Actionable Combined High Utility Itemset Mining

Jingyu Shao, Junfu Yin, Wei Liu and Longbing Cao
Advanced Analytics Institute, Faculty of Engineering and Information Technology
University of Technology, Sydney, 100 Broadway, Ultimo, NSW, AU, 2007
http://www-staff.it.uts.edu.au/ lbcao/publication/publications.htm
{Jingyu.Shao,Junfu.Yin}@student.uts.edu.au; {Wei.Liu,Longbing.Cao}@uts.edu.au

Abstract

The itemsets discovered by traditional *High Utility Itemsets Mining (HUIM)* methods are more useful than frequent itemset mining outcomes; however, they are usually disordered and not actionable, and sometime accidental, because the utility is the only judgement and no relations among itemsets are considered. In this paper, we introduce the concept of combined mining to select combined itemsets that are not only high utility and high frequency, but also involving relations between itemsets. An effective method for mining such actionable combined high utility itemsets is proposed. The experimental results are promising, compared to those from traditional HUIM algorithm (UP-Growth).

Introduction

A retail or marketing manager often keeps an eye on which brand sells well or what commodity has high utility, and a promotion mixture is proved to be more profitable than a single item. Here, a good promotion mixture is actionable if it is not only profitable (high utility) but also well-selling (high frequency). In this paper, we call such promotion mixtures *actionable itemsets* (Adomavicius and Tuzhilin 1997). Unfortunately, this kind of patterns cannot be easily captured by the existing methods for reasons below.

Typical *Frequent Itemsets Mining* (*FIM*) methods only discover frequent itemsets. *High Utility Itemsets Mining (HUIM)* is more useful to discover those goods with high profits. In *UP-growth* (Tseng et al. 2010), which is a state-of-the-art algorithm in the HUIM area, the results are disordered, even though a UP-Tree with four strategies is proposed to make it much efficient and lossless.

However, to the best of our knowledge, there is no method to extract patterns with both above features simultaneously. More importantly, no HUIM methods consider relations between itemsets, which addresses the disordering issue in outcomes. In *Combined Pattern Mining* (Cao 2011), the concept of combined mining was proposed for solving this sort of problems by considering the interactions between relevant individual itemsets. Involving this concept in actionable HUIM, this paper proposes a combination of two structures, called *Utility-Increasing Incremental Itemsets* (UIII) and *Frequency-Maximum Incremental Itemsets* (FMII), for mining combined actionable high utility itemsets. Figure 1 shows an example of UIII structure: a scatterplot of HUIs, mined by traditional HUIM algorithm (UP-Growth), is firstly built with item-length increasing. The X axis is the length of each itemset, and the Y axis is the utility of each itemset. For example, X_0 is a length-one itemset (only one item included) with utility of 624879, X_a is a length-two itemset (composed of X_0 and ΔX_1) with utility of 876934, and X_b is another length-two itemset with utility of 257377. The dotted lines are linked to each itemset composing several incremental itemset paths. In this figure, only X_a (from X_0) and X_c (from X_b) are UIIIs.

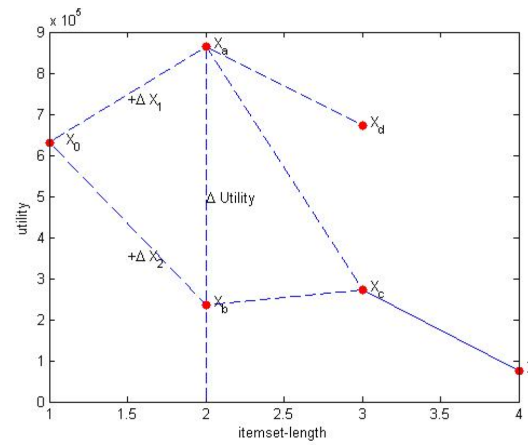

Figure 1: Example of a 4-item itemset with utility dynamics

Combined High Utility Itemsets (CHUI) consist of different itemsets that are extracted from the same itemset component in terms of their utility and frequency perspectives.

$$\begin{cases} X_0 + \Delta X_1 = X_a \\ X_0 + \Delta X_2 = X_b \end{cases} \quad (1)$$

In the CHUI, we call X_a, X_b *Derivative Itemset (DI)*, which is denoted as X in the rest of the paper for global view, X_0 *Underlying Itemset (UI)* and the two variant itemsets ΔX_1 and ΔX_2 *Additional Itemset (AI)* (also denoted as ΔX). Furthermore, $\Delta X_1 \cap \Delta X_2 = \emptyset$ and thus $X_a \neq X_b$.

In the rest of the paper, we propose two functions to mine the range of utility increase (called *contribution*) and the representative degree by frequency (called *weight*) and evaluate the interestingness of each itemset by mean value of the two functions.

Definition Declaration

Notations used in this paper are defined as follows.

- D: The database including all original transactions.
- X,ΔX, X_0: The itemset composed of one or more non-repetitive and non-ordered items.
- $Supp(X)$: The support of an itemset X.
- $u(X_0)$: The utility of X_0 in D.

