Fast Algorithm for Mining Item Profit in Retails Based on Microeconomic View[1]

Xu Xiujuan[1], Jia Lifeng[1], Wang Zhe[1], Zhang Hongyan[1,2], Liang Shuang[3], Zhou Chunguang[1]
[1]*College of Computer Science, Key Laboratory of Symbol Computation and Knowledge Engineering of the Ministry of Education, Jilin University, Changchun 130012, China*
[2]*Changchun Institute of Technology 130023, China*
[3]*Department of Mechanical Engineering, National University of Singapore, 119260, Singapore*
xuxiujuan666@yahoo.com.cn cgzhou@jlu.edu.cn

Abstract

The microeconomic framework for data mining assumes that an enterprise chooses a decision maximizing the overall utility over all customers. In item selection problem, the store wants to select J item set S that maximizes the overall profit. Based on the microeconomic view, we propose a novel algorithm ItemRank to solve the problem of item selection with the consideration of cross-selling effect which has two major contributions. First, we propose customer behavior model, and demonstrate it with the data of customer-oriented business. Second, we propose the novel algorithm ItemRank which is implemented on the basis of customer behavior model. According to the cross-selling effect and the self-profit of items, ItemRank algorithm could solve the problem of item order objectively and mechanically. We conduct detailed experiments to evaluate our proposed algorithm and experiment results confirm that the new methods have an excellent ability for profit mining and the performance meets the condition which requires better quality and efficiency.

1. Introduction

The microeconomic framework [1] is considered as one of the most promising of these models though only few theoretical frameworks for data mining have been proposed in the literature. Meanwhile, mining association rule [2] is a basic but important operation in previous literatures. However, there are only a few studies concerning how association rule can be beneficial in more specific targets. Recent investigations [3] in the retailing market have shown an increasing interest on how to make decisions by unitizing association rules, which is needed better knowledge about items. Consequently, profit mining was first proposed by Ke Wang *et al.* [4] to solve the problems mentioned above based on microeconomic view.

The problem of optimal item selection [5] is one of main profit mining fields. It guides retailers to discard the items that are losing profit and introduce new items by upgrading the item type timely. Consequently, a meaningful and discrete subset is mined out for maximizing the profit. The cross-selling effect [6] of items has also been noticed by current retailers, because the profit of an item is not only involved in the item itself, but also is influenced by its relative items. Given that some items fail to produce high profit, they might stimulate customers to buy other high-profit items. Consequently, the cross-selling factor which could be achieved by the analysis of historical transactions should be involved in the problem of item selection. Searching for such a relation of items has becoming an important issue. However, the current method of mining associate rules is not enough to support profit mining.

The rest of the paper is organized as follows. In Section 2, we introduce the related work concerning profit mining, followed by the preliminary definitions of optimal item selection problem in Section 3. In Section 4, we demonstrate ItemRank algorithm step by step. First, customer behavior model is proposed and introduced in detail. Second, ItemRank algorithm is described systematically and theoretically. Experimental evaluation is given out in Section 5. In Section 6, we draw the conclusion of ItemRank algorithm.

[1] This work was supported by the Natural Science Foundation of China (Grant No. 60175024) and the Key Science-Technology Project of the National Education Ministry of China (Grant No. 02090). To whom correspondence should be addressed. Email: cgzhou@jlu.edu.cn.

Proceedings of the 2005 International Conference on Cyberworlds (CW'05)
0-7695-2378-1/05 $20.00 © 2005 IEEE

2. Related Work

Many important literatures concerning profit mining have been published. Brijs *et al*. proposed the PROFSET Model [3], took advantage of the cross-selling effect of items to solve the problem of item selection, and pointed out that two important criteria [3] that items in retails shops should meet the basic sale request and bring higher profits is requested to being taken into account during the process of mining profit from retail market. Thereby, how to balance the relationship of these two principles is the core problem of profit mining. Ke Wang *et al*. proposed HAP algorithm [6] to solve the problem of item ranking with the consideration the influence of confidence and profit. Raymond *et al*. proposed the maximal profit problem of item selection (MPIS) [8] which has the goal of mining out a subset of items with the maximal profit and then ameliorates these above drawbacks. However, MPIS problem is too difficult to implement, because it is a NP-hardness problem even in the simplest situation. In other words, although MPIS algorithm could find the best solution, the cost of time is too expressive to be tolerated.

