A Clustering based Prediction Scheme for High Utility Itemsets

Piyush Lakhawat, Mayank Mishra and Arun Somani
Department of Electrical and Computer Engineering, Iowa State University, Ames, Iowa, 50011 U.S.A.

Keywords: High Utility Itemset Mining, Clustering, Itemset Prediction.

Abstract: We strongly believe that the current Utility Itemset Mining (UIM) problem model can be extended with a key modeling capability of predicting future itemsets based on prior knowledge of clusters in the dataset. Information in transactions fairly representative of a cluster type is more a characteristic of the cluster type than the the entire data. Subjecting such transactions to the common threshold in the UIM problem leads to information loss. We identify that an implicit use of the cluster structure of data in the UIM problem model will address this limitation. We achieve this by introducing a new clustering based utility in the definition of the UIM problem model and modifying the definitions of absolute utilities based on it. This enhances the UIM model by including a predictive aspect to it, thereby enabling the cluster specific patterns to emerge while still mining the inter-cluster patterns. By performing experiments on two real data sets we are able to verify that our proposed predictive UIM problem model extracts more useful information than the current UIM model with high accuracy.

1 INTRODUCTION AND MOTIVATION

Itemset mining is an important problem in data mining. The key objective in itemset mining is to identify the frequently occurring patterns of interest in a collection of data objects. Itemset mining is among the areas of data mining which have received high interest in the last decade (Liao et al., 2012). There are two primary reasons for these developments. First, there is a primary need to extract highly repetitive patterns from data in many data mining applications. Second, data mining problems from various domains can be easily modelled as an itemset mining problem. As a result, various application areas like market basket analysis (Ngai et al., 2009), bioinformatics (Alves et al., 2009; Naulaerts et al., 2015), website click stream analysis (Ahmed et al., 2009; Li et al., 2008) etc. have witnessed significant use of itemset mining techniques.

The first model (Agrawal et al., 1994) of itemset mining problem was based on identifying patterns solely on their occurrence frequency. However, a subsequent model emerged (Chan et al., 2003; Liu et al., 2005; Tseng et al., 2010; Tseng et al., 2015) in which utility values were assigned to the data elements based on their relative importance in the analysis. The pattern identification criterion in this new model is a combination of occurrence frequency and utility value. In this work, we enhance the effectiveness of the Utility Itemset Mining model by adding a prediction aspect to it. Having reasonably accurate knowledge of possible future itemsets is of immense value in all applications of Utility Itemset Mining where data is scarce or dynamic in nature and where discovery of knowledge sooner and with lesser amount of data adds much more value to them.The key intuition for this work arises from the existence and knowledge of clusters present in the data. In this work, we show that prior knowledge of the clusters present in the data has high potential to guide the future itemsets discovery.

Building on this idea we propose a prediction scheme for high utility itemsets which captures frequency, utility and cluster structure information to predict the possible future itemsets with high accuracy. Experiments shows that we are able predict a good number of future itemsets with high accuracy over the baseline scheme. While Utility Itemset Mining is not a machine learning problem, but if it were then our contribution would be analogous to the Bayesian version of this problem with the cluster structure acting as the Prior.

Before going into mathematical details of the scheme, we first illustrate the key idea of our work with a small example along with how our contribution adds to the existing itemset mining framework. Itemset mining originated as a formal problem called as

Lakhawat P., Mishra M. and Somani A.
A Clustering based Prediction Scheme for High Utility Itemsets.
DOI: 10.5220/0006590001230134
In *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KDIR 2017)*, pages 123-134
ISBN: 978-989-758-271-4

Frequent Itemset Mining (FIM) from the market basket analysis domain (Agrawal et al., 1994). In FIM, data objects are called transactions. Each transaction contains a set of items along with a transaction ID. Set of items in a transaction are a subset of global set of item types. An itemset is defined as a set of one or more item types. The goal of FIM is to find all itemsets which are present in more than a fixed number (say Φ) of transactions.

For illustration, consider a small example of FIM is presented in Figure 1. Dataset D represents a set of transactions from a retail store. Set I represents various item types. The set of Frequent Itemsets contains all itemsets which are present in two or more (as Φ = 2) transactions in the dataset D. In real world scenarios when the threshold values (Φ) are large, a frequent itemset of type (A, B) leads to an example association rule of type A $\rightarrow$ B. The practical implication of such an association rule depends on the application domain. In market basket analysis it can imply a customer buying item A is likely to buy item B as well, so A and B should be advertised together.

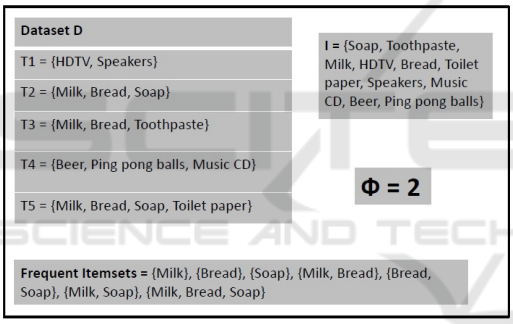

Figure 1: Frequent Itemset Mining.

FIM problem lacks the ability model the relative importance of various item types. For example in Figure 1, above a HDTV and a pack of soap have the same unit of importance. While in reality the profit yield of a unit sale of a HDTV is expected to be much more than that of soap. For a frequency threshold of two, the HDTV and Speakers could not make it to the list of frequent itemsets. Another limitation of FIM is the inability to model the occurrence frequency of items in a particular transaction. For example, it is possible that in transaction T2 from Figure 1 the customer bought one pack of bread, while in transaction T3 customer bought two packs of bread. The FIM problem model is unable to differentiate between these. To overcome these limitations Utility Itemset Mining (UIM) emerged as an evolved version of FIM.

