A Study on Multi-criteria Association Analysis in Transaction Data

Jin Gyu Shin[1], Dong-Cheol Lee[2] and Doug Won Choi[1]

[1] Department of Systems ManagementEngineering, Sungkyunkwan University
Suwon, Korea
[2] Department of Management Information Systems, Jeju National University
Jeju-si, Jeju-do, Korea

sjg0311@paran.com, dchlee@jejunu.kr, dougwchoi@paran.com

Abstract. This paper suggests a new method reflects multiple criteria to overcome the weakness of Apriori algorithm which counts simply frequency of itemsets in transaction database. We can reduce subjectivity by selecting measurable factors and using average of aggregated weights instead of minimum support as threshold. A method proposed in this paper may be applied to distribution or sales in companies.

Keywords: association analysis, Apriori

1 Introduction

The association analysis is useful for discovering relationships among items which often appear at the same time in the commerce transaction database. It is based on customers' buying patterns and can be used to predict the marketability of the products and establish marketing strategies, etc. The Apriori algorithm is generally used to find associations of items [1], [2]. In this study, an advanced methodology of the Apriori algorithm is proposed and a new algorithm is suggested by using multi-criteria to find associated items.

The majority of association rule mining techniques have been developed to explore simple frequent item sets for computational efficiency [7]. The support which is an important measure to find associated items is based on counting the frequency of the item. But it is not sufficient to only "the frequency measure" to find the meaning of the associated items which maximize the marketing return of companies or organizations. If there is a product which is not sold well and is more profitable, we may not find it when looking at the sales volume. But researches take into account multiple criteria to discover the valuable associated items are rare in the area of the association analysis. Previous research papers tried to develop techniques which assigned weight to each item and calculated the frequency with the weight of each item [4], [5], [6]. The problem of these methods is assigning the subjective weight to each item [3] or the aggregated item weights is heavily influenced by support which makes meaningful purpose as the weights lessen. The focus of the proposed method is selecting measurable factors to eliminate the subjectivities in the process of assigning weight to each item and reflecting the independent weights of factors.

From the enterprise point of view, some measures can be considered. In this paper, we set profit, marketing value, and customer satisfaction and important considerations in distribution or when sales companies sell products besides sales volume. Sales volume data is obtained as a result of transactions. Profit, marketing value, and customer satisfaction of each product affect the transaction. The marketing value is the amount of obtainable assets which trade on the open market for the goods or services. So if a company invests a lot of money on the advertising of goods or services, which affects sales volume. In addition to the marketing value, the profit and customer satisfaction may also be relevant measures in the enterprise as this same view.

Another difference proposed by the Apriori algorithm with other algorithms is normalization in the procedure for data are obtained at different scales. And the average of the aggregated weights is used as threshold instead of randomly determined minimum support to reduce subjectivity.

2 Multi-criteria Association Analysis

Let $''^1 = 1\ I_l n, I_2 n, I;, \ldots \}$ denote the item set consisting of n-itemsets. For example, $//=\{(i), (j), (k), \quad \}$, $/2=\{(i, j), (j, k), (k, 1), \ldots \}$, $I^3=\{(i, j, k), (k, 1, m), \quad \}$ and the subscript t of $I_t n$ indicates the t-th element of f. For instance, if n=3 in this example, $11^3 = (i, j, k)$, $/2^3 = (k, 1, m)$.

The following symbols represent the strategic factors used in this paper and these are derived through interviews and literature survey.

