Data-driven Solutions for Building Environmental Impact Assessment

Qifeng Zhou, Hao Zhou
School of Information Science and Engineering
Xiamen University, Xiamen, China
Email: zhouqf@xmu.edu.cn
colinzhou2013@gmail.com

Yimin Zhu
Department of Construction Management
Louisiana State University
Baton Rouge, LA 70803
Email: yiminzhu@lsu.edu

Tao Li
School of Computing and Information Sciences
Florida International University
Miami, FL 33199
Email: taoli@cs.fiu.edu

***Abstract*—Life cycle assessment (LCA) as a decision support tool for evaluating the environmental load of products has been widely used in many fields. However, applying LCA in the building industry is expensive and time consuming. This is due to the complexity of building structure along with a large amount of high-dimensional heterogeneous building data. So far building environmental impact assessment (BEIA) is an important yet under-addressed issue. This paper gives a brief survey of BEIA and investigates potential advantages of using data mining techniques to discover the relationships between building materials and environment impacts. We formulate three important BEIA issues as a series of data mining problems, and propose corresponding solution schemes. Specifically, first, a feature selection approach is proposed based on the practical demand and construction characteristics to perform assessment analysis. Second, a unified framework for solving constraint-based clustering ensemble selection is proposed to extend the environmental impact assessment range from the building level to the regional level. Finally, a multiple disparate clustering method is presented to help sustainable new buildings design. We expect our proposal would shed light on data-driven approaches for environment impact assessment.**

I. Introduction

Environmental sustainability is becoming an increasingly important issue worldwide. In the situation of environmental deterioration, global warming is the most severe environmental crisis. Global warming is the consequence of long term build up of greenhouse gases(GHG), such as CO_2, CH_4, NO_2, in the higher layer of atmosphere [1]. Studies have shown that the biggest contributor to GHG emissions is the built environment, which accounts for more than 38% of global carbon dioxide emissions [2]. In addition, building systems consume about 40%-50% global energy use, 12% of global potable water use and 40% of solid waste in developed countries [3]. More attention has been paid to buildings sustainable development.

Life cycle assessment (LCA) is a widely used method for evaluating the environmental loads of products during their whole life cycle [4]. Although the importance of obtaining environment-related product information by LCA is broadly recognized, performing LCA is not a straightforward process. A typical analysis of buildings may consist of several thousand input variables. Moreover, due to the complexity and the long life cycle of buildings, applying LCA in the building industry has became an important research area within LCA practice [5].

Prior studies have shown that processing the large-scale, high-dimensional, heterogeneous building inventory data is a big challenge when applying LCA to BEIA [5], [6]. Domain experts and researchers are eagerly waiting for new ways to cope with the challenge. In addition to the inventory data, many related industries have also accumulated a large amount of environmental monitoring data. In recent years, data mining, as a technique for discovering interesting patterns from large data sets, has been used in sustainable development. These data mining techniques enable data-driven solutions which may greatly promote the development in these industries. Using these technologies, researchers can easily discover the changes in the ecological environment and make corresponding policies in practice.

In this paper, we study three important issues in building sustainable development and formulate them as a series of data mining problems. Note that recognizing the important component/material impact equivalent is the basis of computing environment impact and reducing the assessment cost. Therefore, the first important issue is identifying the characteristics of each building (such as building type, building form, building size, building location, construction time) and selecting a small and proper building attribute sets. Large scale building group environment impact assessment is another important yet difficult problem. In order to reduce the evaluation cost, we formulate building group impact assessment as a clustering problem and propose a constraint-based clustering ensemble mechanism that can incorporate different types of priori knowledge. In addition, how to find the multiple alternative material or component combinations via data analysis is an important issue in new building design. So far, green building design mainly depends on the experience of domain experts. Developing an data-analysis-based eco-construction material selection approach is the key step of implementing the green building design.

The overall framework of data-driven building environmental impact assessment is described in Fig.1.