The UIM version of the problem from Figure 1 is presented in Figure 2. In Figure 2, next to items in the

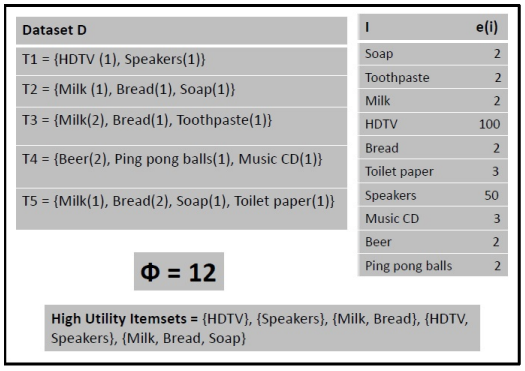

Figure 2: Utility Itemset Mining.

transactions, the parenthesis contain occurrence frequency for the item in that particular transaction. The right side of Figure 2 lists various item types. e(i) represents the relative importance of each item type i. In this case they can be interpreted as the profit associated with the unit sale of that item. The profit associated with an itemset (referred to as absolute utility of the itemset) is the calculated as the sum of profit made by that itemset in all transactions it occurs in. For example, the itemset (milk, Bread) occurs in T2 (profit made = 1 x Milk + 1 x Bread = 4), T3 (profit made = 2 x Milk + 1 x Bread = 6) and T5 (profit made = 1 x Milk + 2 x Bread = 6). Therefore the absolute utility of itemset (Milk, Bread) will be 4 + 6 + 6 = 16. The threshold (Φ) for UIM is a combination criterion of frequency and utility/importance. For the problem in Figure 2, the set of High Utility Itemsets contain all itemsets with absolute utility more than 12 (Φ).

We strongly believe that the UIM problem model can be further extended to add a prediction aspect to it. Let us consider the example in Figure 2. Suppose we have reasonable confidence that the customer for transaction T4 is a college student. Then the information present in T4 is more representative of a customer class of college students than the entire customer population. Leveraging this knowledge can help us predict a latent behavior of college students if present in the data. While ignoring this knowledge leads to information loss due to generalization. This motivated us to investigate ways for leveraging the knowledge of clusters present in data in current UIM model.

1.1 Motivation for a Prediction Enabled UIM Model

Datasets which can be modeled as transactional data have frequently occurring (repeating) patterns of interest in them. This is the key information which itemset mining techniques strive to extract from these datasets. For example, in retail transactions datasets

this information means items which are frequently bought together by customers. However, on the same transactional datasets clustering analysis is performed to study the cluster structure of these datasets. For retail transactions data sets this is the basis of the customer segmentation analysis (Ngai et al., 2009), where similar sets of transactions are clustered together to identify and study various customer types present in the data.

Clustering of transactional type datasets is performed in various biomedical applications as well. Gene expression data is one such example data type which is analyzed using both itemset mining (Alves et al., 2009; Naulaerts et al., 2015) and clustering techniques (Andreopoulos et al., 2009). This implies that itemset mining and clustering study different aspects of the same data set. *While itemset mining abstracts the dataset in form of itemsets, clustering abstracts it in form of clusters of transactions*. Figure 3 presents an illustration of the above idea. If we imagine the dataset to be a solid cylinder, then a top/plan view (corresponding to itemset mining) will show a circle (correspondingly itemsets). While a side/elevation view (corresponding to clustering) will show a rectangle (correspondingly clusters).

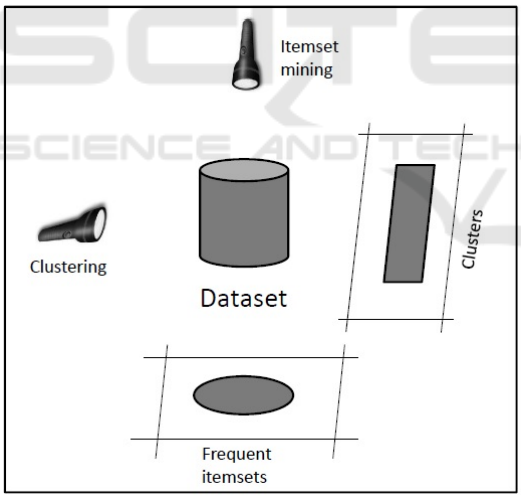

Figure 3: Illustration of different abstractions of dataset.

Performing clustering or itemset mining analysis while ignoring the other creates a handicap as we do not use all available information fully. Recent transactional data clustering techniques are starting to adapt to this fact. For example, a recent transactional clustering algorithm proposed in (Yan et al., 2010) introduces the idea of weighted coverage density. Coverage density is a metric of cluster quality which is used to guide clustering algorithms. Recognizing the fact that frequently occurring patterns are a key characteristic of transactional data, the authors in (Yan et al., 2010) assign weights to items in the coverage density function based on their occurrence frequency. This leads to clusters which are more practically useful. There are two issues if we consider the current UIM problem model and the clustering problem:

1. If we divide the entire set of transactions into clusters and perform itemset mining in each cluster separately, we might miss an inter-cluster pattern.
2. If we perform itemset mining in the whole dataset disregarding clustering, a pattern highly specific to a cluster might be missed due to no support from any other cluster.

This directs to us that we need to somehow implicitly use knowledge the cluster structure while performing itemset mining.

1.2 Need to implicitly use the Cluster Structure

An implicit use of cluster structure of data in itemset mining can potentially address these issues. The knowledge of cluster structure can help identify transactions which are highly representative of a cluster type. The cluster types usually represent some real world entity (for example type of customer). The information in these special transactions is more characteristic of their cluster type than the entire data. Therefore subjecting these transactions to the common threshold in the UIM problem is not appropriate. To overcome this problem, we conclude that some extra importance must be provided to these special transactions. We do this by introducing a new clustering based utility in the definition of the UIM problem model. The modified UIM problem model enables the cluster specific patterns to emerge while still mining the inter-cluster patterns. In essence, we develop a mechanism to enhance the importance (utility) of certain transactions which translates into inflation in utility of certain itemsets. Those itemsets which are enough inflated to cross the threshold will constitute the predictions. This modification in the model can integrate into all UIM techniques as it does not affect the itemset mining part of the techniques.

