Extracting Diverse-Frequent Patterns with Unbalanced Concept Hierarchy

M. Kumara Swamy[1], P. Krishna Reddy[1], and Somya Srivastava[2]

[1] Centre of Data Engineering,
International Institute of Information Technology-Hyderabad (IIIT-H)
Gachibowli, Hyderabad, India- 500032
[2] Amazon Development Centre, Bangalore, India - 560055
kumaraswamy@research.iiit.ac.in, pkreddy@iiit.ac.in, somya@amazon.com

Abstract. The notion of frequent patterns is employed to extract interesting information about the association among the items in a transactional database. Normally, a pattern may contain items which belong to different categories of a particular domain. For certain types of applications, it may be useful to distinguish between the patterns with items belonging to different categories and the patterns with items belonging to the same category. The existing approaches do not consider the notion of diversity in the pattern. In this paper, we propose an approach to assign the rank, called *DiverseRank*, to the pattern by considering that the items of the pattern belong to unbalanced concept hierarchy. The experiment results show that the proposed approach could extract interesting patterns based on diversity.

Key words: Frequent patterns, diverse rank, concept hierarchy

1 Introduction

In the field of data mining, the process of frequent pattern mining [1, 2] has been widely studied. The related concepts of frequent pattern mining are as follows. Let $I = \{i_1, i_2, \cdots, i_n\}$ be set of n items and D be a database of m transactions. Each transaction contains n items and each transaction is identified with unique identifier. Let $X \subseteq I$ be a set of items, referred as an item set or a *pattern*. A pattern that contains k items is a k-item pattern. A transaction T is said to contain X if and only if $X \subseteq T$. The *frequency* or *support* of a pattern X in D, denoted as $f(X)$, is the number of transactions in D containing X. The support X, denoted as $S(X)$, is the ratio of its frequency to the $|D|$ i.e., $S(X) = \frac{f(X)}{|D|}$. The pattern X is frequent if its support is no less than the user-defined minimum support threshold, i.e., $S(X) \geq minSup$.

The techniques to enumerate frequent patterns and generate a large number of patterns which could be uninteresting to the user. Research efforts are on to discover interesting frequent patterns on constraint-based and/or user-interest by using various interestingness measures such as closed [18], maximal [7], periodic [8, 14], top-k [13], pattern-length [17] and cost (utility) [6].

For certain types of applications, it may be useful to distinguish between the frequent patterns having items belonging to different categories and the patterns with items belonging to the same category. The existing frequent pattern extraction approaches fail to distinguish the patterns based on the diversity of the items within it. In this paper, we have made an effort to propose an improved approach to rank the frequent patterns by analyzing the extent to which the items in the patterns belong to different categories. Generally, in real life scenarios, the concept hierarchies are unbalanced. We have proposed an approach to assign the *DiverseRank* to frequent patterns by considering unbalanced concept hierarchy. The proposed approach is a general approach which can be applied by considering both balanced and unbalanced concept hierarchies. Experiments on the real-world data set show that the proposed *DiverseRank* measure could identify interestingness of patterns with respect to diversity.

In the literature, the concept hierarchies have been used to discover the generalized association rules in [23, 24] and to discover multiple-level association rules in [5]. In [19], a keyword suggestion approach based on the concept hierarchy has been proposed to facilitate the user's web search. The notion of *diversity* has been widely exploited in the literature to assess the interestingness of summaries [20, 21, 25]. In [22], an effort has been made to extend the *diversity-based* measures to assess the interestingness of the data sets using the diverse association rules. The diversity is defined as the variation in the items' frequencies and not according to the categories of items within it. As a result, the *diversity-based measures* cannot be directly applied to mine the *diverse-frequent* patterns. Moreover, the work in [22] has focused on comparing the data sets using diverse association rules.

In [16], an effort has been made to rank the patters based on the category of an item by considering balanced concept hierarchy. In this paper, we have proposed a generalized approach to assign *DiverseRank* to patterns by considering that the items in the pattern belongs to unbalanced concept hierarchy. Through experiments, we have demonstrated the effectiveness of the proposed approach.

