A Study of the Relationship between Credits in the LEED-EB&OM Green Building Rating System

Jack C. P. Cheng and Lucky J. Ma

***Abstract*—LEED (Leadership in Energy and Environmental Design) is a credit-based green building rating system. Considering that a better understanding of the relationships between credits would help managers better achieve green building certification, this study analyzed 1381 projects that have been certified in LEED-Existing Building versions 2008 and 2009. The credits achieved by those projects were analyzed using data mining techniques to discover hidden inter-relationships and the effects on high-scoring sustainable design strategies. The data mining results were compared with the credit pairs provided by LEED AP consultants from the engineering perspective. Additional hidden credit pairs were also discovered.**

***Index Terms*—Correlation Rules, Data Mining, Green Building, Leadership in Energy and Environmental Design (LEED).**

I. Introduction

A. Background

According to the U.S. Department of Energy Building Energy Datebook, buildings account for 39 percent of all energy consumption and 48 percent of greenhouse gas emission in the U.S. [1]. Among different building categories, existing buildings have quite a large environmental impact. The most recent data published by the U.S. Department of Energy show that new constructions each year only add roughly 1 percent to the U.S. building stock[2]. Thus, compared to new green buildings, retrofitting existing buildings to an acceptable green level will bring much more energy saving and reduction in environment impacts.

The LEED (Leadership in Energy and Environmental Design) Green Building Rating System is a performance-based tool for determining a building's environmental impact, and the LEED for Existing Building: Operations and Maintenance (LEED-EB&OM) emphasizes the certification of green existing buildings.

B. Problem Statement and Research Objectives

The credits that comprise LEED are designed to value a building's sustainable performance, and the number of credits generally determines the level of achievement. Although credit selection is critical to the success of certification, few studies have been conducted to address the relationships between particular credits.

This study tries to study the hidden relationship between credits in LEED-EB&OM V2009 & V2008 by conducting data mining analysis. By identifying credits with strong associations, credit bundles that are commonly achieved can be identified. These credit selection biases may reveal interesting implications for current sustainability practitioners and for the future development of LEED.

C. Literature Review

In 2004, the U.S. General Services Administration (GSA)released a study of the cost for applying LEED based on typical GSA construction projects[2]. In this report, "synergistic credits" were raised for achieving cost-effectiveness. The author defined synergistic credits as achieving a combination of LEED credits which cost significantly different from the sum cost of achieving one by one. However, as green building technologies developed fast in recent years, his results were outdated.

In 2008, Benjamin J. Thomas[1] did research on the association rules between credits in the LEED for new construction. He defined credit bundles as credits with hidden relationships. His final results were applicable and reasonable, but it had an obvious limitation that his model for association rule mining was not comprehensive. The author actually only used high *importance* values to select the potential synergic credits, which would miss many valuable rules.

This research tries not only to use the advantages of the previous research but also to update the mining model to conduct more comprehensive research on mining the hidden relationship between credits.

II. Methodology

The data mining technique was used in this study by analyzing the LEED-EB&OM projects' credit database. The data mining technique is the analysis of observational data sets in order to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner[3].In order to find credit bundles with strong inter-relationships, frequent patterns, association and correlation rules were targeted.

A. Frequent Patterns

Frequent patterns are patterns or attributes that appear frequently in a data set[4]. If the occurrence possibility is larger than the required threshold, then it is a frequent pattern. The occurrence possibility of certain patterns in data mining is called support, and can be represented as in (1), when A and B represent two attributes or events. A higher *support* value means a higher occurrence possibility of all the attributes; in other words, the more likely the attributes are related.

Manuscript received January 31, 2013; revised May 6, 2013.

The authors are with the Department of Civil and Environmental Engineering, The Hong Kong University of Science & Technology, Hong Kong, China (e-mail: {cejcheng, jmaae}@ust.hk).