Revisiting the example in Figure 2, the new predictive UIM model gives extra importance/utility to the items in transaction T4 by identifying it as a special transaction (representative of a college student). Let us assume that the Music CD bought by this college student is of a current hit album. Then the pattern of this Music CD bought along with typical college student items is likely to repeat. This will lead to eventual discovery of this Music CD as high utility item. The predictive UIM model will facilitate a

sooner (using less data) discovery of such items.

In rest of the paper, we first discuss the key works done on the itemset mining problem. Then we formally describe the itemset mining problem followed by the definition of our new clustering based utility to extend the UIM model. We then have a discussion on the use clustering algorithm followed by the experiments on to real data sets before we conclude.

2 RELATED WORK

The problem of itemset mining was first introduced by Agrawal et al in (Agrawal et al., 1994) as frequent itemset mining in context of market basket analysis. They introduced the idea of a downward closure property for generating the potential (candidate) frequent itemsets of size k using the already discovered frequent itemsets of size k-1. This is also popularly known as the apriori technique. This helped to substantially reduce the search space for the frequent itemsets. Building up on this idea many subsequent works extended it by introducing sampling techniques (Toivonen et al., 1996), dynamic itemset counting (Brin et al., 1997), parallel implementations (Agrawal and Shafer, 1996) etc.

A limitation of "apriori" logic based techniques is that sometimes they can generate a large number of candidate itemsets. Since each candidate itemset requires a scan over the entire dataset it also slows the mining process significantly. A popular technique to overcome this issue has been proposed in (Han et al., 2000) called FP-Growth. It performs itemset mining by generating a tree structure rather than candidate generation. There are also techniques proposed which mine the dataset in vertical format (that is list items with sets of transactions) rather than the traditional horizontal format (list of transactions with items). One such work is propose by Zaki in (Zaki, 2000).

Frequent itemset mining lacked important modeling capabilities like relative importance of various items (called utility) and the frequency of an item in a particular transaction, leading to the emergence of utility itemset mining in (Chan et al., 2003; Liu et al., 2005; Tseng et al., 2010; Tseng et al., 2015) among others, where itemsets are mined on the basis of utility support in the dataset rather than frequency support. This makes the problem model more realistic and of higher practical value.

The downward closure property for candidate generation does not apply directly for utility mining. This led to the idea of a transaction weighted utility, which enabled the apriori type candidate generation again. This was the basis of the initial work done in utility mining with subsequent techniques proposed on various strategies for pruning the search space.

The problem of candidate set explosion is also present in these works due to the use of "apriori" logic. To counter this (Tseng et al., 2010) proposes a tree based model called UP-Growth for Utility mining which traverses the dataset only twice.

Recently in (Tseng et al., 2015) authors proposed Utility mining algorithms which use a closed set representation for itemsets which is very concise and yet shows competing performance.

3 ITEMSET MINING PROBLEM MODEL

In this section we formally define the itemset mining problem. We first define the problem of Frequent itemset mining (FIM) followed by Utility itemset mining (UIM).

$$I = \{a_1, a_2, \ldots, a_M\} = \text{Set of distinct item types} \quad (1)$$

$$D = \{T_1, T_2, \ldots, T_N\} = \text{Transaction dataset} \quad (2)$$

where each $T_i = \{x_1, x_2, \ldots\}, x_k \in I$

$$\text{itemset}(X) \text{ of size k} = \{x_1, x_2, \ldots, x_k\} \quad (3)$$

$$SC(X) = |\{T_i \text{ such that } X \in T_i \wedge T_i \in D\}| \quad (4)$$

$$\text{Frequent itemsets} = \{X \text{ such that } SC(X) \geq \Phi\} \quad (5)$$

As mentioned earlier, FIM lacks two key modelling capabilities. It cannot model difference in relative importance of various item types and the frequency of an item type in a transaction. UIM overcomes these limitations. UIM problem builds up on the FIM problem with additional information of external and internal utilities for items. External utility is a measure of unit importance of an item type. This is a transaction independent utility. Internal utility is a transaction specific utility. This is typically the frequency or some measure of quantity of an item in the transaction.

$$eu(a_i) = \text{ external utility of item type } a_i \quad (6)$$

$$iu(a_i, T_j) = \text{ internal utility of } a_i \text{ in } T_j \quad (7)$$

The absolute utility of an item in a transaction is defined as the product of its internal and external utility.

$$au(a_i, T_j) = eu(a_i) * iu(a_i, T_j) \quad (8)$$

Absolute utility of an itemset in a transaction is the sum of absolute utilities of its constituent items.

$$au(X, T_j) = \sum_{x_i \in X} au(x_i, T_j) \quad (9)$$

Absolute utility of a transaction (also called transaction utility) is the sum of absolute utilities of all its constituent items.

$$TU(T_j) = \sum_{x_i \in T_j} au(x_i, T_j) \tag{10}$$

Absolute utility of an itemset in the dataset D is the sum of absolute of that itemset in all transactions that it occurs in.

$$au(X, D) = \sum_{X \in T_j \wedge T_j \in D} au(X, T_j) \tag{11}$$

The set of HUI is the collection of all itemsets which have absolute utility more than or equal to in the dataset D.

$$\text{set of HUI} = \{X \text{ s.t. } au(X, D) \geq \Phi\} \tag{12}$$

The following three concepts are used in the solution techniques of UIM to achieve a downward closure property for efficient candidate generation similar to the FIM problem: Transaction weighted utility (TWU) of itemset X in dataset D is the sum of transaction utilities of transactions in which the itemset X occurs.

$$TWU(X, D) = \sum_{X \in T_j \wedge T_j \in D} TU(T_j) \tag{13}$$

Set of high transaction weighted utility itemsets (HTWUI) is a collection of all itemsets which have transaction weighted utility more than or equal to Φ in the dataset D.

$$\text{Set of HTWUI } = \{X \text{ s.t. } TWU(X) \geq \Phi\} \tag{14}$$

TWDC property ((Tseng et al., 2015; Liu et al., 2005)):"The transaction-weighted downward closure property states that for any itemset X that is not a HTWUI, all its supersets are low utility itemsets."