$11(r)$: normalized profitability of t-th itemset $/1^1$, n=1, 2, 3, ...
$11(I_t n)$: normalized strategic marketing value of It^n
$Z(It^n)$: normalized customer satisfaction for It^n
$W(I_i n)$: strategic importance adjusted aggregate weight for item set $I_t n$

Let P_i, p_i be the raw profitability of itemset i, j and $/^3=\{(i, j, k), (k, 1, m)\}$. Then = $(i, j,$ k), $/2^3 = (k, 1, m)$ and we define $P(I_1^3) = P_i + P_j + Pk$ for the sake of computational simplicity. Now we know that $P(12^3) = Pk+P/-1-P_{,n}$. And the relationship between normalized profitability $11(//^1)$ and raw profitability $P(/,^3)$ is defined as follows.

$$1(/1^1) = P(It^n)\ I\ ;^{13}(^1 1^1), \ j=1..n$$

Let v,, v_j be the raw strategic marketing value of itemset i, j. In the same token, we can define the relationship between normalized strategic marketing value $la(I_t n)$ and raw strategic marketing value $V(I_t n)$ as follows.

$$= V(I_t n) / V(I^n)$$

And say that C,, C_1 are the raw value of customer satisfaction of itemset i, j. Then the relationship between normalized customer satisfaction z(V) and raw customer satisfaction $C(I_t n)$ can be defined as follows.

$_x(I,n) = C\ (I_t\text{-})\ l$;

s(hn) is the normalized frequency of occurrence of itemset I (i.e., normalized support). If itemset III appears ***f;*** times in the database D, then

$S(//^1) \quad / E_i(t)$

Let w_k be the weight placed on item characteristics kE x, S} and W(If) be the aggregate weight for item set ***I;*** then $W(I_t")$ is defined as follows.

$W(it^n) = oln(it^n)\pm(^0p(it^n)\pm^04/t^n)\pm(^0Ait^n)$

where $(1)_7, + \quad + co_x + (O_s = 1$

Now the above formula shows that $W(//^1)$ is the aggregation of three strategic factors (profitability, strategic marketing value, and customer satisfaction) and one support value, where each factor is weight adjusted by its respective strategic importance.

Lastly, we need to define the minimum aggregate weight (threshold) R, that is required of an 'n-candidate itemset' to be an 'n-frequent itemset.' Let 1/"I denote the cardinality of n-candidate itemset. In this paper, for the sake of simplicity, we define R_n as follows.

$R_n = 1\ I$

Aggregate weight W(r) of n-candidate itemset is used instead of itemset frequency count. Those candidate itemsets which have $\boldsymbol{W(I_t nW?_n}$ become n-frequent itemsets.

3 Conclusion

We try to reflect more factors in association analysis besides frequency in association. We focused on the semantic and strategic aspects of association analysis, but it is not follows Apriori algorithm's main principle (downward closure property) which is that subsets of a frequent itemset are also frequent [6]. This principle is for reducing computational complexity of frequent itemset generation. Therefore, it is necessary to find more measures which are governed downward closure to follow Apriori algorithm. And the proposed method is needed to apply real commerce transaction data for validation.

Acknowledgments This work was supported by the research grant of Jeju National University in 2012

References

1. Agrawal, R., Imielinski, T., Swami, A.: Database Mining: A performance perspetive, IEEE Transactions on Knowledge and Data Engineering, 5 (1993) pp. 914--925
2. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Database, Proceedings of ACM SIGMOD, vol.22, no.2 (1993)

3. Yue, S., Tsang, E., Yeung, D., Shi, D.: Mining Fuzzy Association Rules with Weighted Items, IEEE International Conference on Systems, Man and Cybernetics, vol.3, pp.1906--1911 (2000)
4. Park, J.S., Yu, W.K., Hong, K.H.: Association rule mining and its applications, Journal of the Korea information science society, vol.16 (1998)
5. Cai, C.H., Ada, W.C., Chen, C.H., Kwong, W.W.: Mining Association Rules with Weighted Items, Proc. of 1998 Intl. Database Engineering and Applications Symposium (IDEAS'98) (1998), Cardiff, Wales, U.K., pp. 68--77
6. Wu, J., Li, X.: An Effective Mining Algorithm for Weighted Association Rules in Communication Networks, Journal of Computers, vol.3, no 10 (2008) pp. 20--27
7. Han, J., Kamber, M.: Data Mining Concepts and Techniques, Morgan Kaufmann (2006)