Proposed Method: CHUIM

Two parameters *contribution* and *weight* are calculated to extract the actionable combined high utility itemsets.

The Contribution of UI to DI

The contribution of X_0 transferring to X is defined as:

$$C(X) = \frac{2}{1+e^{-\mathcal{R}}} - 1 \;\; u(X) > u(X_0) \tag{2}$$

By default, the itemsets are UIIIs. In Eq. (2), $\mathcal{R} = \frac{u(X)}{u(X_0)}$. In other words, C reflects the transferability from X_0 to X. A logistic function is to make the values in the range of $[0, 1]$.

The Weight of DI

The weight of DI X_a is to measure the co-occurrence extent (in another word, the impact of one on another) between UI and AI, which is defined as:

$$W(X) = \frac{Supp(X)}{Supp(X_0 \cup \Delta X)} \tag{3}$$

$Supp(X)$ is the support of the co-occurrence of X_0 and ΔX, and $Supp(X_0 \cup \Delta X)$ is the support of incidents either X_0 or ΔX appears, which is defined as:

$$Supp(X_0 \cup \Delta X) = Supp(X_0) + Supp(\Delta X) - Supp(X_a)$$

The Interestingness Measure

The *Quadratic Mean (QM) (also known as Root-Mean Square)* of $C(X)$ and $W(X)$ is used to measure the significance of an itemset X in terms of both utility and frequency perspectives, which is defined as:

$$QM(X) = \sqrt{\frac{C^2(X) + W^2(X)}{2}} \tag{4}$$

QM is chosen as the integration function of $C(X)$ and $W(X)$ because it represents the sample standard deviation of the difference between predicted values and observed values, thus the result won't be effected heavily by the smaller value. It is easy to prove $QM^2(X)$ = $(\overline{X})^2$+$\sigma^2(X)$. Here, $\overline{X}$ is the arithmetic mean, and $\sigma(X)$ is the standard deviation. In figure (1), for example, the resultant itemset is X_a in Eq. (1) with both high utility and co-occurrence because the frequency of X_c is low although it has increasing utility.

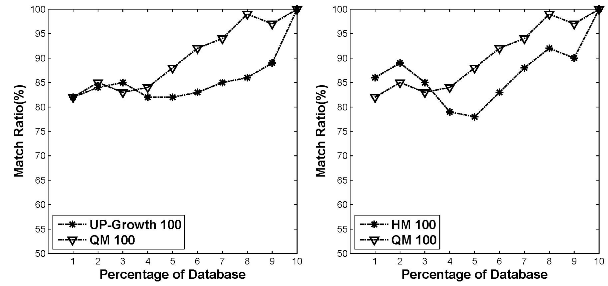

Figure 2: A two-algorithm-comparison on match percentage

Preliminary Outcomes

To evaluate our proposed approach, we conduct experiments on the real datasets downloaded from (Brijs et al. 1999). The utilities of the items are randomly generated as (Tseng et al. 2010). The database is split into 10 parts randomly. The first part contains 10% transactions in the database and each later part contains 10% more than the former part.

The top 100 experimental results are selected and shown in figure 2. The figure on the left shows the comparison between UP-Growth and QM, while the figure on the right shows the result of QM and another mean function Harmonic Mean (HM) on $C(X)$ and $W(X)$. The X axis is the k_{th} part of the database, and the Y axis is the match ratio, which means the ratio of the exact patterns found in the k_{th} part and the $(k+1)_{th}$ part. As can be seen from the figures, the QM method outperforms both HM and UP-Growth.

Conclusion and Future Work

In this paper, we have proposed a novel actionable combined high utility itemset mining method, which integrates the value of frequent pattern mining with high utility pattern mining and involve relations between itemsets. Preliminary experimental results show that the outcomes are more actionable and stable, and avoid accidentally high utility findings that often appear in the classic pattern-growth approaches. We'll further explore more datasets with different characteristics, and a variety of functions and measurements to improve the stability of the resultant actionable combined high utility itemsets.

References

Adomavicius, G. and Tuzhilin, A. 1997. Discovery of actionable patterns in databases: the action hierarchy approach. In *Proc. of Int'l Conf. on ACM SIGKDD*, 111-114.

Brijs, T.; Swinnen, G.; Vanhoof, K.; and Wets, G. 1999. Using association rules for product assortment decisions: a case study. In *Proc. of Int'l Conf. on ACM SIGKDD*, 254-260.

Cao, L.; Zhang, H.; Zhao, Y.; Lou, D.; and Zhang, C. 2011. Combined Mining: discovering informative knowledge in complex data. *IEEE Transactions on SMC Part B* 41(3): 699-712.

Tseng, V. S.; Wu, C. W.; Shie, B. E.; and Yu, P. S. 2010. UP-Growth: an efficient algorithm for high utility itemset mining. In *Proc. of Int'l Conf. on ACM SIGKDD*, 253-262.