The goal of UIM is to find the set of all high utility itemsets for a given Φ. Here threshold Φ is a combination criterion of utility and frequency rather than a solely frequency based one in FIM. Figure 2 shows a small example illustrating UIM. The *iu* (internal utility) values for all items are written in parenthesis next to it in the example.

4 A NOVEL CLUSTER BASED UTILITY TO ENHANCE THE UIM MODEL

We discussed in the first section that the goal is to extend the current UIM problem model to add prediction capability to it by implicitly using the cluster structure of data in itemset mining. *Certain transactions are more representative of a cluster type over others.* The information in these special transactions is more characteristic of their cluster type than the entire data. Therefore we do not wish to subject these transactions to the common threshold in the UIM problem. To overcome this problem, we develop a mechanism to attach extra utility to these transactions. We do this by introducing a new clustering based utility in the definition of the UIM problem model. This addition translates into predicting capability of the UIM model.

We define this new utility by calling it cluster utility of a transaction (and the items in it). This is a transaction specific utility for items and is same for all items in a transaction. We introduce following two new concepts in the UIM model before we define the cluster utility.

C as the set of all given clusters. Each cluster is defined as: $C_j = \{T_1, T_2, \ldots\}$. Cluster C_j is a subset of transactions from D.

We also introduce an affinity metric which represents the degree of similarity between a cluster C_j and a transaction T_i.

$$affinity(T_i, C_j) = \text{similarity b/w} T_i \text{ and } C_j \tag{15}$$

These additions to the UIM problem model assume that a fairly accurate cluster structure is given and an appropriate affinity metric is provided. The accuracy here defines an attribute that a cluster structure which portrays the characteristics (repetitive patterns) of interest in the dataset. By appropriateness of the affinity metric we mean a metric which captures the type of similarity (based on constituent items) between a cluster and a transaction that is of interest in the analysis. These assumptions are fairly reasonable as there is a large body of work directed towards of categorical (transactional) clustering. These clustering techniques define subsets of transactions as clusters in the same way as we define them in our predictive UIM problem model. Use of some version of a similarity metric is common for these techniques (Huang, 1998; Guha et al., 1999; Chen and Liu, 2005). The affinity metrics used in them can be used in our extended UIM problem model by interpreting a transaction as single element cluster.

$$cu(a, T_i) = 1 + k * max\{affinity(T_i, C_j) \forall C_j \in C\} \tag{16}$$

In equation 16, k is a tunable parameter and decides how aggressively the cluster information is used in the predictive UIM. Note that the cluster utility is same for all items in a transaction. The rationale behind this definition is to decide the cluster utility of a

transaction based on the cluster which is most similar to it.

We integrate this new internal utility in the calculation of the absolute utilities. The new definition of absolute utility of an item a in a transaction T_i is given by the following:

$$au(a,T_i) = eu(a) * iu(a,T_i) * cu(a,T_i) \quad (17)$$

This implicitly changes the definitions of $au(X,T_i)$, $TU(T_i)$, $TWU(X,D)$, Set of HTWUI, $au(X,D)$ and the set of HUI. All techniques for UIM use the absolute utilities as the building blocks to search for high utility itemsets (Chan et al., 2003; Liu et al., 2005; Tseng et al., 2010; Tseng et al., 2015), so this enhanced predictive UIM problem model will integrate into all of them.

4.1 Impacts of the Enhanced Predictive UIM Problem Model

The following are the impacts of making the above updates to the current UIM model.

1. Assuming that the affinity function to have range [0, 1]. The cluster utility of any item will fall in range [1, 1+k]. Cluster utility closer to 1 will imply their respective transaction to be almost non-representative of any given cluster type. Higher values will imply more similarity of their respective transaction with some given cluster.
2. Since the new definition of absolute utility of an item in a transaction is the product of cluster utility, internal utility and external utility, all absolute utilities will either increase or remain same in the new predictive model.
3. For the same threshold Φ, the predictve model will always find equal or more number of HUI than the current model. Also the set of HUI found by the current model will always a subset of the HUI found by the predictive model.
4. Higher values of parameter k will aggressively use the cluster information and therefore produce more number of HUI. This is recommended when additional emphasis on cluster specific patterns is required.
5. The additional (predicted) itemsets found should be interpreted in the following two ways.
   - When more data arrives later, the additional itemsets found by the model at a previous time are likely to be found in the list of HUI of the current model at that time. The interpretation of this is that a certain pattern(s) are present in particular cluster(s), but with the given amount of data they do not have enough utility support to appear in the list of HUI of the current UIM model. However, with the numbers accumulating with time they will soon show up in the list of HUI in the future. The predictive UIM model recognizes them and helps them getting discovered sooner (with fewer data).
   - If the data is static (or no new data will be available at a later point in time), the additional HUI found in the predictive model are the ones which missed out in the list of HUI of current model due to being specific to only one (or very few) cluster(s) present among many and hence could not gather enough numbers to cross the threshold. However such additional HUI can have application specific importance. For example, a purchase pattern for a specific customer type can be used to create targeted advertisements for those customers.
6. Making this addition modifies the definition of various absolute utilities. However, the use of absolute utilities to find the set of HUI remains the same. Therefore this new model has to ability to be able to be integrated into all UIM techniques.
7. Each cluster in the cluster structure of the data usually represents some real world entity. This has the following implications.
   - Once a satisfactory cluster structure is obtained it can be reused for same type of data. This is because the purpose of cluster structure is only to identify if a particular transaction is fairly representative of a cluster type. This means that the computational expense of clustering need not be repeated every time.
   - The entire dataset might not be needed to obtain an accurate cluster structure. If the size of the dataset is much bigger compared to the cluster structure present in it, then a randomly sampled fraction of dataset is sufficient to capture the cluster structure.
8. The predictive model always finds equal or more HUI than the current model, it can potentially extract the complete set of HUI based on the current model while using fewer data. It can also find additional useful HUIs which the current model missed. This translates into earlier access to actionable information and access to additional useful information.