The rest of the paper is organized as follows. In the next section, we explain concept hierarchy and the notion of diversity of the pattern. In section 3, we explain the approach to extract diverse frequent patterns with balanced concept hierarchy. In section 4, we present the proposed approach. In section 5, we present experimental results. The last section contains summary and conclusions.

2 About Concept Hierarchy and Diverse-Frequent Patterns

The notion of concept hierarchy plays the main role in assigning diverse rank to frequent patterns. We first explain about concept hierarchy and then diverse frequent patterns.

2.1 Concept hierarchy

A concept hierarchy is a tree in which the data items are organized in a hierarchical manner. In this tree, all the leaf-level nodes represent the items. The items are mapped to next-level items, called categories. The categories, in turn are mapped to next-level nodes. Finally, all categories merge at the *root*. In this paper, we consider the hierarchies like a tree structure, i.e., a lower level node is mapped to only one higher level node.

Figure 1 represents a concept hierarchy. In this, the nodes *orange*, *apple* and *cherry* mapped to *fruits*. Similarly, *drinks* and *fruits* are mapped to *fresh food.* Finally, *fresh food* and *house hold* are mapped to the *root*.

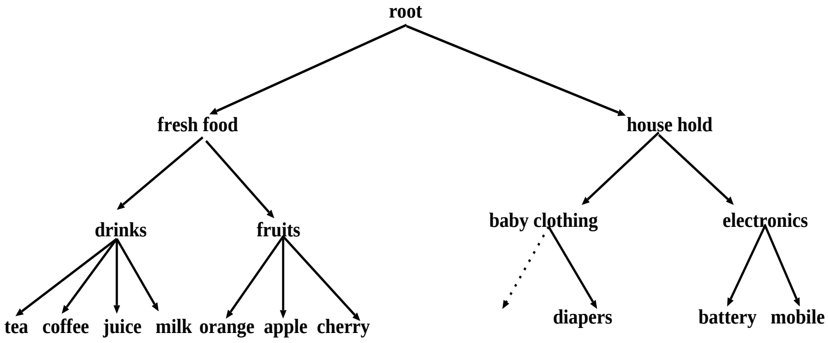

Fig. 1. Balanced Concept Hierarchy

The hight of item and height of concept hierarchy are defined as follows.

- **Height of an item i_k ($h(i_k)$):** Consider an item i_k in the concept hierarchy. The height of an item i_k is denoted as $h(i_k)$ and is equal to the number of levels or edges on the path from *root* to item i_k.
- **Height of a concept hierarchy C ($h(C)$)**: Let C be the concept hierarchy. The hight of the concept hierarchy, denoted as h(C), is equal to the length of the longest path, number of levels, from the *root* to a leaf node.

The concept hierarchies can be balanced or unbalanced.

- **Balanced concept hierarchy:** Balanced concept hierarchy is a tree in which the hight of all leaf level nodes is the same. The hight of balanced concept hierarchy is equal to the hight of a leaf-level item. Figure 1 is an example of balanced concept hierarchy. The hight of all leaf-level nodes is the same.
- **Unbalanced concept hierarchy:** Unbalanced concept hierarchy is also a tree in which the hight of at least one of the leaf level node is different from the hight of other leaf-level nodes. The hight of unbalanced concept hierarchy is equal to the hight of the leaf-level node having largest hight. Figure 2 is the example of the unbalanced concept hierarchy. In this figure, it can be observed that the items *whole milk, 2% milk fat-free milk* are at level-4, the items *pepsi, coke* are at level-3 and the items *shampoo, hair spray, hair oil* are at level-2.