DOI: 10.7763/IJET.2013.V5.592

$$\text{Support}(A, B) = \text{P}(A \cap B) = Support\_Count\ (A\ \&\ B) \quad (1)$$

1) Apriori Algorithm

Apriori is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994for mining frequent patterns [5]. It uses a breadth-first search and hash tree structure to count candidate patterns. First, it generates candidate patterns of length *i* from patterns of length *i-1*. Then it prunes the candidates with infrequent sub-patterns. According to the downward closure lemma, the candidate set contains all frequent *i-length* patterns. After that, it scans the transaction database to determine frequent patterns among the candidates[3]. Fig. 1 shows the pseudo-code of the Apriori algorithm.

```
i=0;
Ci = {{A} | A is a variable};
while Ci is not empty do
        database pass:
                for each set in Ci, test whether it is frequent;
                let Li be the collection of frequent sets from Ci;
        candidate formation:
                let Ci+1 be those sets of size i+1 whose all subsets are frequent;
end.
```

Fig. 1. Apriori algorithm [3]

2) Maximal Patterns

A frequent pattern is called maximal if no superset is frequent, based on the user defined *support* threshold[6]. For example, if pattern {A, B, C} is frequent, and its superset {A, B, C, D} is not frequent, then {A, B, C} is a maximal pattern. This is a good way to find out closed frequent patterns by self-adjusting the number of attributes.

B. Association Analysis

As well as frequent pattern mining, Association Rule Mining was also motivated by market basket analysis, and is based on observational data to find out which products tend to be purchased together. The rule, for example, can be represented like "if buy beer, then buy bread". Actually, association rule mining relies on frequent patterns, but besides the *support* measure, association analysis also needs *confidence* measurement which could be represented as in (2).It is the occurrence possibility of B, given A; thus a higher *confidence* value means more likely there exist hidden relationships between *A* and *B*.

$$\text{Confidence}(A => B) = P(B|A) = \frac{Support\_Count\ (A\ \&\ B)}{Support\_Count\ (A)} \quad (2)$$

If a rule (e.g. *A*=>*B*) meets the minimum requirement of the *support* and *confidence* measure, it is named as an "interesting association rule".

C. Correlation Analysis

According to recent research on association analysis, *support* and *confidence* alone are not always enough to discover interesting rules [4]. Thus, a correlation measure can be used to augment the support-confidence framework for association rules. A correlation rule is measured not only by *support* and *confidence* but also correlation between A and B. This research mainly used 4 measures: *lift*, *difference of confidence*, *chi-square,* and *importance*.

1) Lift

Lift is a simple correlation measure[4]. According to the possibility theory, the occurrence of A is independent of the occurrence of B if $P(AB) = P(A) * P(B)$. Otherwise, they are dependent and correlated as events. The lift between the occurrence of A and B can be computed by

$$\text{Lift}(A => B) = {P(AB)}\Big/{P(A) \cdot P(B)} \quad (3)$$

It is revealed in (3) that, if lift>1, then the occurrence of *A* and the occurrence of *B* are positively correlated; otherwise, negatively correlated. In other words, the larger the lift value is, the more interesting the rule is.

2) Difference of Confidence (DOC)[7]

This correlation measure is to compare the posterior and the prior *confidence* of an association rule[6]. Since the former should differ considerably from the latter to make the rule interesting, which means the occurrence of *A* has a significant impact on the occurrence of *B*. In other words, the larger the value of DOC, the more interesting the correlation rule is. It is given as follows.

$$\text{DOC}(A => B) = |\text{Confidence}(A => B) - \text{Confidence}(B)| \quad (4)$$

The *confidence* of *B* actually is the *confidence* of All=>*B*, which equals the *support* of *B*.

3) Chi-Square Measure[4]

The *chi-square* measure is well known from statistics. Along with the *correlation coefficient* and *covariance*, the *chi-square* measure is usually used to analyze the correlation relationship between two attributes. Consider that the credit database is in a binomial (0 or 1) format, thus the *chi-square* measure is used in this research. The governing equation of the *chi-square* measure is

$$\chi^2_{A,B} = \sum \frac{(Observed - Expected)^2}{Expected} \quad (5)$$

$$\text{Expected}_{A^+,B^+} = \frac{A^+_{Sum} \cdot B^+_{Sum}}{Sum} \quad (6)$$

where$\chi^2_{A,B}$ is the *chi-square* value between two attributes *A* and *B*. The *Observed* and *Expected* are the observed and expected data of *A* or *B*, and the expected value could be calculated through (6). While $Expected_{A^+,B^+}$ represents the count of cases when *A*=1 and *B*=1, A^+_{Sum} represents the count of cases when *A*=1, andB^+_{Sum} represents the count of cases when *B*=1. *Sum* means the total number of cases.

