Engineering Applications of Artificial Intelligence 77 (2019) 9–20

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

An indexed set representation based multi-objective evolutionary approach for mining diversified top-k high utility patterns

Lei Zhang [a], Shangshang Yang [a], Xinpeng Wu [a], Fan Cheng [a,*], Ying Xie [a], Zhiting Lin [b]

[a] Key Laboratory of Intelligent Computing & Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei 230039, China

[b] School of Electronics and Information Engineering, AnHui University, Hefei 230029, China

ARTICLE INFO

Keywords:
High utility pattern mining
Diversity
Multi-objective optimization
MOEA/D

ABSTRACT

How to discover top-k patterns with the largest utility values, namely, mining top-k high utility patterns, is a hot topic in data mining. However, most of the existing works for mining top-k high utility patterns consider each pattern separately during the mining process, thus many mined patterns are highly similar and lack diversity. In this paper, we propose to mine top-k high utility patterns with high diversity for enhancing users' satisfaction in recommendation. Specifically, we first introduce a simple measure of *coverage* to quantify the diversity of the whole set, that is, the top-k patterns as a complete entity. Then we propose an **i**ndexed **s**et **r**epresentation based **m**ulti-**o**bjective **e**volutionary **a**pproach named ISR-MOEA to mine diversified top-k high utility patterns, due to the fact that the two measures utility and coverage are conflicting. In ISR-MOEA, an indexed set individual representation scheme is suggested for fast encoding and decoding the top-k pattern set. Experimental results on six real-world and two synthetic datasets demonstrate the effectiveness of the proposed approach. The proposed approach can obtain several groups of top-k pattern set with different trade-offs between utility and diversity in only one run, which would further enhance the satisfaction of users.

1. Introduction

Top-k high utility pattern mining, namely discovering k patterns with the largest utility value (or profit) from transaction database, has recently attracted a lot of research work in data mining area (Wu et al., 2012; Yin et al., 2013; Zihayat and An, 2014; Ryang and Yun, 2015; Tseng et al., 2016). In the task of utility mining, each item is associated with a utility (e.g. unit profit) and can appear many times (e.g. quantity) in different transactions. The importance of a pattern can be measured by its utility in terms of weight, value or other information specified by users. Most of the existing works for mining top-k high utility patterns focus on improving the efficiency of the mining algorithms and the mined patterns are considered separately during the mining process. In other words, the item difference between any two patterns is not considered at all in these proposed methods. Thus, the recommended top-k patterns may be very similar and lack diversity.

Take the database shown in Table 1 as an example, if the measure utility is only considered, then the returned top-2 high utility patterns are $\{Jacket, Shoe, Belt\}$ and $\{Jacket, Shoe, Hat\}$. It can be found that there exist many similar items such as $Jacket$ and $Shoe$ among the 2 recommended patterns. In other words, the recommendation of the top-2 high utility patterns lacks diversity, which undermines users' satisfaction. In fact, the following 2 patterns $\{Jacket, Shoe, Belt\}$, $\{Watch, Suitcase, Scarf\}$ may be much better than the top-2 high utility patterns since a practical decision maker system should not only make high utility but also diversified pattern recommendations to improve overall satisfaction of users (Hammar et al., 2013; Wu et al., 2016). Fig. 1 gives two such examples for goods recommendation, where the recommendation-2 by considering both utility and diversity is much better than the recommendation-1 by only considering utility.

To this end, this paper proposes to mine both diversified and high utility patterns for recommendation. In this problem, the recommended top-k patterns are considered as a complete entity, denoted as S. In addition to using utility to measure the profits of S, we introduce a simple measure of coverage to quantify the diversity of S. By considering both utility and coverage, we can obtain both high utility and diversified patterns. However, the measure utility and coverage conflict with each other. In other words, a higher utility of S will lead to lower coverage, whereas a lower utility of S often leads to higher coverage. Due to