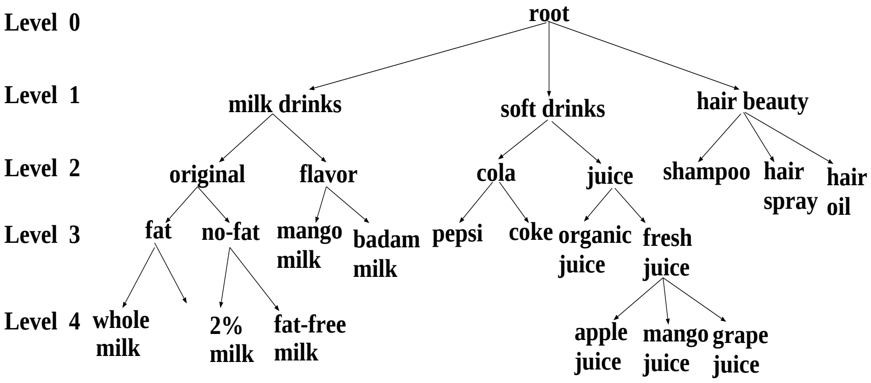

Fig. 2. Unbalanced Concept Hierarchy

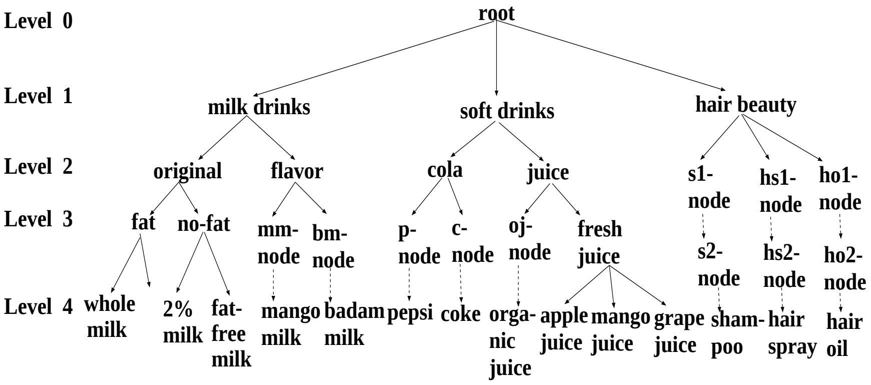

Fig. 3. Extended Balanced Concept Hierarchy for the Figure 2

2.2 Diverse-Frequent Patterns

Given a frequent pattern obtained from a transactional database, its diversity is based on the category of items within it in a given concept hierarchy. All the items of a frequent pattern belong to the leaf level of a concept hierarchy. The categories of the items are extracted from concept hierarchy. If the items of a frequent pattern belong to the same/few categories in a concept hierarchy, we consider that the pattern has low diversity. Relatively, if the items belong to different categories, we consider that the pattern has more diversity. It can be observed that a given frequent pattern moves from the leaf level to the *root* level through a merging process. Several lower level items of a pattern are merged into the corresponding higher level items. In Figure 2, consider {shampoo, hair spray, hair oil}, {pepsi, organic juice, hair oil}, and {whole milk, organic juice, hair oil}. The frequent pattern {whole milk, organic juice, hair oil} is relatively more diverse than the pattern {shampoo, hair spray, hair oil} and {pepsi, organic juice, hair oil}. Since, the items in the pattern {whole milk, organic juice, hair oil} have only one common parent *root*, hence, it is more diverse than the other patterns. Similarly, {pepsi, organic juice, hair oil} is more diverse than the pattern {shampoo, hair spray, hair oil} as the items *pepsi, organic juice* a common parent as *soft drinks*, and finally the merges at *root*. Finally, the {shampoo, hair spray, hair oil} has least diversity of the other patters, as all the items share a common parent at immediate level.

3 Overview of Approach to Calculate *DiverseRank* with Balanced Concept Hierarchy

It can be observed that we take patterns as an input and explain how to assign the *DiverseRank*. The proposed approach assigns *DiverseRank* for a given frequent pattern. In this paper, we use the words *pattern* and *frequent pattern* interchangeably.

The input to the approach [16] is frequent pattern and a concept hierarchy for the items in the transactional database. In case of balanced concept hierarchy, all the items of the database are located at the same level, that is at the leaf level.