Table I shows an illustrative example of calculating the *chi-square* value in this study. SSc2 and SSc3 means two credits in LEED-EB&OM. SSc2 represents the credit for the building exterior and hardscape management plan, and SSc3 represents the credit for integrated pest management, erosion control, and landscape management. The underlined value is the observed count, and the one in the bracket is the expected count.

TABLE I: CHI-SQUARE VALUE CALCULATION SAMPLE

	SSc3=1	SSc3=0	CountSum-SSc2
SSc2=1	**960**(898.4)	**225**(286.6)	1185
SSc2=0	**87**(148.6)	**109**(47.4)	196
CountSum-SSc3	1047	334	CountSum=1381

The *chi-square* value of this case is around 123. The larger the *chi-square* value, the more likely the variables are related[4].

4) Importance

This measure was introduced in Benjamin J. Thomas's study[1], and is computed as follows.

$$Importance(A => B) = \log(\frac{P(B|A)}{P(B|notA)}) \qquad (7)$$

It is a straight-forward measurement on the impact of the occurrence of A over the occurrence of B. When *importance* is larger than 0, it means the occurrence of A has a positive impact on the occurrence of B; otherwise, a negative impact. In short, the larger the *importance* value is, the more interesting the rule may be.

III. Data Mining Process

This study followed the process model of the Cross-Industry Standard Process for Data Mining (CRISP-DM), which is commonly used by expert data miners. CRISP-DM is organized into a set of six phases: (1) business understanding, (2) data understanding, (3) data preparation, (4) modeling and mining, (5) evaluation, and (6) deployment or verification[1].

Because the business understanding phase focuses on gaining a perspective of the business and translating it into a data mining problem, and Section II has already introduced it, so other phases are discussed in the following.

A. Data Understanding and Collection

1) LEED-EB&OM versions

There are three versions of LEED-EB&OM rating system-V2.0, V2008, V2009. In this study, V2008 and V2009 were used because they have a similar credits structure, which is quite different from that of V2.0. The most measureable change between V2008 and V2009 was the introduction of Regional Priority Credits. Except for this, most of the differences were just changes to the credit weight or name of the credit.

2) Data collection

	A	B	C	D	E
1	SSc1	SSc2	SSc3	SSc4	SSc5
2	0	0	0	0	1
3	0	0	1	0	1
4	0	1	0	0	0
5	0	1	1	0	0
6	0	1	1	0	1
7	0	1	1	0	0
8	0	1	0	0	1
9	0	1	0	2	0

Fig. 2. Excerpt from the credit achievement database.

TABLE II: Cases Statistics

Category	Platinum	Gold	Silver	Certified	Total(1381)
V2008	23	338	231	120	712
V2009	55	326	202	86	669

The credit database was collected from the U.S. Green Building Council (USGBC) website (Project Directory: http://new.usgbc.org/projects). In order to better fit the general association and correlation analysis, the credit database was presented in a binomial tabular format. Each column represents an individual credit in the rating system, and each row represents a project case. Fig. 2 provides an excerpt from the database, while the numbers represent the relevant points. Table II introduces how many projects were included in this study.

B. Data Preparation

1) Data cleaning

Firstly, in order to integrate the cases from V2008 and V2009, both the Innovation in Operation and the Regional Priority credits categories were pruned, because these two categories are too specific to generate general rules and hidden information. Secondly, the prerequisite credit columns were also eliminated because all the cases would have to achieve them as prerequisite requirements.

2) Data integration

There are still two problems left before modeling and analysis. One is about the multiple points. For example, SSc4has a value of15 points and EAc1 has 18 points. Another is the credit difference between V2008 and V2009. For instance, the Sustainable Purchasing – Ongoing Consumables credit in V2009 only takes 1 point, while in V2008 it takes 3 points.