IEEE ICSC 2015, February 7-9, 2015, Anaheim, California, USA
978-1-4799-7935-6/15/$31.00 ©2015 IEEE

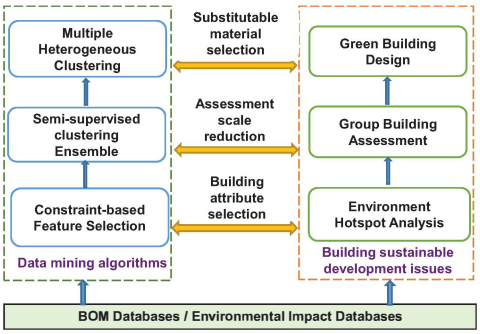

Fig. 1. The framework of data-driven solutions for BEIA

II. Related Work

Various approaches and tools have been used for BEIA based on the building characteristics and the assessment requirements. A common tool used for environmental assessment is LCA. Although LCA as a basic tool for building assessment has been widely used since 1990, it is less developed in building industry than in other industries [7]. One of the main reasons is its limited capacity of collecting, processing, and analyzing large amount of complex building data.

Several researchers formulate LCA as a multi-objective programming (MOP) problem [8]. Given a set of decision variables (e.g., building material, components, energy), LCA aims to optimize more than one objective functions with constraints as formulated below:

$$\min_{x \in X} f(x) = (f_1(x), f_2(x), ..., f_m(x))^T; \quad s.t. \quad \sum Q \cdot X \leq E, \tag{1}$$

where $f_i(x)$'s are different objective functions such as environmental burden, ecological impact, and the economic benefits. $Q = \{q_1, ...q_n\}$ is the coefficient matrix, $E = \{e_1, ...e_j\}$ is the maximum construction/consumption cost or constraint conditions (e.g., the balance between material and energy).

There are three general approaches to solve the LCA problem: process analysis, input/output analysis, and hybrid analysis [9]. Several other data analysis studies in LCA, including impact variation analysis, simulations methods (such as Monte Carlo) to test robustness, empirical methods (such as typologies) to evaluate the data quality, have been conducted in the context of uncertainty analysis, data quality assessment and sensitivity studies [5], [10].

Recently, data mining techniques have been introduced into LCA for simplifying data analysis and impact assessment. Currently, data mining is mainly applied to the following sub-problems:

- **Data collection and completion:** Frischknecht et al. [11] set up a web-based LCA database which can automatically collect and preprocess inventory data with different formats. Usually, environment impact dataset may contain missing or invalid entries. Different data mining techniques are used for filling in the missing values [12].
- **Reconstructing inventory data structure:** Sundaravaradan et al. [13] adopt the non-negative least squares algorithm to infer inventory trees from a database of environmental factors. They aim to address the problem in which only the overall environment impacts are reported but the inventory data is not available.
- **Energy/economic consumption analysis:** Some researchers adopt data mining algorithm to realize optimal building energy consumption [14]. Slawomir Golak et al. [15] applied artificial neural networks (ANN) to evaluate the economic and ecological effects of new products design process.
- **Environmental impact assessment:** Some researchers applied ANN to assess the environmental impacts of product design alternatives. Compared with statistic methods, ANN-based method simplified the valuation parameters and improved the evaluation performance [8], [16]. Hossain et al. [12] proposed a framework for sustainable design that simplifies the effort of estimating the environmental footprint from a product bill of materials.

Although the aforementioned data mining techniques addressed many data analysis problems in LCA, they can only deal with the assessment of a single product. There are little research efforts on combining data mining techniques with building LCA. Since the typical BEIA data is of large-scale, high-dimensional and heterogeneous, performing LCA in BEIA is labor intensive and impractical. Different from the general LCA methods, this paper treats building LCA as a learning problem and infer the relationships between the building materials and environment impact facts from the large amount of available data.

III. Data Mining Solutions for Building Environment Impact Assessment

The following studies investigate the data-driven solutions for three important BEIA issues and give the corresponding algorithms or solution framework.