We define the terms *Balanced Frequent Pattern* and *Projection of Frequent pattern.*

Definition 1. ***Balanced Frequent Pattern (BFP)****: Consider a frequent pattern* $Y = \{i_1, i_2, \cdots, i_n\}$ *with 'n' items and a concept hierarchy of height 'h'. The* Y *is called balanced frequent pattern, if the height of all the items in* Y *is equal to 'h', i.e.,* $\forall j \in \{1, 2, \cdots, n\} : h(i_j) = k$*, where* k *is a constant* $(0 \leq k \leq h)$*.*

Definition 2. ***Projection of frequent pattern****: For a given frequent pattern* Y *and the corresponding concept hierarchy* C*, the projection of* C *for the pattern* Y *is called* $P(Y/C)$*. The* $P(Y/C)$ *contains the portion of* C *which include all the paths of the items of* Y *from the leaf to root.*

Let Y be BFP, the *DiverseRank* Y is calculated by capturing how the items are merged into higher level items from leaf-level to *root* in $P(Y/C)$. Two notions are employed to compute the diversity of a frequent pattern: *Merging Factor (MF)* and *Level Factor (LF)*.

The notion of generalized-frequent pattern is used to calculate the MF.

Definition 3. ***Generalized-Frequent Pattern (GFP(Y,l, P(Y/C))):*** *Let* Y *be a frequent pattern,* l *be a level and* h *be the height of* $P(Y/C)$ *(where* $0 \leq l \leq h$*). The* $GFP(Y, l, P(Y/C))$ *indicates the* GFP *of* Y *at level* l *in* $P(Y/C)$*. Assume that the* $GFP(Y, l+1, P(Y/C))$ *is given. The* $GFP(Y, l, P(Y/C))$ *is calculated based on the* GFP *of* Y *at the level* $(l+1)$*. The* $GFP(Y, l, P(Y/C))$ *is obtained by replacing every item at level* $(l+1)$ *in* $GFP(Y, l+1, P(Y/C))$ *with its corresponding parent at the level* l *with duplicates removed, if any.*

The MF value indicates how the items of a frequent pattern are merge from lower level to higher level.

$$MF(Y,l) = \frac{|GFP(Y,l,P(Y/C))| - 1}{|GFP(Y,\ l+1,P(Y/C))| - 1} \tag{1}$$

where, Y is a frequent pattern and l is level of a concept hierarchy.

The LF value of P(Y/C) indicates the contribution of items at each level.

$$LF(l,P(Y/C)) = \frac{2*(h-l)}{h*(h-1)} \tag{2}$$

where, h is height of a concept hierarchy, l is a level of a concept hierarchy, $1 \leq l \leq (h-1)$ and $h \neq \{0,1\}$.

The *DiverseRank* of frequent pattern Y for a given concept hierarchy, Y is calculated by summing up the product of MF and LF at every level from leaf-level to the *root* of P(Y/C).

$$DRB(Y,C) = \sum_{l=h-1}^{s+1} MF(Y,l,P(Y/C)) * LF(l,P(Y,C)) \tag{3}$$

where, Y is frequent pattern, h is height of a concept hierarchy, l is a level of a concept hierarchy, s is the level where the number of items in the GFP of Y.

4 Proposed Approach to Calculate *DiverseRank* with Unbalanced Concept Hierarchy

If the items of a pattern contain leaf-level nodes of unbalanced concept hierarchy, the hight of the items is less than or equal to the hight of balanced concept hierarchy. If the hight of all the items are equals to the hight of balanced concept hierarchy, we can extend the approach proposed for balanced frequent patterns.

We explain the approach to assign the *DiverseRank*, if the hight of at least one of the item in the pattern is less than the hight of balanced concept hierarchy. We call such a pattern as unbalanced frequent pattern. The basic idea is to measure the extent of *unbalanced-ness* with respect to the corresponding diversity or *DiverseRank* of the balanced frequent pattern. We first convert the unbalanced concept hierarchy to balanced concept hierarchy, called, extended unbalanced concept hierarchy, by adding dummy items and edges. We calculate the *DiverseRank* of extended frequent pattern by applying the approach proposed for balanced concept hierarchy. Next, we reduce the rank in proportionate to the *unbalanced-ness* of the unbalanced frequent pattern by removing the contribution of dummy items and edges. So, the *DiverseRank* of unbalanced frequent pattern is relative to the *DiverseRank* of the corresponding balanced frequent pattern considering all the nodes are at the level equal to the hight of the unbalanced (or extended) concept hierarchy.

We define the following terms: Unbalanced Frequent Pattern and extended Concept Hierarchy.