Two ways for overcoming the problems are proposed in this study. Firstly, the credits that only represent the increments of the same design implementation were omitted, or in other words, integrated them as 1 (Table III).For example, in V2009, the credit for Additional Indoor Plumbing Fixture and Fitting Efficiency ranges from 0~5 representing different percentage of water saving, and in V2008, same idea is conveyed by 3 credits (WEc2.1, WEc2.2 and WEc2.3). If a building gets more than 1 point in WEc2 V2009 or one of the three credits in V2008, we treat it as 1 credit.

TABLE III: Credits or Points Integrated as 1Credit

Context	V2008	V2009
Additional Indoor Plumbing Fixture and Fitting Efficiency	WEc2.1~WEc2.3 (1~3=>1)	WEc2 (1~5=>1)
Water Efficient Landscaping	WEc3.1~WEc3.3 (1~3=>1)	WEc3 (1~5=>1)
Cooling Tower Water Management	WEc4.1~WEc4.2 (1~2=>1)	WEc4 (1~2=>1)
On-site and Off-site Renewable Energy	EAc4.1~EAc4.4 (1~4=>1)	EAc4 (1~6=>1)
Sustainable Purchasing - Ongoing Consumables	MRc1.1~MRc1.3 (1~3=>1)	MRc1
Sustainable Purchasing - Reduced Mercury in Lamps	MRc4.1~MRc4.2 (1~2=>1)	MRc4
Solid Waste Management - Ongoing Consumables	MRc7.1~MRc7.2 (1~2=>1)	MRc7
Occupant Comfort - Daylight and Views	IEQc2.4~IEQc2.5 (1~2=>1)	IEQc2.4
Green Cleaning - Custodial Effectiveness Assessment	IEQc3.2~IEQc3.3 (1~2=>1)	IEQc3.2
Green Cleaning - Sustainable Cleaning Products and Materials	IEQc3.4~IEQc3.6 (1~3=>1)	IEQc3.3
Existing building commissioning - investigation and analysis	EAc2.1 (2=>1)	EAc2.1 (2=>1)
Existing building commissioning - implementation	EAc2.2 (2=>1)	EAc2.2 (2=>1)
Existing building commissioning - ongoing commissioning	EAc2.3 (2=>1)	EAc2.3 (2=>1)

Secondly, if the multiple credits had significant differences between the high points and low points, then they were separated into different categories. For example, EAc1-Optimize Energy Efficiency Performance (18 points in total), the official guideline mentions that if the calculation of energy saving was based on its historical data, it could get possible credits between 1~9; and if it was also based on comparable buildings, then 10~18 points were acceptable. Thus, this credit was divided into 2 columns. One represents scores 1~9, and the other represents scores 10~18. Detailed treatment is presented in Table IV.

TABLE IV: SPECIAL MULTIPLE CREDIT INTEGRATION

Context	V2008	V2009
Alternative Commuting Transportation	<50% reduction – Low >50% reduction – high	<50% reduction ~ Low >50% reduction ~ high
Optimize Energy Efficiency Performance	1~7 – Low 8~15 – High	1~9 – Low 10~18 – High

After data integration, a binomial matrix database with 46 columns (46 credits) and 1382 rows (credit title and 1381 cases) was finally formed.

C. Modeling and Mining

1) Mining Strategy

The information in this study was mined in two ways. The first was from those with high correlation measure values, including high *lift* value, high *DOC* value, high *chi-square* value and high *importance* value. Secondly, *maximal patterns mining* was used to deal with high frequent patterns. The framework of the mining strategy is presented in Fig. 3.

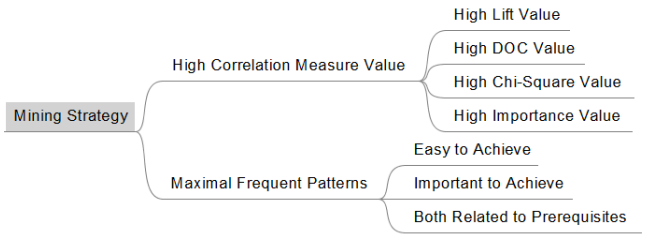


Fig. 3. Mining Strategy

2) Process Control

Technically, the 46 credits will generate 2^46 rules. Obviously, most of them are redundant or not applicable. Three ways are proposed to control the process.