A. Constraint-based Building Attribute Selection

Building bill of material (BOM) provides a comprehensive listing of the components, but it dose not include their environmental impacts. Hence, we need map the BOM nodes to nodes in an environmental impact (EI) database (as shown in Table I and Table II). This process can be completed using a classifier such as naive Bayes or other classification methods [12]. Since a typical building BOM contains thousands of nodes, it is not easy to simplify the inputs of building LCA and identify the major contributors to the environment deterioration (footprint). As a result, building attribute selection is needed to identify the materials, components or their combinations having major impacts to the environment. Building attribute selection can be treated as a feature selection problem. However, unlike the classical feature selection, performing BEIA attribute selection should consider both the feature importance and the feature costs (e.g., economy benefit and scarcity of material). We formulate it as constraint-based building attribute selection and then introduce a principled approach which can naturally incorporate the constraints (e.g., feature costs) into the process of feature selection.

TABLE I
AN EXAMPLE OF BOM DATASET

Name	No.	Description	Unit	...
concrete	150001C325	Ordinary portland cement(32.5Mpa)	t	...
concrete	150001C425	Ordinary portland cement(42.5Mpa)	t	...
glasses	530011G001	Ordinary glass-6 mm	m^3	...
lime	1280010M10	M5	m^3	...
...	...	...	...	...

TABLE II
AN EXAMPLE OF EI DATASET

Component	$CO_2(kg)$	$SO_2(kg)$	Water(m^3)	...
Solid clay brick(1000)	5.04E+02	3.15E+01	3.42E+00	...
Lime(1t)	4.58E+02	2.68E+01	3.40E-01	...
Normal concrete block($1m^3$)	1.46E+02	3.86E+01	9.62E-01	...
...	...	...	...	...

Note that Support Vector Machine (SVM) has the excellent performance for feature selection of high-dimensional small sample problem [17]. We thus incorporate the feature costs into the SVM optimization process which can be modeled as follows:

$$\begin{aligned} &\min_{w,b,\mu,\xi,\tau} \ \beta_1 \sum_{j\in N} \mu_j + \beta_2 \sum_{i\in M} c_i\xi_i + \beta_3 \sum_{j\in N} p_j\tau_j; \\ &s.t. \quad y_i(\omega \cdot x_i - b) \geq 1 - \xi_i, \xi_i \geq 0, i \in M, \\ &-\mu_j \leq \omega_j \leq \mu_j, \tau_j \in 0, 1, j \in N, \end{aligned} \tag{2}$$

where p_j denotes the cost of j_{th} feature (the larger cost, the larger p), N is a large positive number. Since τ_j is a discrete value 0 or 1, it can be approximated with a Sigmoid function. The above model can be solved using convex quadratic programming.

Computing the environment impact of a component or material is actually a regression problem. For simplicity, a regression problem can be converted to a classification problem. In addition, although our constraint-based SVM feature selection method is described in the context of classifiers, it can also be implemented using support vector regression.

B. Region-level Impact Assessment using Constraint-based Clustering Ensemble

Large scale building environmental impact assessment (such as a region or a city) is an attractive problem with the rapid extension and growth of population in urban area. The number of different buildings in a region is often very large. It is labor intensive to perform LCA on every single building in a region. A few studies recently attempted to apply LCA to large scale building environmental impact assessment [18]. These studies demonstrate that extending the building level assessment to the regional level is quite challenging.

We introduce a data-driven solution for regional level buildings LCA. This solution can be briefly described as follows. First, based on building characteristic, buildings having similar environment impacts in a region are clustered into a group . Second, a small number of buildings in each cluster are selected as the representative buildings and their environment impacts are assessed using LCA. Third, computing the environment impacts of each cluster by scaling the impacts of representatives. Finally, the environment impacts of the whole region can be obtained by accumulating the impacts of all building clusters. The key problem in the solution is how to generate robust and stable clusters. In this work, we propose a semi-supervised clustering ensemble framework as shown in Fig.2. The proposed framework utilizes cluster ensemble (i.e., combining multiple base clustering results) and semi-supervised learning (i.e., incorporating priori knowledge) to improve the robustness and performance of regional level environment impact assessment.