5 CHOICE OF CLUSTERING TECHNIQUE

Since the proposed predictive UIM problem model assumes the knowledge of an accurate cluster structure and an appropriate similarity metric as discussed in the previous section, it is important to choose a suitable clustering technique. There is a large body of work directed towards clustering of categorical (transactional) data. The clustering techniques return the clusters in form of sets of transactions with similar transactions in each set. A majority of these techniques (Huang, 1998; Guha et al., 1999; Chen and Liu, 2005) employ some similarity metric between the clusters to guide the clustering process using divisive, agglomerative or repartitioning algorithms. The same affinity metrics can be used in the enhanced UIM model by interpreting a transaction as single element cluster. The choice of clustering technique used can be subjective based on the preferences and requirements of the application domain.

Review suggests that certain categorical (transactional) clustering algorithms perform clustering on the basis of frequently occurring patterns in the transactions. Such schemes may be applicable when the external utility information is not very important. However in most real world applications, various item types have different relative importance in the analysis. This is the reason for emergence of UIM as an evolved version of FIM. A better suited clustering technique for use in this enhanced UIM problem model should be based on high utility patterns in the data rather than high frequency ones. We have developed a clustering technique which successfully captures the high utility patterns in the data (Lakhawat et al., 2016). This clustering technique, though not a contribution of the current work, is chosen here due to its strong applicability. An overview of it is provided in the Appendix at the end of the paper. In the next section we perform experiments on two real datasets to evaluate results of the predictive UIM problem model.

6 EXPERIMENTS ON REAL DATASETS

We perform an analysis of the results from the predictive UIM problem model proposed here. We use two real datasets called BMSWebView1 (obtained from (BMSWebView1, 2016)) and Retail dataset (provided by (Brijs et al., 1999) and obtained from (Retail-Dataset, 2016)). BMSWebView1 is a real life dataset of website clickstream data with 59,601 transactions in it. Retail dataset contains 88,163 anonymized transactions from a Belgian retail store. We randomly generated the external utilities (between 1-50) for various item types in both the datasets by using a uniform random number generator. It is common to generate utility values when evaluating algorithms for UIM (Tseng et al., 2015). To obtain the cluster structure to be used for the predictive UIM problem model, we use the utility based categorical clustering algorithm discussed earlier and in the Appendix. For finding the high utility itemsets (HUIs) we implemented a popular UIM technique called the two-phase method (Liu et al., 2005). It essentially finds all the potential HUI using the transaction weighted downward closure property we discussed in an earlier section and then scans the dataset to determine the actual HUIs.

6.1 Experimental Design

We created the following experimental design to compare the effectiveness of our predictive UIM problem model with the current UIM problem model:

1. We create the following 4 versions of both the data sets:
   - Containing first 25% of the data.
   - Containing first 50% of the data.
   - Containing first 75% of the data.
   - Containing the complete data.

   We interpret the complete dataset as all the information which future holds. The purpose of this step is to create scenario where as more data arrives with time it leads to more itemsets being discovered.
2. For each of these datasets we find the set of HUI using the current UIM model. For the retail dataset we use Φ = 50,000 and for the BMSWebView1 data set we use Φ = 20,000. The choice of these threshold values is based on discovering a manageable number of HUI. Higher values of Φ lead to fewer HUI and vice versa. This step establishes the checkpoints for the itemsets discovered by the current UIM model for each version of both the datasets.
3. We generate two cluster structures for both the Retail dataset and the BMSWEbView1 dataset by using 1% and 5% of uniformly randomly sampled data using our clustering algorithm as described before. This step results in a total of 4 cluster structures which will be used to model the predictive UIM problem for each version of the two datasets. The purpose of selecting two different

fractions of datasets in clustering is to observe their effect in the discovery of itemsets.

4. Next we assign the cluster utility to each transaction and their constituent items based on the chosen cluster structure. We do this assignment in a conservative, plain or aggressive manner based on the following criterion:

$$\text{conservative } k = \begin{cases} 0 \text{ if affinity}(T_i,C_j)<0.25 \\ 1 \text{ otherwise} \end{cases} \quad (18)$$

$$\text{moderate } k = 1 \quad (19)$$

$$\text{aggressive } k = \begin{cases} 1 \text{ if affinity}(T_i,C_j)<0.5 \\ 2 \text{ otherwise} \end{cases} \quad (20)$$

5. After assigning the cluster utility we calculate the new values for all absolute utilities. We then find out the set of HUI for each of the above cases based on our predictive UIM problem model (for their respective values) and compare them with the ones found when using the current UIM problem model on the same version of dataset. The key information pieces of interest are:
   - **HUI Found:** This is the number of HUI found by the predictive UIM model for each version of both datasets for the two cluster structures. This will always be equal to or more than the number HUI found using the current UIM problem model.
   - **Additional HUI Found:** This is the additional number of HUIs found by the predictive UIM problem model over the current UIM problem model. This is the most important information of interest. This represents additional itemsets the new model was able to extract using the knowledge of cluster structure of the dataset.
   - **HUI not in Future Data:** This is the number of HUI found by the predictive UIM problem model which are not present in the list of HUI for the current UIM model when using the complete dataset. The HUI in this category represent patterns which are very cluster specific and could not find enough support from the complete data set to cross the threshold . While these itemsets cannot be called high utility itemsets (HUI) in the conventional definition, they do have high utility with respect to their cluster type and they might be very close to crossing the threshold for the current UIM problem model as well. This attribute of these itemset makes a useful set of information.