3. Problem Definition

Based on the microeconomic view [1] the framework considers a store with a possible set of items, depending on the decision chosen, contributes different amounts to the overall utility of a decision from the point of view of the enterprise. It is assumed that the contribution of an item is a possibly complicated, function of the data available on that item. The store chooses an item set that maximizes the overall utility over all transactions.

This section introduces many preliminary but important definitions which is essential for the further understanding of problem of item selection.

3.1. Basic Definition

Given a set I of item and a set of transactions, each transaction is a subset of I. An association rule has the form $X \rightarrow I_j$, such that $X \subseteq I$ and $I_j \in I$; the support threshold of association rule is the fraction of transactions containing all items in X and item I_j ; the confidence of association rule is the fraction of the transactions containing all items in set X that also contain item I_j.

3.2. Problem Definition

MPIS is actually the problem of selecting a subset S from a given item set to maximize the estimated profit of the resulting selection.

Given the dataset which contains m transactions: t_1, $t_2, \ldots, t_m$, and the item set I which includes n items: I_1, $I_2, \ldots, I_n$. The profit of item I_a in transaction t_i is denoted by $prof(I_a, t_i)$. A subset of I, denoted by S, means a set of selected items. We also define two item sets: $t_i' = t_i \quad S$ and $d_i = t_i - t_i'$. Consequently, t_i' represents the selected items in the transaction t_i and d_i represents the no-selected items in the transaction t_i. If d_i is empty, all items in t_i are selected and the profit of t_i is unchanged. If t_i' is empty, we don't selected any item in t_i and thus t_i generates no profit. If both t_i' and d_i are not empty, we stipulate d_i, $\Diamond d_i$ and t_i' as follow: $d_i = \{Y_1, Y_2, \ldots, Y_q\}$, $\Diamond d_i = \{Y_1 \vee Y_2 \vee Y_3 \vee \ldots \vee Y_q\}$, and $t_i' = \{I_1, I_2, \ldots, I_k\}$, where Y_i ($1<i<q$) represents a single no-selected item.

Definition 1 Total Profit of Item Selection [8]: The total profit of an item selection S is defined as below formula:

$$P = \sum_{i=1}^{m} \sum_{I_a \in t_i'} prof(I_a, t_i)(1 - csfactor(d_i, I_a)) \qquad (1)$$

We specify the cross-selling effect (denoted by *csfactor*) of some items for other items by *loss rule* [8] which has a form of "$I_a \rightarrow \Diamond d_i$". For any an item I_a contained by t_i', the loss rule $I_a \rightarrow \Diamond d_i$ indicates that a customer who buys the item I_a must also buy at least one of the items in d_i. Based on the reasoning above, the higher the confidence of $I_a \rightarrow \Diamond d_i$, the more likely the profit I_a in t_i should not be counted. Consequently, the profit in selected set S can also be defined as follow:

$$P = \sum_{i \in T} \sum_{a \in t_i'} prof(I_a, t_i)(1 - conf(I_a \rightarrow \Diamond d_i)) \qquad (2)$$

In this paper, the loss rule $I_a \rightarrow \Diamond d_i$ is treated as a special kind of association rule. Therefore, the confidence of this rule is defined in a similar manner as that of association rule.

Definition 2 The confidence of loss rule [8]: The confidence of a loss rule, $I_a \rightarrow \Diamond d_i$, is defined as below formula

$$\textit{The confidence of loss rule} = \frac{\textit{no. of } \tan \textit{sactions containing } I_a \textit{ and any element in } d_i}{\textit{no. of transactions containing } I_a} \qquad (3)$$

4. ItemRank Algorithm

In this section, a new algorithm called ItemRank which successfully solves the item ranking problem which considers the influence of cross-selling among items is demonstrated step by step. First, customer behavior model is proposed and constructed for the preparation of the novel algorithm. Second, the novel algorithm which is in charge of ranking items is demonstrated theoretically.