Definition 4. ***Unbalanced Frequent Pattern (UFP):*** *Consider a pattern Y and an unbalanced concept hierarchy U of height h. A pattern is called unbalanced frequent pattern, if the height of at least one item in Y is not equal to 'h'.*

Definition 5. ***Extended Unbalanced Concept Hierarchy (EBFP):*** *For a given unbalanced concept hierarchy U with hight h, we convert the unbalanced concept hierarchy into balanced concept hierarchy by adding dummy edges and nodes such that the hight of each leaf-level node is equal to h.*

We have shown the unbalanced concept hierarchy in Figure 4(i) and the corresponding EBFP in Figure 4(ii). In this figure, the solid lines indicate the original edges and the dotted lines indicate the dummy edges, the items with item at higher level nodes indicate the dummy nodes.

The approach to calculate the *DiverseRank* of UFP is as follows. We convert the unbalanced concept hierarchy to the corresponding extended concept hierarchy. Next, we apply the methodology developed for balanced frequent patterns. We adjust the *DiverseRank* by removing the effect of dummy nodes and edges.

Definition 6. ***Projection of Unbalanced Frequent Pattern (PUFP):*** *The projection of UFP is equal to the portion of the corresponding extended concept hierarchy by including all the paths of corresponding items from leaf to root.*

The PUFPs are shown in the 4(iii) to (vi) for the frequent patterns {a,b}, (iv) {b,c}, (v) {b,d}, (vi) {c,d}. Consider the Figure 4 (iii), the items 'a' locate in level-3 and the item locate at 'b' at level-2. In this case, we add a virtual node to bring the item 'b' from level-2 to level-3. We calculate the diversity of the pattern {a,b} for the extended unbalanced concept hierarchy. Finally, the diversity is calculated by removing the effect of dummy edge. Let the Figure 4 (iv) be considered as a pattern {b,c}, the items 'b' and 'c' locate at level-2. Two dummy edges are added to bring both the items to level-3 to calculate diversity. Even though the pattern is balance, the diversity is adjusted with respect to the full balanced. Hence, the $DR(UFP) < DR(BFP)$. The Figure 4(v) for a pattern {b,d}, the items 'b', 'd' are there at level-2 and level-1 respectively in frequent pattern {b,d}. The two items are bring down to the level-3 to calculate the diversity. We remove the effect of dummy edges from diversity with the adjustment factor.

The Figure 2 shows a UFP and Figure 3 shows its corresponding EBFP. After the conversion, the diversity of the frequent pattern can be calculated with diversity of the BFP and adjust the ratio of adjustment factor of UFP.

Adjustment factor: We define the adjustment fator at each level. The *Adjustment Factor (AF)* is an adjustment of the dummy edges with respect to the original edges at the level l. The AF for a frequent pattern Y at a level l should depend on the fact that with respect to the number of edges at l in the EBFP, how many dummy edges at the same level l in the unbalanced concept hierarchy. Therefore, the AF for a frequent pattern Y at level l is the ratio of the number of edges in Y and the number of edges in the corresponding EBFP of Y at l. The AF for Y at l is denoted as $AF(Y,l)$ and is calculated by the following formula.

The value of AF at a level should lie between 0 and 1. For the levels, where the generalization for all the items is same for the EBFP and UFP, the value of AF should be 1.

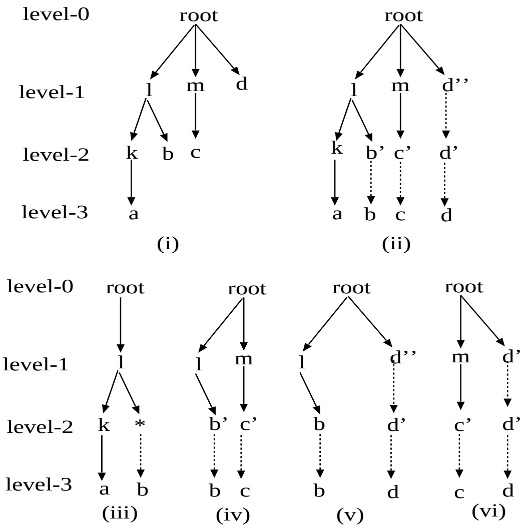

Fig. 4. (i) Unbalanced Concept Hierarchy, (ii) Extended Concept Hierarchy, and Projected Concept Hierarchies for patterns (iii) {a,b}, (iv) {b,c}, (v) {b,d}, (vi) {c,d}.

$$AF(Y,l) = \frac{|EUFP(Y,l)|}{|EEBFP(Y,l)|} \tag{4}$$

where $|EUFP(Y,l)|$ is the number of edges in the frequent pattern Y at level l, $|EBUFP(Y,l)|$ is the number of edges in the corresponding EBFP Y at level l and $0 \leq l < h$.