These results from the above experiment are presented in Table 1 and Table 2.

Table 1: Experiment results: Retail dataset.

Fraction of transactions used in clustering	Cluster utility assignment criterion	HUI found (in 25%, 50% and 75% data respectively)	Additional HUI found (in 25%, 50% and 75% data respectively)	HUI not in future data (in 25%, 50% and 75% data respectively)
0.01	Conservative	28, 53, 91	5, 14, 27	0, 0, 0
0.01	Moderate	30, 61, 99	7, 22, 35	0, 0, 0
0.01	Aggressive	32, 70, 107	9, 31, 43	0, 2, 6
0.05	Conservative	30, 64, 102	7, 25, 38	0, 1, 2
0.05	Moderate	31, 65, 110	8, 26, 46	0, 1, 5
0.05	Aggressive	33, 79, 126	10, 40, 62	0, 3, 19

Table 2: Experiment results: BMSWebView1 dataset.

Fraction of transactions used in clustering	Cluster utility assignment criterion	HUI found (in 25%, 50% and 75% data respectively)	Additional HUI found (in 25%, 50% and 75% data respectively)	HUI not in future data (in 25%, 50% and 75% data respectively)
0.01	Conservative	15, 47, 105	7, 29, 64	0, 5, 23
0.01	Moderate	17, 54, 121	9, 36, 80	0, 6, 38
0.01	Aggressive	17, 57, 124	9, 39, 83	0, 6, 41
0.05	Conservative	15, 49, 107	7, 31, 66	0, 5, 25
0.05	Moderate	17, 54, 121	9, 36, 80	0, 6, 38
0.05	Aggressive	17, 57, 125	9, 39, 84	0, 6, 42

6.2 Key Inferences from the Experimental Results

The following inferences are drawn from the obtained results.

1. Increasing the fraction of transactions used in clustering results in increase of number of HUI found and additional HUI found. This is expected, as with more transactions being used in clustering the cluster structure found is expected to be closer to the true cluster structure of the dataset. This results in more transactions finding higher affinity values with their respective clusters. Higher affinities imply higher cluster based utilities, which further implies higher absolute utilities for itemsets. Higher absolute utilities mean more itemsets are likely to cross the threshold Φ.

   Figures 4 to 7 show the graphical illustrations. The Y-axis shows the HUI found in Figure 4 and Figure 5. Additional HUI found are shown on the Y-axis in Figure 6 and 7. Four different predictive UIM problem models are shown in these figures based on two cluster structures and two cluster utility assignment criterion. The X-axis for these figures shows the dataset version used. Figure 4 and Figure 5 also shows the HUI found when using the current UIM problem model.

2. Varying the cluster utility criterion from conservative to moderate to aggressive results in increase in the number of HUI found and additional HUI found. This is expected, as this stepped variation

results in increase of cluster utility for the transactions. Increase in cluster utility results in increase of absolute utility for itemsets at each step. Increase in absolute utility for itemsets means more itemsets are likely to cross the threshold . A graphical illustration is shown in Figure 4. There are few HUI found (for the predictive model) which are not present in the list of HUI for the complete data (when using the current model) for cases of aggressive cluster utility assignment and especially when using 75% of data. This should interpreted in the correct perspective. Aggressive cluster utility assignment should be used when the analysis is especially focused on discovering all possible cluster specific patterns along with the global patterns. As the current UIM problem model completely disregards the cluster structure, comparison with it in this case becomes less relevant. Furthermore, when we use the 75% version of the data with the predictive UIM problem model, the complete data set is inadequate to verify the validity of the additional HUI discovered and more data might be needed to do so.

3. The predictive UIM problem model extracts significantly more (30% to 50% more for most cases in our experiments when being conservative or moderate in cluster utility assignment) actionable information (HUI) from the data compared to the current UIM problem model. While most of additional HUI found by the new model are found by the current model when additional data is available, few which are not found, are also useful itemsets. These itemsets represent patterns which are specific to cluster types and were not discovered by the current model due to the information loss problem discussed in Section 1. Overall the predictive UIM model leverages the knowledge of the cluster structure while mining for itemsets based on utility and frequency for improved information extraction.

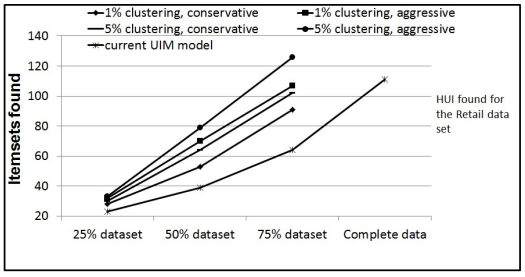

Figure 4: HUI found for the Retail dataset.

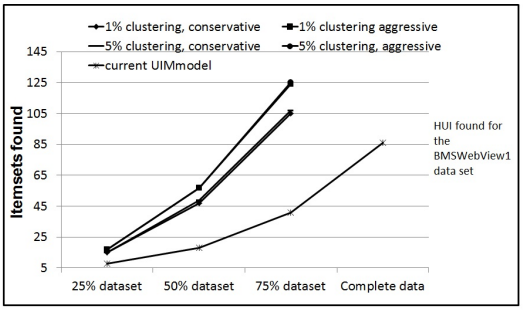

Figure 5: HUI found for the BMSWebView1 dataset.