4.1. Customer Behavior Model

Here we pay attention to the individual behavior, i.e. reflected information of a transaction. In the domain of profit mining, it is assumed that a customer enters the store and selects items randomly. When a customer finishes selecting a set of items, he or she will begin to select another set randomly. Therefore, the possibility of an item to be selected, in the view of our algorithm, is the rank of the item (ItemRank, abbr IR).

Based on the discussion so far, it is necessary for a retail shop to choose some items to buildup a list which includes items with a high possibility to be brought, and the list guarantees that the higher possibility the items in it are bought, the higher IR of the item.

In microeconomic view [1], the transaction of an individual has an influence with the others and the others also influence it. Customer behavior model is proposed on the basis of the regressive relationship that if an item is influenced by many high-profited ones, it must be a high-profited item. Customer behavior model is denoted by a directed graph G = (V, E), where the set V of vertices denotes the items and the set E of directed edges denotes cross-selling effect between items. N_i denotes the out-degree of item I_i. In customer behavior model, we stipulate that an edge from I_i to I_j means that I_i gets the authority from I_j, and item I_j is relatively important if there are many such edges. If many items point to an item, its IR will be high, or there are some high IR items pointing to it. When we enlarge other items based on items in the list, these items will have high IR. Consequently, the weights of items could be transferred among the items based on the association rules.

4.2. ItemRank Process

Based on customer behavior model, ItemRank is proposed on the basis of PageRank [9] [10] algorithm to solve item selection problem. The initial item rank IR(I_i) is defined as the following equation (4)

$$IR(I_i)=\frac{IR(I_1)}{N_1}+...+\frac{IR(I_n)}{N_n}=\sum_{(i,j)\in E}\frac{IR(I_j)}{N_j} \tag{4}$$

However, the equation (4) fails to consider the confidence of items, *conf(I_i→I_j)*, and the profit of items itself, *prof(I_i)*. The number of frequent items is far more than the selected items, so the potential cross link, $I_i \rightarrow I_j$, depends on not only the confidence of items, but also the profit of items itself. Consequently, if the link $I_i \rightarrow I_j$ exists, we use "*conf(I_i→I_j)×prof(i)*" to evaluate the IR of item I_j, and if item I_j is not selected, the lose profit of itself is also denoted by the value of *conf(I_i→I_j) × prof(i)*, which is viewed as the trust weighs of subsequent items of I_j, so the equation (4) could be rewritten as

$$IR(I_i)=\sum_{(i,j)\in E}\frac{IR(I_j)\times prof\ (I_j)\times conf\ (I_i\rightarrow I_j)}{N_j} \tag{5}$$

The equation (5) is the accurate result only under the precondition that all items must become a strongly connected graph. However, the precondition is difficult to be satisfied because the relation between items is not always in the ideal situation. Moreover, the phenomena of *rank sink* and *rank leak* [9] might happen for the occurrence of loop in the custom behavior model. Consequently, we introduce a damping factor *d*, *d*∈ [0,1], to avoid the phenomenon of *rank sink*. So the equation (5) is remodified to:

$$IR(I_i)=\frac{1-d}{m}+d\sum_{(i,j)\in E}\frac{IR(I_j)\times prof(I_j)\times conf(I_i\rightarrow I_j)}{N_j}, \tag{6}$$

where m is the number of nodes in the directed graph.

The equation (6) guarantees that the rank of item is decided by the trust weights of its subsequent items and all items proportionally.