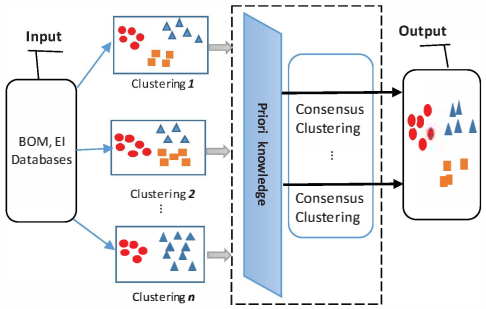

Fig. 2. The Framework for regional level BEIA

Clustering ensemble provides a framework for combining multiple base clusterings of a dataset into a single consolidated clustering. Compared to individual clustering algorithms, clustering ensemble can generate more robust and accuracy clustering results. Prior knowledge obtained from domain experts can provide some valuable information when performing BEIA. Recently, incorporating prior knowledge in unsupervised learning evokes the development of semi-supervised clustering, also known as constrained clustering [19]. A natural way of incorporating the constraints in cluster ensemble is to use them in generating base clustering solutions. However, the prior knowledge often changes with different building assessment conditions (e.g., building region, building style, and assessment task). It is computational expensive to re-apply constrained clustering for data analysis when faced with large datasets. Therefore, a better way for making use of the constraints in cluster ensemble is to use them in the ensemble selection process (as shown in Fig.2), instead of directly incorporating them in the base clustering algorithms.

This framework can be implemented as a combinatorial optimization problem in terms of the consistency under the constraints, the diversity among ensemble members, and the overall quality of ensembles.

C. Finding Sustainable Design Alternatives Based on Multiple Heterogeneous Clustering

Green building design should consider what kind of building structures or building systems can meet the environmental requirements. To fulfil this task, one key problem is how

to discover sustainable design alternatives from large amount of building materials or components. A straightforward solution is perform clustering on the BOM dataset and the EI dataset, respectively. Clustering on the BOM dataset will group functionally similar components, and clustering on the EI dataset will group components with similar impact factors. In order to find design alternatives, the clustering results should have the following properties: (1) The components with similar environment impacts should be grouped in the same clusters. (2) The components with similar functions should be distributed across different environment impact clusters. To find design alternatives, we desire the clustering results on the EI dataset is different from the clustering results on the BOM dataset. The disparateness between the two clustering results (functionality and environment impacts) would provide us the design alternatives. However, since the relationship between BOM dataset and EI dataset is one-to-one, traditional clustering methods such as K-means or hierarchical clustering will not produce clustering results with significant difference. We thus propose a disparate clustering algorithm to find the design alternatives (as shown in Algorithm 1).

Algorithm 1 Finding design alternatives

Input: BOM dataset, EI dataset

1: Mapping the BOM dataset to EI dataset;
2: Clustering on the BOM dataset according to the text descriptions(function);
3: Multiple clustering on the EI dataset;
4: Calculating the difference of two clustering results;
5: Finding the design alternatives.

Output: Designing alternatives

Multiple clustering can be implemented using different attributes in K-means or by changing the cluster merging rules in hierarchical clustering [20]. We can use a contingency table to capture the relationships between members in BOM clusters and EI clusters. In the ideal case, the contingency table in BOM clustering would be close to a diagonal matrix and the contingency table in EI clustering would be close to a uniform distribution [12]. Then an objective function can be used to evaluate the clustering results. For example, the following relative entropy can be used as the objective function:

$$R = D(C||U), \tag{3}$$

where C is the distribution function of the contingency table over EI clusters, and U is the uniform distribution. Relative entropy is often useful as a "distance" between two distributions and our goal is to minimize the objective function (in other words, we hope the distribution C is similar to U).

IV. Conclusion

This paper provides an overview of building sustainable development and gives a brief analysis of the challenges when applying LCA to building environment impact assessment. Three important issues (i.e., building attribute selection, region-level assessment, and sustainable building design) are formulated as a series of data mining problems, and the corresponding solution schemes are proposed. In our future work, we will perform empirical verification of our proposed solutions on benchmark datasets and real-world building BOM datasets.

Acknowledgement: This work is supported by the Key Laboratory of System Control and Information Processing, Shanghai Jiao Tong University under Grant No.SCIP2012007 and Natural Science Foundation of China under Grant No.61403318 and No.61271337.