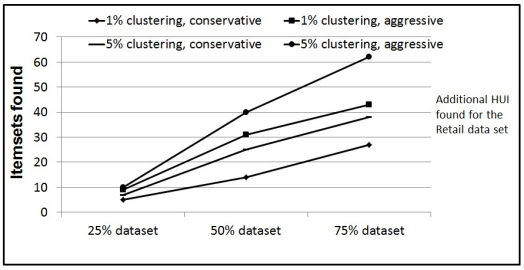

Figure 6: Additinal HUI found for the Retail dataset.

6.3 A Note on Prediction Accuracy

Since we propose this new UIM model as a predictive one, we need to address the accuracy of this prediction with respect to a baseline. Since the current UIM model does not do any prediction, it cannot be considered a baseline. As in our model we are inflating the utility of certain transactions (and hence itemsets), we need to establish that the decision to do it to chosen transactions is better than doing sp uniformly to all transactions. In other words, how much the accuracy suffers if we were to inflate the utility of every transaction in the data. We performed an Itemset search by doing this (inflation by a factor of 3) and discovered that the accuracy suffers heavily. Specifically accuracy here means how many of the predicted itemsets (Addition HUI found) are indeed found to be present in the future data. The inflation by factor of 3 is a baseline for our aggressive cluster utility assignment. For the Retail dataset accuracy dropped to 50.2% (from 96.2%) and 24.9% (from 84.9%) when working on 50% and 75% data respectively. While the for the BMSWebView1 dataset it dropped to a 44.7% (from 89.5%) and 19.6% (from 66.4%) when working on 50% and 75% data respectively. The performance of our predictive model is significantly better (refer Table 1 and Table 2) than these.

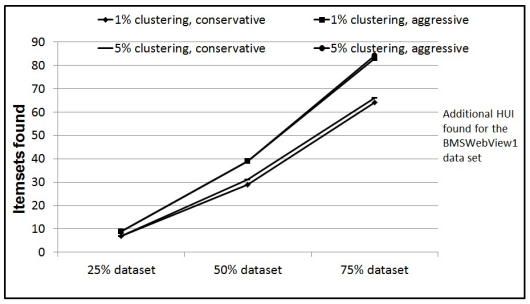

Figure 7: Additional HUI found for the BMSWebView1 dataset.

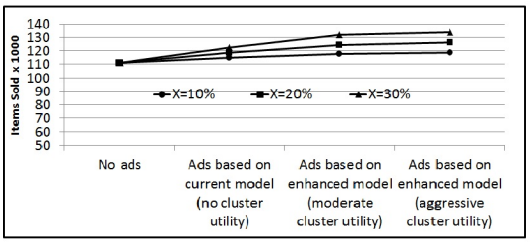

Figure 8: Example impact of UIM model used.

7 EXAMPLE PRACTICAL IMPACT OF THE ENHANCED PREDICTIVE UIM PROBLEM MODEL

Data is used to guide forecasting, planning and decision making in almost all science and business applications. Availability of actionable information is time critical for various reasons ranging from generating more profit for businesses or early release of a drug. Faster processing of the data is one of the ways to achieve actionable information sooner. However when availability of data is the bottleneck (which is the case for many applications in present times), it is most important to extract as much actionable information from the data as possible. With all the data available as well it is always preferred to extract as much useful information from it as possible. We perform an illustrative experiment to demonstrate that the benefit of the predictive UIM problem model.

For illustration, let us assume that for a retail store with no advertising 1000 of items in each HUI are sold every month. With correct advertising assume a X % increase in the sales. By correct advertising we mean advertising based on discovered HUI from the data. Therefore the sales achieved by the store in a month will be based on their choice UIM problem model used in the analysis. For this analysis we use 50% of the Retail dataset with $\Phi = 50000$ and 10% of the transactions for clustering. The results are shown in Figure 8.

8 CONCLUSION AND FUTURE WORK

We establish that the current Utility Itemset Mining (UIM) problem model can be extended by adding a key modeling capability of prediction by capturing cluster specific patterns in the dataset. All transactions possess information in them regarding the degree to which they belong to a cluster of similar objects from the entire data. If a transaction is fairly representation of cluster type then the information in it is more characteristic of their cluster type than the entire data. Therefore ignoring this knowledge and subjecting these transactions to the common threshold in the UIM problem leads to information loss.

We identify that an implicit use of cluster structure of data in the UIM problem model will address the above limitation. We do this by introducing a new clustering based utility in the definition of the UIM problem model and modifying the definitions of absolute utilities based on it. This modified predictive UIM problem model enables the cluster specific patterns to emerge while still mining the inter-cluster patterns and can integrate into all UIM techniques. Through performing experiments on two real data sets we are able to verify that our proposed predictive UIM problem model extracts more useful information than the current UIM model. This enhancement in the UIM problem model leads to improved information extractions by facilitating a sooner (using less data) discovery of HUI and also discovery of cluster specific useful patterns.

For the future work, we plan to study the impact of our new model specific to various applications types in further detail. We also are developing a thorough information theoretic analysis of our model in conjunction with various clustering and UIM techniques.

ACKNOWLEDGEMENTS

The research reported in this paper is funded in part by Philip and Virginia Sproul Professorship Endowment at Iowa State University. The research computation is supported by the HPC@ISU equipment at Iowa State University, some of which has been purchased through funding provided by NSF under MRI grant number CNS 1229081 and CRI grant number 1205413. Any opinions, findings, and conclusions or recommendations expressed in this material are

those of the author(s) and do not necessarily reflect the views of the funding agencies.

REFERENCES

Agrawal, R. and Shafer, J. C. (1996). Parallel mining of association rules. *IEEE Transactions on Knowledge & Data Engineering*, (6):962–969.

Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499.