Firstly, the number of items on the right hand side (RHS) of a rule is controlled as one. Because more than one item on the RHS is of little real use. Consider a rule {A, B=> C, D} (name it Rule 1).If the number of items on the RHS is not controlled, all its sub-rule as well as Rule 1 will be generated, say {A, B=> C} – Rule 2, {A, B=> D} – Rule 3. Actually Rule 2 and Rule 3 should be more specific and clear to illustrate the relationship between A, B, C, D. Consider that the possible small amount of extra information in the Multi-item-in-RHS gained is not worth having to coping with a much bigger rule set, it will be better to only look into rules with one attribute on the RHS.

The second is setting the total items in a rule to no more than 3. Considering the 48 credits, there must be numerous rules with 4, 5, 6 or even more items in the LHS (left hand side), but the redundancy is high, say {A, B, C, D=> E} and {A, B, C, E=> D}. Most just switch positions, and it also seems impossible that so many credits have correlations. Furthermore, even the most professional manager will not consider 4 or more credits at the same time.

The last control is about the *maximal frequent patterns* mining. According to the definition, the *support* threshold of *maximal frequent patterns* mining has a negative effect on the number of attributes in the patterns. If the *support* threshold is low, the maximal number of attributes in one pattern will be high. As the *support* threshold rises, the number of attributes that can meet the higher threshold drops. After experimentations, 0.81 was finally set as the *support* threshold in order to control the number of attributes in one pattern to be no larger than three.

IV. EVALUATION AND DISCUSSION

A. Comparison with Professional Opinions

From the LEED user forum (http://www.leeduser.com/), 63 rules about the relationship between LEED-EB&OM V2009 credits were collected. These rules were raised by LEED AP/engineering professionals based on their engineering experience. A simple comparison between these professional rules and the data mining outcomes was made. The result is presented in Table V. Each column represents the number of matched rules within the Top N rank based on one measurement.

TABLE V: MATCHED RULES

Measures	Lift	DOC	Chi-Square	Importance	Total
Top 50	16	18	20	11	29(46.0%)
Top 100	17	20	23	11	33(52.4%)
Top 250	23	26	30	17	44(69.8%)
Top 500	30	34	35	31	54(85.7%)
Top 921	44	48	47	44	63(100%)

TABLE VI: MATCHED BUNDLES

Measures	Lift	DOC	Chi-Square	Importance	Total
Top 50	9	11	12	8	21(41.2%)
Top 100	10	12	15	8	25(49.0%)
Top 250	15	18	21	13	35(68.6%)
Top 500	21	25	26	24	43(84.3%)
Top 921	34	36	35	33	51(100%)

The Bundles in Table VI represent rules without direction. If both forward and backward rules exist, they are regarded as one bundle. The total number of bundles from professional opinions was 51.The last column in Table V and Table VI represents the total number of matched rules or bundles in the study after removing duplicates.

As shown in Table V and Table VI, if taking all the measures into consideration, all the professional rules and bundles could be mined out from the Top 921 ranks (2070 rules in total) and around 50% from the Top 100. It proves the validity of data mining technique.

B. Evaluation and Verification

Fig. 4 shows the relationship between the number of matched rules and the number of the Top N rules in the result.

Those in the Top 100 are more likely to be interesting and reasonable, and match around half number (52.4% and 49.0%) of the rules from the professionals.

In order to check whether the rest of the rules within the Top 100 reveal reasonable relationships or not, the Top 100 rules from the 4 measures were integrated. After removing the duplicates and the previous "matched" ones, there were preliminarily171 rules. In order to make it more manageable, the "repeated" rules were deleted as suggested by Barry and Lin off suggested [8]. That is if rule {A=>B} and rule {B=>A} both exit, then it is called "repeated" and only one rule is picked for detailed analysis. Last, adding the 17 bundles from maximal frequent pattern mining gave 188 in total. A brief result summary is presented in Table VII.