Example 1: Consider the frequent pattern $Y = \{$*whole milk*, *pepsi*, *coke*, *shampoo*$\}$ in Figure 3. The item which lies deepest in hierarchy is *whole milk*. The height of *whole milk* is 4. As l is between $[0,h)$, we calculate the number of edges at level 3. At the level 3, the number of edges in Y is 1 and the number of edges in corresponding EBFP is 4. Thus, $|EUFP(Y,4)| = 1$ and $|EEBFP(Y,4)| = 1$, i.e., $AF(Y,4) = \frac{1}{4} = 0.25$. At level 3, the value of $|EUFP(Y,3)| = 3$ and $|EEBFP(Y,3)| = 4$. As a result, the value of AF at level 4 is $\frac{3}{4} = 0.75$. Similarly, $AF(Y,2) = \frac{|EUFP(Y,2)|}{|EEBFP(Y,2)|} = \frac{3}{3} = 1$. Also, $AF(Y,1) = \frac{3}{3} = 1$ and $AF(Y,0) = \frac{2}{2} = 1$.

Diverse rank of UFP: From the above discussion, the *DiverseRank* of UFP is a function of MF, AF and LF. The formal definition is as follows.

Definition 7. ***DiverseRank of a frequent pattern Y:*** *Consider Y as a frequent pattern and an unbalanced concept hierarchy of height h_d. The DiverseRank of Y, denoted by $DR(Y)$, is given by the following equation.*

$$DR(Y) = \sum_{l=h_d-1}^{s+1} [MF(Y,l) * AF(Y,l)] * LF(l) \tag{5}$$

where h_d is the height of the item which lies at deepest level in the hierarchy, s is the level where the length of GFP becomes 1, $MF(Y,l)$ is the MF of Y at level l, $LF(l)$ is the LF at level l and $AF(Y,l)$ is the *AdjustmentFactor*(AF) of Y at level l.

5 Experiment Results

The experiments were carried out on the classical R "groceries" market basket analysis data set. The groceries data set contains 30 days of point-of-sale transaction data from a typical local grocery outlet. The data set contains 9,835 transactions and 169 items. The average transaction size in the data set is 4.4. The maximum and minimum transaction size is 32 and 1 respectively.

We conducted the experiments on real-world concept hierarchy to calculate the diversity of the frequent patterns and we also conducted the experiments in clustering. The experimental results are presented as follows.

Real-world concept hierarchy: To generate a concept hierarchy for the items, a web-based Grocery-API provided by Tesco [15] (a United Kingdom Grocery Chain Store) is used. Some of the items that were not listed in the concept hierarchy of Tesco are added manually by consulting the domain experts. The total number of nodes in the concept hierarchy were 220 and the height of the concept hierarchy is 4. The distribution of items at different levels of the concept hierarchy is shown in Table 1.

Table 1. Distribution of items in the unbalanced concept hierarchy.

Level No.	No. of items
0	0
1	1
2	29
3	104
4	34

Table 2. Distribution of items in the simulated concept hierarchy.

Level No.	No. of items
3	14
4	38
5	32
6	12
11	3
12	28
13	38
14	3

In Table 3, we present the list of the top 10 *diverse-frequent* 3-patterns. The second column in Table 3 is the value of *DiverseRank* for the given frequent pattern in the unbalanced concept hierarchy. In the third column, we calculated the *DiverseRank* of its UFP for the given pattern. We call this as the DR of the UFP. The fourth column measures the difference in the values of DR for the given pattern and the DR of the corresponding UFP. The last column is the support value of the pattern. Similarly, Table 4 contains the list of top 10 frequent 3-patterns with respect to the frequency of the patterns (support) along with their *DiverseRank*, the *DiverseRank* of the corresponding UFP and the difference between the two values. Highest support of a frequent 3-pattern is 2.3 (%) and the highest *DiverseRank* value of a frequent 3-pattern is 1. From the two tables, one can observe that there are no common patterns between them.

