect analysis encountered two limitations in digging out effective defect management strategies: first, proposing hypotheses or developing models for defect mechanism in these methods is time-consuming and subjective when the data sets of defects are large and complicated; second, as a defect is usually caused by a combination of various factors, it is difficult to identify hidden patterns by traditional methods.

The hidden relationships among defect factors mentioned above are usually ignored due to inadequate analytical tools adopted in traditional quality management practice [13,14]. According to the current construction practice in China, work defects are documented and corrected or repaired by inspectors of individual disciplines. Cross-discipline analyses of construction defects have rarely been conducted due to the massive volume of quality inspection data and the fragmented trade structure of construction project organization. There is a need to develop an approach for automated extraction of association rules from massive defect data. The extracted defect rules can be utilized to predict unknown defects based on identified defects and to assure the quality of components where defect might occur by taking preventive actions in advance. Consequently, the quality management of construction projects can be improved.

Association rule mining has been applied to discover useful patterns in large databases by previous researches [14–19]. However, there were also challenges that need to be conquered before a successful application can be implemented in the construction industry:

* Corresponding author. Tel.: +886 3 5186748; fax: +886 3 5370517.
E-mail address: wenderyu@chu.edu.tw (W. Yu).

http://dx.doi.org/10.1016/j.autcon.2014.12.016
0926-5805/

- In traditional data mining, massive amounts of transaction data are collected. On the contrary, defect data in construction industry are sparse and make it difficult to extract association rules because the set of items in one instance is too small comparing with the whole set of items in the database and the cases are difficult to obtain [20,3].
- Previous works usually focused on mining rules at a single concept level [14–19]. It is difficult to find many strong association rules at a primitive concept level [21,16]. And, the extracted rules convey less information when items in a rule are more specific [21,22].
- In the traditional mining methods, the number of rules and the usefulness of mining results vary with the threshold values for support and confidence, both of which are defined by users [23,24]. If the threshold values are set too high, some useful patterns will be pruned. In contrast, too low values will lead to the mining result full of useless patterns.
- In the traditional methods, many discovered association rules are redundant as they only capture the irregularities and idiosyncrasies of data [25]. Such rules have obstructed the interpretation of regularities in the domain. Unfortunately, the number of extracted rules makes it impossible for decision makers to prune redundant rules manually.

To tackle the above-mentioned problems, this paper proposes a solution that incorporates concept hierarchy [21,22] and GA [26–28] with the traditional Apriori algorithm [29,24]. The proposed algorithm is applied to a defect database containing data from defect-oriented disputes to identify hidden patterns of defects. The rules make project managers to understand defect causation and develop quality improvement plans.

The rest of the paper is organized as follows: Section 2 presents a review of relevant works in literature; in Section 3, the architecture of the rule mining system in the research is developed and the model and implementation procedure of the proposed GA-based multi-level association rule mining approach are described in details; in Section 4, a case study is demonstrated to illustrate the usefulness and applicability of the proposed approach in the construction industry; in Section 5, assumptions and limitations made in model development and findings from a case study are addressed and discussed; Section 6 concludes the findings of the research and suggests future research directions.

2. Review of relevant works

2.1. Construction defects

Due to the outdoor site and the unique nature of work, defects are almost inevitable in construction project. According to Watt [1], a defect is defined as "a failing or shortcoming in the function, performance, statutory or user requirements of a building, and might manifest itself within the structure, fabric, services or other facilities of the affected building". Previous studies on defects can be classified into three categories: (1) classifying defects [20,30]; (2) identifying causations of defects [2,10,4,12,13]; and (3) reuse of historical defect information [5]. Works of the three categories are briefly reviewed as follows.

Depicted with different attributes, defects can be categorized by values of attributes, which constitutes classification system of defects [20]. Classification of defects provides the foundation for the analysis of housing defects [20]. Georgiou et al. [31] divided defects into the following three categories: (1) technical defects—arising from poor workmanship or defective material of an element; (2) esthetic defects—relating to the appearance of an element; and (3) functional defects—rendering the building unusable. Sommerville and McCosh [32] furthered this study by adding omissions into the division to represent defects of omitting parts of a building. Macarulla et al. [20] proposed a two-level classification system of defects including main level with general words and second level with specific concepts to embrace the whole lifecycle of projects. Other criteria used to classify defects include elements where defects were found [31,2], materials of the elements [2], and stages when defects occurred [2,3].

Besides classifications, the causes of defects in construction have been analyzed to develop defect prevention strategies [4]. Josephson and Hammarlund [10] collected 2879 defects from seven building projects and calculated the distribution of the defect origin and the defect causes in each project, as well as the number of defect causes for different actors. Chong and Low [33] analyzed defect records of 79 buildings to determine the percentage of defect causes in 10 elements and summarize the major causes of defects in each element. Sassu and Falco [34] provided detailed description of some defects, such as water infiltration in the basement, leakage in pitched roofs and water damage in flat roofs, and also showed the distribution of the causes for each defect. Forcada et al. [35] carried out a contingency and correlation analysis to test association between defects and sources and between defects and origins. Although such statistical methods provide reliable and useful information for assessing the defects and improve the quality, the relationship among defect factors and the combination of factors are neglected. Aware of the weakness of statistical methods, Aljassmi and Han [4] adopted a fault tree to formulate the structure of the complex causes resulting in defects and to quantify the influences of defect causes in terms of frequency and magnitude. They then extended the fault tree to a project pathogens network to generate the construction defects effectively [36]. Love et al. [37] used a system dynamics modeling to simulate the influences between factors related to rework. Love et al. [38] developed a structural model of the causes of rework by distributing questionnaires to reveal path coefficients predicting rework. Palaneeswaran et al. [39] proposed artificial neural networks (ANN) to map the causes and effects of rework.

A historical defect case comprising of problem and solution descriptions provides comprehensive and effective knowledge for defect analysis. Compilation of historical defect cases to be stored in the database can be viewed as a vital preprocessing step for information reuse. Such an approach has been widely employed in different domains, including chemical accident analysis [40], safety management [14], and accident adjudication [41]. The defect specific domain ontology, which is tied with a well-classified defect data collection system, could be developed for the effective utilization of defect information through defining the relationships among various knowledge sectors in defect data [5]. A defect domain ontology to search and retrieve project defect information has been proposed by Park et al. [5]. Their defect domain ontology is connected with a comprehensive defect data template. It could help users to identify and easily access the most relevant and critical defect information. For the managers, the defect domain ontology could be utilized as a foundation to prepare a proactive and project-specific defect management plan. Similar to Park et al. [5], Lee et al. [15] used a relational database to store data related to product quality and defects. Such schemes, while incorporating with data mining techniques, provide promising solutions for finding useful defect management strategies for construction industry.

2.2. Algorithms for mining association rules

Data mining is a technique used to extract knowledge from an existing data set and transform it into a human-understandable structure [42]. As a specific technique of data mining, association rule mining is employed to discover the interesting relations between variables in large databases. An association rule is an implication in the form of '$X \rightarrow Y$', where X and Y are sets of items and the intersection of them is empty. X and Y represent the "If" part and the "Then" part, respectively. For example, an association rule extracted from construction defect database indicates if defects in X exist then defects in Y also occur.

Apriori algorithm is one of most commonly used methods for the mining of association rules [29]. The Apriori algorithm divides a rule mining process into two steps: First, the database is scanned to find all the itemsets with support values above the predefined minimum; Second, a rule is generated if it satisfies the predefined minimum confidence. Rule