Note that both the sum of possible distribution of items and the sum of IR of items are equal to 1.

$$Matrix\ \ A=\begin{bmatrix} a_{11}, & ..., & a_{1n} \\ ... & ... & ... \\ a_{n1}, & ..., & a_{nn} \end{bmatrix},\ Unit\ \ Matrix\ \ I=\begin{bmatrix} 1, & ..., & 1 \\ ... & ... & ... \\ 1, & ..., & 1 \end{bmatrix}, \tag{7}$$

$$such\ that\ the\ element\ of\ A, a_{ij}=prof\ (i)\times conf\ (i\rightarrow j)/N_j,$$
$$only\ if\ there\ is\ an\ edge\ from\ i\ to\ j. Otherwise\ , a_{ij}=0.$$

$$\begin{aligned} x=\begin{bmatrix} IR(I_1) \\ ... \\ IR(I_n) \end{bmatrix} &=\frac{1-d}{m}\times\begin{bmatrix} 1 \\ ... \\ 1 \end{bmatrix}+d\times A\times\begin{bmatrix} IR(I_1) \\ ... \\ IR(I_n) \end{bmatrix}=\frac{1-d}{m}\times\begin{bmatrix} 1 \\ ... \\ 1 \end{bmatrix}\times 1+d\times A\times\begin{bmatrix} IR(I_1) \\ ... \\ IR(I_n) \end{bmatrix} \\ &=\frac{1-d}{m}\times\begin{bmatrix} 1 \\ ... \\ 1 \end{bmatrix}\times(\begin{bmatrix} 1, & ..., & 1 \end{bmatrix}\times\begin{bmatrix} IR(I_1) \\ ... \\ IR(I_n) \end{bmatrix})+d\times A\times\begin{bmatrix} IR(I_1) \\ ... \\ IR(I_n) \end{bmatrix} \\ &=(\frac{1-d}{m}\times\begin{bmatrix} 1 \\ ... \\ 1 \end{bmatrix}\times\begin{bmatrix} 1, & ..., & 1 \end{bmatrix}+d\times A)\times\begin{bmatrix} IR(I_1) \\ ... \\ IR(I_n) \end{bmatrix} \\ &=(\frac{1-d}{m}\times\begin{bmatrix} 1 \\ ... \\ 1 \end{bmatrix}\times\begin{bmatrix} 1, & ..., & 1 \end{bmatrix}+d\times A)\times x \\ &=(\frac{1-d}{m}\times\begin{bmatrix} 1, & ..., & 1 \\ ... & ... & ... \\ 1, & ..., & 1 \end{bmatrix}+d\times A)\times x \\ &=Bx \end{aligned} \tag{8}$$

$$whereB=\frac{1-d}{m}\times I+d\times A.$$

Consequently, the item order result could be achieved by computing principal eigenvector of the matrix B. When the iteration number or the IR value of is invariable, ItemRank algorithm is terminated.

Algorithm ItemRank:

```
Input: (1) N items; (2) n transactions; (3) frequent items; (4) all association rules;
Output: (1) S selected items; (2) the corresponding profit coming from S selected items.
(1) Compute the out-degree for each item $I_i$.
(2) Compute the value of matrix B;
(3) Get the principal eigenvector of the matrix B;
(4) Order the principal eigenvector of the matrix B;
(5) Return the items' order.
```

5. Experimental Results

To evaluate ItemRank algorithm, we use the IBM synthetic generator [11] to create a synthetic database, T10.I4.N1K.D10K, with stronger cross-selling factors. Item profit should be normal distributed but we still generated profit random average distribution for simple [6]. Profit of Items are generated as follow: 80% of items have a medium profit ranging from \$1 to \$5, 10% of items have a high profit ranging from \$5 to \$10, 10% of items have a low profit ranging from \$0.1 to \$1. The total profit in this dataset is \$320092. In T10.I4.N1K.D10K, there are 601 frequent items and 10 frequent pairs with minimum support threshold of 0.5%, 872 frequent items and 15398 frequent pairs have the min support 0.1%. It is obvious that there is stronger cross-selling effect between items when support is 0.1%. All experiments are conducted on a 2.8GHZ Intel PC with 512MB main memory, using the Microsoft Windows XP.

5.1. Results for Synthetic Data

ItemRank algorithm is comprehensively compared with HAP algorithm and naïve algorithm [6], because HAP algorithm is the state of the art in the consideration of item selection with cross-selling effect in the data mining literature. ItemRank algorithm outperforms both HAP and naïve approach with both 0.5% and 0.1% min support threshold on the basis of the experiment evaluation.