* Corresponding author.
E-mail addresses: zl@ahu.edu.cn (L. Zhang), yangshang0308@icloud.com (S. Yang), ahuwxp@163.com (X. Wu), chengfan@mail.ustc.edu.cn (F. Cheng), xieyingahu@126.com (Y. Xie), ztlin@ahu.edu.cn (Z. Lin).

https://doi.org/10.1016/j.engappai.2018.09.009
Received 14 April 2018; Received in revised form 14 August 2018; Accepted 13 September 2018

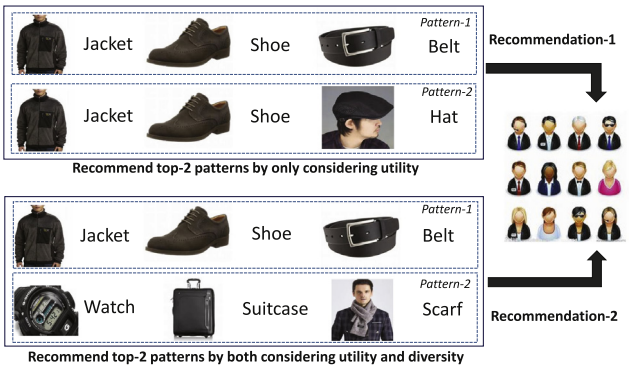

Fig. 1. The examples of two recommended top-2 patterns.

the conflicting property between utility and coverage, an indexed set representation based multi-objective evolutionary approach termed ISR-MOEA is then proposed to obtain diversified top-k high utility patterns. In summary, the main contributions of this paper are described as follows:

- A novel task in utility mining named *mining diversified top-k high utility patterns* is proposed, motivated by the fact that the recommended top-k patterns should not only have high utility but also be diversified. In the suggested task, a measure of coverage is firstly introduced to quantify the diversity of top-k patterns. Then, this task can be formulated as a 2-objective optimization problem since that the two measures utility and coverage are conflicting. To the best of our knowledge, this is the first work by incorporating diversity into top-k high utility mining.
- To tackle this new problem, an indexed set representation based multi-objective evolutionary algorithm named ISR-MOEA is then proposed for mining diversified top-k high utility patterns. In ISR-MOEA, a simple yet effective population initialization strategy is proposed to guarantee that all generated solutions are useful and hold a good diversity for recommendation. In addition, two effective evolutionary operators (i.e. crossover and mutation) are proposed to speed up the convergence of populations.
- The effectiveness and efficiency of the proposed algorithm ISR-MOEA are verified on six real world and two synthetic datasets with different characteristics. Experimental results show the superior performance of our method over several state-of-the-art baseline methods in mining diversified top-k high utility patterns, which indicates that the proposed ISR-MOEA is a competitive and promising method for the task of mining diversified top-k high utility patterns.

The rest of the paper is listed as follows. In Section 2, we introduce the preliminaries about pattern miming, problem formulation, multi-objective evolutionary approach and related work. In Section 3, we give the proposed algorithm in detail. We present our experimental results in Section 4 and conclude the paper in Section 5.

2. Backgrounds and related work

In this section, we first present some preliminaries about utility pattern mining, then give the problem formulations and discuss multi-objective evolutionary approaches. Finally, we give the related work.

Table 1
A toy transaction table for a goods dataset.

t_{id}	Transaction	uti(t_i)
t_1	$\{Jacket[12], Shoe[15], Belt[30], Hat[24]\}$	81
t_2	$\{Jacket[14], Shoe[15], Belt[26], Watch[12]\}$	67
t_3	$\{Jacket[12], Shoe[9], Hat[20], Watch[16]\}$	57
t_4	$\{Jacket[8], Hat[4], Suitcase[15]\}$	27
t_5	$\{Watch[24], Suitcase[27], Scarf[32]\}$	83

2.1. Preliminaries about utility pattern mining

Let transaction database $\mathcal{D} = \{t_1, t_2, \dots, t_m\}$ be a set of transactions, where each transaction t_{id} is a set of items associating with weights (e.g., profits obtained by selling the item). Suppose there are n distinct items in $\mathcal{D}$, denoted as $\mathcal{I} = \{i_1, i_2, \dots, i_n\}$. A *pattern*, denoted as P, is a non-empty set of items. The transactions that contain a pattern P are the *supporting transactions* of P, which is denoted as $\mathcal{D}_P$. The *support* of a pattern P, denoted as $sup(P)$, is the percentage of transactions in $\mathcal{D}$ that contains P. P is called a *frequent pattern*, if $sup(P)$ is no less than a user-specified minimum support threshold *min_sup* ($0 \le min\_sup \le 1$).