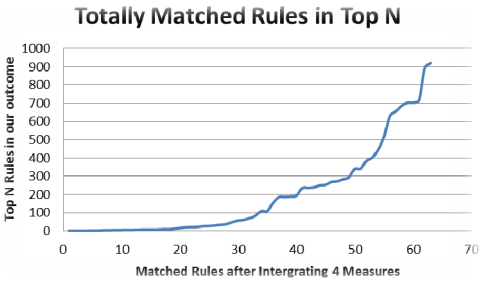


Fig. 4. Totally Matched Rules in Top N

TABLE VII: SUMMARY OF THE ADDITIONAL BUNDLE MINING PROCESS

Process	From High Correlation Measure Value	From Maximal Frequent Pattern	Total	We Think Reasonable
Bundle Num.	171	17	188	25

Finally, 188 bundles were looked into and those we thought reasonable were selected (25 remained).Then the final 25 bundles were sent to 5 LEED AP/professionals/green building scholars to verify whether they were applicable.

Two green building scholars replied, and Table VIII briefly summarized their opinions. They agreed with most of our suggestions. 16 bundles were rated practical by both, and there were no bundles they both thought inapplicable. Table IX lists several practical examples.

TABLE VIII: OPINIONS FROM TWO GREEN BUILDING SCHOLARS

	Scholar A	Scholar B	Both
Think practical	21	20	16
Not applicable	4	5	0

TABLE IX: BUNDLES EXAMPLES

Bundles	Possible Relation
{EQc3.1,WEc4}	When settling the building cleaning plan and program try to address the chemical treatment in the cooling tower.
{SSc6,WEc4}	Use collected rainwater in the cooling tower.
{EQc31,EQc33,EQc34}	Set cleaning programs or plans in which the cleaning products and materials should meet the requirements of IEQc3.3 and the cleaning equipment should meet the requirements of IEQc3.4.

V. CONCLUSION

A new relationship mining strategy is proposed in this study to discover reasonable credit bundles from the LEED-EB&OM certified projects' credit database and to compare the result with professional opinions. With a 100% match within the top 921 rules, it proved the validity and effectiveness of the data mining technique. What's more, additional reasonable credit bundles were discovered and most were supported by two green building researchers.

On the other hand, there were also some limitations in this study. First, we did not further mine the patterns with more than 3 attributes. Secondly, the cost of obtaining the credit bundles has not been taken into consideration. The budget may have a significant impact on selecting credits. So, further research should be conducted based on the limitations above.

REFERENCES

[1] B. J. Thomas, *Mining Association Rules Between Credits in the Leadership in Energy and Environmental Design for New Construction (LEED-NC) Green Building Assessment System*, in Department of Systems and Engineering Management 2008, Air Force Institute of Technology. pp. 108.

[2] S. Winter. *GSA LEED Cost Study*. (2004). United States General Services Administration (GSA). [Online]. Available: http://www.greenbiz.com/toolbox/reports_third. cfm.

[3] D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*, MIT press, 2001.

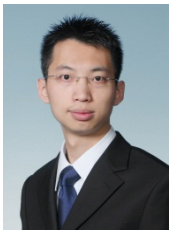
[4] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2006.

[5] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th Int. Conf. Very Large Data Bases*, *VLDB*, 1994.

[6] C. Borgelt, *Apriori–Finding Association Rules/Hyperedges with the Apriori Algorithm*, Working Group Neural Networks and Fuzzy Systems, Otto-von-Guericke-University of Magdeburg, Universitätsplatz, 2004. 2.

[7] M. Hahsler *et al.*, *arules: Mining Association Rules and Frequent Itemsets.* [Online]. Available: http://cran. r-project. org/package= arules. R package version 0.6-8, 2008.

[8] M. J. Berry and G. S. Linoff, *Data Mining Techniques: For Marketing, Sales, And Customer Relationship Management*, Wiley Computer Publishing, 2004.



Jack C. P. Cheng is an assistant professor in Department of Civil and Environmental Engineering at the Hong Kong University of Science and Technology. He received his Bachelor and Master of Philosophy degrees at the Hong Kong University of Science and Technology, Hong Kong. He then received his Doctor of Philosophy degree at Stanford University, USA. His research areas include building information modeling (BIM), knowledge management, data retrieval and mining, green building, sustainable construction and built environment, and service computing for construction management.

Lucky J. Ma was born in Zhejiang, Mainland China. He is now pursuing Master of Philosophy degree at the Hong Kong University of Science and Technology, Hong Kong, China. He received his Bachelor Degree from Huazhong University of Science and Technology, Wuhan, China. His research areas include data mining, green building, and building information modeling (BIM).