Form Fig.1, the naïve approach gives the lowest profitability of all algorithms because it simply calculates the profits generated by each item for all transactions and selects some items which generate the greatest profits without considering any cross-selling effect between items. Suppose that all three algorithms get more profit as the increasing of the number of items selected, however, ItemRank could get higher profits than HAP and naïve approach. Fig.2 presents the running time of those algorithms when support is 0.5%. Notice that naïve approach costs least time and naïve approach need no iteration. Although both ItemRank and HAP algorithm need to more or less 20 iterations to get the basic steady eigenvalues of the matrixes, ItemRank algorithm outperforms HAP algorithm as the increasing of item selection.

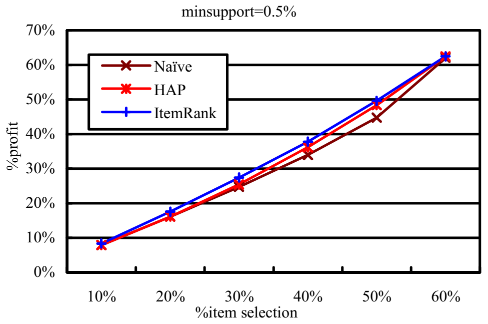

Figure 1. Results with minsupp 0.5%

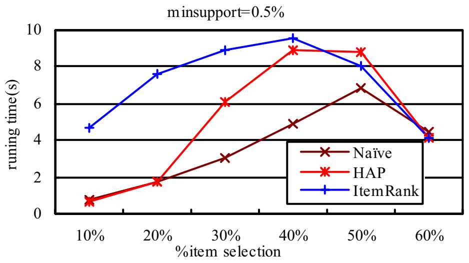

Figure 2. Run Times with minsupp 0.5%

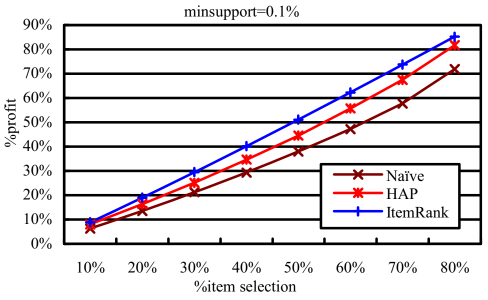

Figure 3. Results with minsupp 0.1%

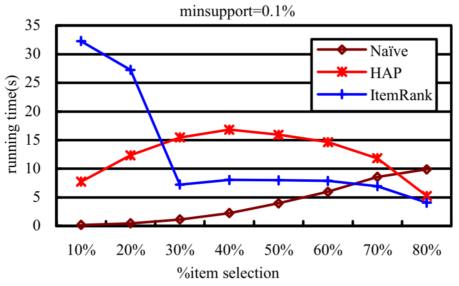

Figure 4. Run Times with minsupp 0.1%

With the support of 0.5%, the damping factor *d* has little effect for items rank, so ItemRank algorithm gets

0-7695-2378-1/05 \$20.00 © 2005 IEEE

almost the same result for different *d* from 0 to 1. However, when the support is 0.1%, ItemRank can get best result when *d* is 0.2. Meanwhile, if *d* is more than 0.3, ItemRank algorithm could get only the same result as HAP. Consequently, we set the damping factor *d* equal to 0.2 for the higher profit. Fig.3 and Fig.4 both indicates that ItemRank has the more excellent performance than HAP and naïve approach when strong cross-selling factor exists between items. Although ItemRank spends the most time when selecting 10% and 20% items, ItemRank is more efficient than both HAP and naïve algorithm with the increasing of selected items.

5.2. Results for HAP Worst-case Set

We now illustrate the goodness of ItemRank by using the specially generated dataset where HAP may get the worst results because of the weaknesses of HAP [12] (In [12], they call the dataset HAP worst-case data set 2 and show how to generate it synthetically). We generate HAP worst-case set by the following parameters: 1,000 items and 10,000 transactions. When minimize support is 0.05%, all items are frequent and there are 2194 frequent pairs. The profit distribution is the same with the above dataset.

