**1 Introduction**

The main purpose of Knowledge Discovery in Database (KDD) is to discover meaningful and useful information from a collection of data. Depending on different requirements in various domains and applications, association-rule mining (ARM) is an important and common issue in KDD. Agrawal et al. first designed the well-known Apriori algorithm to mine ARs in a level-wise way. Han et al. then presented FP-growth algorithm to directly mine frequent itemsets without candidate generation. Traditional algorithms of ARM only consider whether or not the item or itemset is present in a transaction.

High-utility itemsets mining (HUIM) incorporates the concept of utility (e.g., worth, value, profit, etc.) of an item or itemset to measure how useful an item or itemset is. An itemset is defined as a high-utility itemset (HUI) if its utility value in the databases is no less than a user-specified minimum utility count. The goal of HUIM is to identify the rare items or itemsets in the transactions, but it can bring valuable profits for retailers.

In real-life applications, data may be uncertainly collected from incomplete data sources, sometimes uncertainty may be the original attribute of the data, such as recommender system of music, video. System can automatically recommend users’ favorite songs which they may not find by themselves. There is correlation between songs are enjoyed by users, therefore, recommendations depend on history in transactions. However, songs in transactions are not always the favorite of users (e.g., a user may find a song which has the same name with his favorite or he may find some songs to listen in random at a time). Therefore, whether the probability of a song in a transaction is the favorite of the user must be considered. For instance, in tabe1 and table2, the probability that “Someone Like You” is the favorite song of the user of Transaction T2 is 0.9. The probability is based on the matching degree between user and song. The higher matching degree is, the larger the probability is.

|  |  |  |
| --- | --- | --- |
| ID | Song Name | Score |
| A | Big Big World | 3 |
| B | Someone Like You | 4 |
| C | Don’t You Remember | 1 |
| D | My Love | 1 |
| E | Time to say goodbye | 8 |
| F | Right Here Waiting | 2 |
| G | Can You | 1 |

1. A Table of Song

|  |  |
| --- | --- |
| User | Record and matching analysis |
| Jack | (A, 0.8) (B, 0.8) (C, 0.7 ) (D, 0.7) |
| Rose | (A, 0.9 ) (C, 0.8) (G, 0.8) |
| Tom | (A, 0.6) (B, 0.7) (C, 0.6) (D, 0.5) (E, 0.8) (F, 0.5) |
| Peter | (B, 0.7) (C, 0.8) (D, 0.5) |
| Linda | (B, 0.8) (C, 0.7) (F, 0.4) |

(b) A table of Record

Figure 1: Example of a larger uncertain transaction database.

Figure 1, in (a) there are the id, name and value which is denoted as the popularity of the song, in (b) there are several transactions in a database, and every transaction is consist of several songs and the probability that it is the user’s favorite.

The existing algorithms of HUIM have been developed to handle a precise database, which is insufficient in real-life applications. An item or itemset is, however, not only present or absent (like or dislike) in the transactions but also associated with an existential (like) probability. In particular, numerous discovered HUIs may not be the interest patterns to help system for recommending favorite songs to users without considering the probability factor. In fact, it is more interested in finding the high existential (like) probability and high popularity patterns.

In this paper, we make the following contributions:

1. To the best of our knowledge, this is the first work to formulate the problem of mining probabilistic high utility itemsits in uncertain databases (MPHU).
2. Due to the challenges from utility constraints, we propose a novel mining framework, called UUH-mine, which only includes an efficient mining algorithm but also contains an effective pruning technique.
3. We verify the effectiveness and efficiency of the proposed methods through extensive experiments on real and synthetic datasets.

The rest of the paper is organized as follows. Preliminaries and our problem formulation are introduced in Section 2. In Section 3, we present a novel mining framework, called UUH-mine. Based on this framework, an efficient mining algorithm and an effective pruning technique is devised. Experimental studies on both real and synthetic datasets are reported in Section 4. In Section 5, we review the existing works and conclude this paper in Section 6.