The utility of one item i in one transaction t is denoted as $uti(i,t)$. The *utility of a pattern* P in one transaction t, denoted as $uti(P,t)$, is the sum of the utilities of P in t, i.e., $uti(P,t) = \sum_{i \in P} uti(i,t)$. The *utility of a pattern* P, denoted as $uti(P)$, is the sum of the utilities of P in its all supporting transactions, i.e., $uti(P) = \sum_{i \in P \subseteq t, t \in \mathcal{D}_P} uti(i,t)$. The *utility* of a database $\mathcal{D}$, denoted as $uti(\mathcal{D})$, is the sum of the utilities of distinct items in its all supporting transactions, i.e., $uti(\mathcal{D}) = \sum_{i \in t, t \in \mathcal{D}} uti(i,t)$. The *relative utility* of a pattern P, denoted as $ruti(P)$, is the fraction of $uti(P)$ to $uti(\mathcal{D})$, i.e., $ruti(P) = uti(P)/uti(\mathcal{D})$. A pattern P is called *high utility pattern* if $ruti(P)$ is not smaller than a user-specified minimum utility threshold *min_uti* ($0 \le min\_uti \le 1$). The above definitions about frequent and utility pattern mining can be referred in Zhang et al. (2016) and Tseng et al. (2016).

Table 1 gives one example of a transaction database with utilities, where there are 7 unique items *{Jacket, Shoe, Belt, Hat, Watch, Suitcase, Scarf}* and 5 transactions $\{t_1, t_2, t_3, t_4, t_5\}$. Suppose $P = \{Jacket, Shoe, Belt\}$, the supporting transaction database of P is $\mathcal{D}_P = \{t_1, t_2\}$ and $sup(P) = 2/5$. Suppose *min_sup* = 0.30, P is a frequent pattern. The utility of $Jacket$ in transaction t_1 is $uti(Jacket, t_1) = 12$. The utility of P in transaction t_1 is $uti(P, t_1) = 12 + 15 + 30 = 57$, the utility of P in transaction t_2 is $uti(P, t_2) = 14 + 15 + 26 = 55$, thus $uti(P) = 57 + 55 = 112$. The utility of $\mathcal{D}$ is $uti(\mathcal{D}) = 315$. The relative utility of P is $ruti(P) = 112/315 \approx 0.36$. Suppose *min_uti* = 0.30, P is a high utility pattern.

Based on the definition of utility for a single pattern, we can easily extend the utility definition for a pattern set, specifically,

Definition 1. The *relative utility* of the pattern set S containing k patterns is defined as

$$ruti(S) = \frac{1}{k} \sum_{P \in S} ruti(P).$$

Mining top-k high utility patterns: Given a transaction database $\mathcal{D}$, the task of *mining top-k high utility patterns* is to find the pattern set S (containing k patterns) with the largest relative utility value.

In the running example, suppose $k = 2$, the task of *mining top-2 high utility patterns* is to find 2-pattern set S: $\{\{Jacket, Shoe, Belt\}$:112, $\{Jacket, Shoe, Hat\}$:92$\}$. From this example, it can be found that the task of mining *top-k high utility patterns* only considers the utility of each pattern in S, and the diversity of items in the whole pattern set S is not considered. In fact, the recommended top-2 patterns $\{Jacket, Shoe, Belt\}$:112, $\{Watch, Suitcase, Scarf\}$:83 may be more interesting since that this set of patterns is more diversified so as to improve the overall satisfaction of users.

Thus, in this paper, we introduce a measure named *coverage* (Zuo et al., 2015), which can be used to measure the diversity of k-pattern set