Table 3. Top 10 *diverse-frequent* 3-patterns along with DR, DR of the corresponding UFP pattern, difference between the two and support value.

Top 10 diverse frequent patterns	**DR of BFP**	**DR of UFP**	**Diff**	**Support (%)**
{soda, whole milk, shopping bags}	0.89	1	0.11	0.7
{rolls-buns, whole milk, newspapers}	0.89	1	0.11	0.8
{rolls-buns, soda, sausages}	1.00	1	0.00	1.0
{rolls-buns, bottled water, other vegetables}	0.89	1	0.11	0.7
{soda, rolls-buns, other vegetables}	1.00	1	0.00	1.0
{rolls-buns, bottled water, yogurt}	0.89	1	0.11	0.7
{rolls-buns, soda, shopping bags}	1.00	1	0.00	0.6
{rolls-buns, soda, whole milk}	0.89	1	0.11	0.9
{rolls-buns, soda, yogurt}	0.89	1	0.11	0.9
{rolls-buns, bottled water, whole milk}	0.89	1	0.11	0.9

Table 4. Top 10 frequent 3-patterns along with DR, DR of the corresponding UFP, difference between the two and support value.

Top 10 diverse frequent patterns	**Support (%)**	**DR of BFP**	**DR of UFP**	**Diff**
{whole milk, other vegetables, root vegetables}	2.3	0.29	0.33	0.04
{yogurt, whole milk, other vegetables}	2.2	0.31	0.33	0.02
{rolls-buns, whole milk, other vegetables}	1.8	0.67	0.75	0.08
{whole milk, tropical fruit, other vegetables}	1.7	0.44	0.5	0.06
{rolls-buns, yogurt, whole milk}	1.6	0.78	0.83	0.05
{yogurt, whole milk, root vegetables}	1.5	0.31	0.33	0.02
{yogurt, whole milk, tropical fruit}	1.5	0.31	0.33	0.02
{whipped sour cream, whole milk, other vegetables}	1.5	0.31	0.33	0.02
{whole milk, pip fruit, other vegetables}	1.4	0.44	0.5	0.06
{soda, whole milk, other vegetables}	1.4	0.67	0.75	0.08

Thus, the results show that the patter having the highest *DiverseRank* value may not be the patterns with the highest support. Similarly, the patterns with the highest support may not have the highest value of *DiverseRank*.

In Tables 3 and 4, the highest difference between the *DiverseRank* of the frequent pattern and the *DiverseRank* of the corresponding UFP is 0.11 and 0.08 respectively. It can be observed that, the difference between the two *DR* values in the tables is very less. This is because the generated concept hierarchy of the transactional data set is not very unbalanced.

Simulated Concept Hierarchy: In this experiment, we simulate the concept hierarchy such that it becomes very unbalanced. To simulate the concept hierarchy, we pushed some of the items deeper in the hierarchy. A random number from the list {1, 1, 1, 1, 2, 2, 2, 9, 9, 9, 10, 10, 10} is chosen to add that many number of edges to increase the height of the respective items. This list is used to ensure that in a frequent pattern, the height difference between highest and deepest item is high. So that one can observe the significant difference between the *DiverseRank* of the UFP and the *DiverseRank* of the frequent pattern.

The height of the new simulated concept hierarchy is 14 and the distribution of items in the concept hierarchy is given in Table 2. The list of some frequent patterns of size-3 with high difference between the values of *DiverseRank* of unbalanced and the corresponding UFP is given in Table 5. From this table, it can be observed that after simulating the hierarchy, the difference between the DR of the unbalanced pattern and the DR value of UFP is as high as 0.4.

Table 5. Frequent 3-patterns with value along with support, DR, DR of corresponding UFP and the difference.