References

[1] A. H. Buchanan and B. G. Honey, "Energy and carbon dioxide implications of building construction," *Energy and Buildings*, vol. 20, no. 3, pp. 205–217, 1994.
[2] M. Comstock, C. Garrigan, S. Pouffary, de Feraudy T., J. Halcomb, and Hartke, "Building design and construction: Forging resource efficiency and sustainable development," United National Environmental Program (UNEP), technical Report, Jun 2012.
[3] N. Raynsford, "The uk's approach to sustainable development in construction," *Building Research & Information*, vol. 27, no. 6, pp. 419–423, 1999.
[4] J. Reap, F. Roman, S. Duncan, and B. Bras, "A survey of unresolved problems in life cycle assessment," *The International Journal of Life Cycle Assessment*, vol. 13, no. 5, pp. 374–388, 2008.
[5] M. M. Khasreen, P. F. Banfill, and G. F. Menzies, "Life-cycle assessment and the environmental impact of buildings: a review," *Sustainability*, vol. 1, no. 3, pp. 674–701, 2009.
[6] Q. Zhou, F. Yang, and T. Li, "Application of data mining in building industry," in *Data mining: where theory meets practice*. Xiamen University, 2013, pp. 332–378.
[7] J. A. Fava, "Will the next 10 years be as productive in advancing life cycle approaches as the last 15 years?" *The International Journal of Life Cycle Assessment*, vol. 11, pp. 6–8, 2006.
[8] K.-K. Seo, "A methodology for estimating the product life cycle cost using a hybrid ga and ann model," in *Artificial Neural Networks–ICANN 2006*. Springer, 2006, pp. 386–395.
[9] K. Zhang, "Analysis of non-linear inundation from sea-level rise using lidar data: a case study for south florida," *Climatic Change*, vol. 106, no. 4, pp. 537–565, 2011.
[10] H. A. U. Haes, R. Heijungs, S. Suh, and G. Huppes, "Three strategies to overcome the limitations of life-cycle assessment," *Journal of industrial ecology*, vol. 8, no. 3, pp. 19–32, 2004.
[11] R. Frischknecht and G. Rebitzer, "The ecoinvent database system: a comprehensive web-based lca database," *Journal of Cleaner Production*, vol. 13, no. 13, pp. 1337–1343, 2005.
[12] M. S. Hossain, M. Marwah, A. Shah, L. T. Watson, and N. Ramakrishnan, "Autolca: A framework for sustainable redesign and assessment of products," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 2, p. 34, 2014.
[13] N. Sundaravaradan, M. Marwah, A. Shah, and N. Ramakrishnan, "Data mining approaches for life cycle assessment," *IEEE ISSST*, vol. 11, 2011.
[14] Y. Yuan, J. Yuan, H. Du, and L. Li, "Pareto ant colony algorithm for building life cycle energy consumption optimization," in *Life System Modeling and Intelligent Computing*. Springer, 2010, pp. 59–65.
[15] S. Golak, D. Burchart-Korol, K. Czaplicka-Kolarz, and T. Wieczorek, "Application of neural network for the prediction of eco-efficiency," in *Advances in Neural Networks–ISNN*. Springer, 2011, pp. 380–387.
[16] J.-H. Park and K.-K. Seo, "A knowledge-based approximate life cycle assessment system for evaluating environmental impacts of product design alternatives in a collaborative design environment," *Advanced Engineering Informatics*, vol. 20, no. 2, pp. 147–154, 2006.
[17] V. N. Vapnik and V. Vapnik, *Statistical learning theory*. Wiley New York, 1998, vol. 2.
[18] S. M. Batouli and Y. Zhu, "A framework for assessing environmental implications of an urban area," in *Construction Research Congress 2014@ sConstruction in a Global Network*. ASCE, pp. 2365–2374.
[19] A. P. Topchy, A. K. Jain, and W. F. Punch, "A mixture model for clustering ensembles." in *SDM*. SIAM, 2004, pp. 379–390.
[20] T. Li, M. Ogihara, and S. Ma, "On combining multiple clusterings: an overview and a new perspective," *Applied Intelligence*, vol. 33, no. 2, pp. 207–219, 2010.