Ahmed, C. F., Tanbeer, S. K., Jeong, B.-S., and Lee, Y.-K. (2009). Efficient tree structures for high utility pattern mining in incremental databases. *Knowledge and Data Engineering, IEEE Transactions on*, 21(12):1708–1721.

Alves, R., Rodriguez-Baena, D. S., and Aguilar-Ruiz, J. S. (2009). Gene association analysis: a survey of frequent pattern mining from gene expression data. *Briefings in Bioinformatics*, page bbp042.

Andreopoulos, B., An, A., Wang, X., and Schroeder, M. (2009). A roadmap of clustering algorithms: finding a match for a biomedical application. *Briefings in Bioinformatics*, 10(3):297–314.

BMSWebView1 (2016). Smpf: An open-source data mining library, accessed: 2016-06-14. http://www.philippe-fournier-viger.com/spmf/index.php?link=datasets.php.

Brijs, T., Swinnen, G., Vanhoof, K., and Wets, G. (1999). Using association rules for product assortment decisions: A case study. In *Knowledge Discovery and Data Mining*, pages 254–260.

Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. In *ACM SIGMOD Record*, volume 26, pages 255–264. ACM.

Chan, R. C., Yang, Q., and Shen, Y.-D. (2003). Mining high utility itemsets. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 19–26. IEEE.

Chen, K. and Liu, L. (2005). The" best k" for entropy-based categorical data clustering.

Guha, S., Rastogi, R., and Shim, K. (1999). Rock: A robust clustering algorithm for categorical attributes. In *Data Engineering, 1999. Proceedings., 15th International Conference on*, pages 512–521. IEEE.

Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without candidate generation. In *ACM Sigmod Record*, volume 29, pages 1–12. ACM.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3):283–304.

Lakhawat, P., Mishra, M., and Somani, A. K. (2016). A novel clustering algorithm to capture utility information in transactional data. In *KDIR*, pages 456–462.

Li, H.-F., Huang, H.-Y., Chen, Y.-C., Liu, Y.-J., and Lee, S.-Y. (2008). Fast and memory efficient mining of high utility itemsets in data streams. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 881–886. IEEE.

Liao, S.-H., Chu, P.-H., and Hsiao, P.-Y. (2012). Data mining techniques and applications–a decade review from 2000 to 2011. *Expert Systems with Applications*, 39(12):11303–11311.

Liu, Y., Liao, W.-k., and Choudhary, A. (2005). A fast high utility itemsets mining algorithm. In *Proceedings of the 1st international workshop on Utility-based data mining*, pages 90–99. ACM.

Naulaerts, S., Meysman, P., Bittremieux, W., Vu, T. N., Berghe, W. V., Goethals, B., and Laukens, K. (2015). A primer to frequent itemset mining for bioinformatics. *Briefings in bioinformatics*, 16(2):216–231.

Ngai, E. W., Xiu, L., and Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, 36(2):2592–2602.

RetailDataset (2016). Frequent itemset mining dataset repository, accessed: 2016-06-14. http://fimi.ua.ac.be/data/.

Toivonen, H. et al. (1996). Sampling large databases for association rules. In *VLDB*, volume 96, pages 134–145.

Tseng, V. S., Wu, C.-W., Fournier-Viger, P., and Yu, P. S. (2015). Efficient algorithms for mining the concise and lossless representation of high utility itemsets. *Knowledge and Data Engineering, IEEE Transactions on*, 27(3):726–739.

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 *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 253–262. ACM.

Yan, H., Chen, K., Liu, L., and Yi, Z. (2010). Scale: a scalable framework for efficiently clustering transactional data. *Data mining and knowledge Discovery*, 20(1):1–27.

Zaki, M. J. (2000). Scalable algorithms for association mining. *Knowledge and Data Engineering, IEEE Transactions on*, 12(3):372–390.

APPENDIX

C is the set of all given clusters. A cluster $C_k \in C$ is essentially a subset of transactions from D.

$$C_k = \{T_1, T_2 \dots T_k | T_i \in D\} \tag{21}$$

$$I_{C_k} = \{a_i | a_i \in T_j \wedge T_j \in C_k\} = \text{item types in } C_k \tag{22}$$

Cluster utility (CU), relative utility (ru) of a category type in a cluster and the $affinity$ between clusters have the following definitions:

$$CU(C_k) = \sum_{T_j \in C_k} TU(T_j) = \text{Cluster utility of } C_k \tag{23}$$

Input: C ;
while $max_{aff} \geq min_{aff}$ **do**
  for $C_i, C_j \in C$ **do**
    if *affinity*$(C_i, C_j) > max_{aff}$ **then**
      max_{aff} = affinity(C_i, C_j);
      $C_{m1} = C_i$;
      $C_{m2} = C_j$;
  merge(C_{m1}, C_{m2});
  update relevant affinities;
for $C_t \in C$ **do**
  if $\frac{CU(C_t)}{max(CU(C_k) \forall C_k \in C)} \leq min_{uty}$ **then**
    delete C_t;
return C;

Algorithm 1: Clustering algorithm for categorical data with utility information.

since it is the sum of utilities of all transactions in it. CU is an overall measure of importance of a cluster,

$$\forall a_i \in I_{C_k}, ru(a_i, C_k) = \frac{\sum_{a_i \in I_{C_k} \wedge T_j \in C_k} au(a_i, T_j)}{CU(C_k)} \quad (24)$$

ru is the relative importance (since utility is a unit of importance) given to a_i among all I_{C_k} in C_k.

For clusters C_i and C_j:

$$affinity(C_i, C_j) = \sum_{a \in I_{C_k} \wedge a \in I_{C_j}} min(ru(a, C_i), ru(a, C_j)) \quad (25)$$

It is the sum of shared utility of common category types among two clusters. min_{aff} and min_{uty} are tunable parameters of the algorithm. min_{aff} decides the termination criterion of the clustering and min_{uty} decides the final selection criterion for the clusters.