In Fig. 5, the parameters in ItemRank algorithm is marked. The iteration number is the number in first bracket and the value of damping factor *d* is in the second bracket.

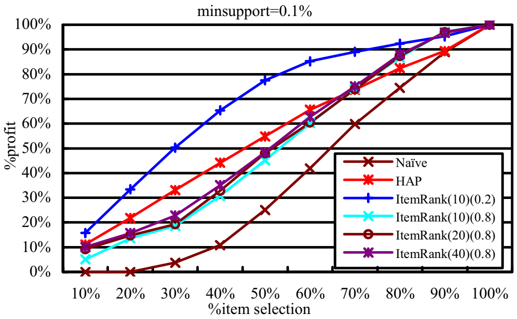

Figure 5. Results with minsupp 0.05%

Obviously, Naïve approach gets the worst result. ItemRank algorithm gets the best result with the damping factor *d* 0.2 after 10 iterations. ItemRank gives the second smallest result with d 0.8 and 10 iterations. If the damping factor d is fixed on 0.8, ItemRank can get higher with the more iterations, However, it could not outperform HAP when the selection items are less than 70%.

From Fig. 5, we maybe draw a conclusion that people usually have 20% chance to purchase a relative item after they have purchased some item.

6. Conclusions and Future work

We use ItemRank algorithm to rank items which are influenced by the cross-selling effect among items based on microeconomic view. Customer behavior model is proposed for the implement of ItemRank algorithm, which presents the better performance than HAP. Comprehensive experiments indicate very satisfied results which confirm that ItemRank algorithm outperforms well-known algorithms: HAP and Naïve.

7. References

[1] J. Kleinberg, C. Papadimitriou, and P. Raghavan, "A Microeconomic View of Data Mining", In: J. Data Mining and Knowledge Discovery, 1998, pp. 311-324.

[2] R. Agrawal, and R. Srikant, "Fast Algorithms for Mining Association Rules", In: Proc. of 20th *VLDB Conference*, 1994, pp. 487-499.

[3] T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets, "The Use of Association Rules for Product Assortment Decisions: a Case Study", In: Proceedings of KDD 98, San Diego (USA), August 15-18, 1999, pp. 254-260.

[4] K. Wang, S. Zhou, and J. Han, "Profit Mining: From Patterns to Actions", Proc. 2002 Int. Conf. on Extending Data Base Technology (EDBT'02), Prague, Czech, March, 2002, pp. 70-87.

[5] R. Wong, and A. Fu, "ISM: Item Selection for Marketing with Cross-Selling Considerations", *PAKDD*, Sydney, Australia, 2004, pp. 431-440.

[6] K. Wang, and M. Su, Item Selection by "Hub-Authority" Profit Ranking. In: Proc. of ACM SIGKDD, 2002, pp. 652-657.

[7] J. M. Kleinberg, "Authoritative Source in a Hyperlinked Environment", In Proc. of the 9th ACM-SIAM Symposium on Discrete Algorithms, ACM, 1998, pp. 668-677.

[8] Wong, R., Fu, A. and Wang, K. "MPIS: Maximal-Profit Item Selection with Cross-selling Considerations", In: Proc. of IEEE ICDM, 2003, pp. 371-378.

[9] S. Brins, and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", In: Computer Networks and ISDN Systems, 30(1-7), 1998, pp. 107~117.

[10] A. Arasu, J. Novak, A. S. Tomkins, and J. A. Tomlin, "Pagerank Computation and the Structure of the Web: Experiments and Algorithms", In: Proc.WWW2002, Honolulu, 2002.

[11] R. Agrwwal, IBM synthetic data generator, http:// www.almaden.ibm.com/cs/quest/syndata.html. 2004.

[12] R. Wong, A. Fu, and K. Wang, "Data Mining for Inventory Item Selection with Cross-selling Considerations", In the Journal of Data Mining and Knowledge Discovery, 2005.