Top 10 diverse frequent patterns	**Support** (%)	**DR of BFP**	**DR of UFP**	**Diff**
{rolls-buns, whole milk, root vegetables}	1.3	0.54	0.93	0.39
{rolls-buns, bottled water, whole milk}	0.9	0.60	1.00	0.40
{rolls-buns, whole milk, yogurt}	1.6	0.54	0.93	0.39
{rolls-buns, soda, whole milk}	0.9	0.60	1.00	0.40
{whole milk, citrus fruit, other vegetables}	1.3	0.47	0.86	0.39
{rolls-buns, whole milk, pork}	0.6	0.56	0.93	0.37
{rolls-buns, whole milk, newspapers}	0.8	0.62	1.00	0.38

6 Summary and Conclusions

Finding interesting frequent patterns is one of the issues in frequent pattern mining. Several interestingness measures have been proposed to extract the subset of frequent pattern according to the needs and demands of the users. Generally, in real life scenarios, the concept hierarchies are unbalanced. In this paper, we have proposed an approach to calculate the DiverseRank of the frequent patterns by considering unbalanced concept hierarchy. The experiments on the real world data set show that the diverse-frequent patterns differ from frequent pattern knowledge.

As a part of future work, we are planning to extend the notion of diverse-frequent patterns to improve the performance of clustering, classification and recommendation algorithms.

References

1. R.Agrawal, T.Imieliński, and A.Swami: Mining association rules between sets of items in large databases, SIGMOD Rec., 22: pp. 207–216, 1993.
2. R.Agrawal, H.Mannila, R.Srikant, H.Toivonen, and A.I. Verkamo: Advances in knowledge discovery and data mining, Chapter Fast discovery of association rules, AAAI, Menlo Park, CA, USA, pp. 307–328, 1996.
3. J.Han, J.Pei, and Y.Yin: Mining frequent patterns without candidate generation, *SIGMOD Rec.*, 29: pp. 1–12, 2000.
4. S.Brin, R.Motwani, and C.Silverstein: Beyond market baskets: Generalizing association rules to correlations, 1997.
5. J.Han and Y.Fu: Mining multiple-level association rules in large databases, *IEEE Transactions on Knowledge and Data Engineering*, Vol.11(5), pp. 798–805 1999.

6. J.Hu and A.Mojsilovic: High-utility pattern mining: A method for discovery of high-utility item sets, *Pattern Recogn.*, 40: pp. 3317–3324, 2007.
7. T.Hu, S.Y. Sung, H.Xiong, and Q.Fu: Discovery of maximum length frequent itemsets, *Inf. Sci.*, 178: pp. 69–87, 2008.
8. R.U. Kiran and P.K. Reddy: Mining periodic-frequent patterns with maximum items' support constraints, In *Proceedings of the Third Annual ACM Bangalore Conference*, COMPUTE '10, New York, USA, pp. 1:1–1:8, 2010.
9. B.Liu, W.Hsu, L.-F. Mun, and H.-Y. Lee: Finding interesting patterns using user expectations, *IEEE Transactions on Knowledge and Data Engineering*, 11(6), pp. 817–832, 1999.
10. K.McGarry: A survey of interestingness measures for knowledge discovery, *Knowl. Eng. Rev.*, 20, pp. 39–61, 2005.
11. E.Omiecinski: Alternative interest measures for mining associations in databases,*IEEE TKDE*, 15(1), pp. 57– 69, 2003.
12. J.Pei, J.Han, and L.V.S. Lakshmanan: Mining frequent itemsets with convertible constraints, IEEE Computer Society, pp. 433–442, 2001.
13. T.M. Quang, S.Oyanagi, and K.Yamazaki: Mining the k-most interesting frequent patterns sequentially, In *IDEAL*, pp. 620–628, 2006.
14. S.K. Tanbeer, C.F. Ahmed, B.-S. Jeong, and Y.-K. Lee.: Discovering periodic-frequent patterns in transactional databases, PAKDD '09, Berlin, Heidelberg, Springer-Verlag, pp. 242–253, 2009.
15. Tesco.: Grocery api. https://secure.techfortesco.com/tescoapiweb/, 2011.
16. S. Somya and R. Uday Kiran and P. Krishna Reddy.: Discovering Diverse-Frequent Patterns in Transactional Databases, COMAD 2011, pp. 69-78, 2011.
17. J.Wang, J.Han, Y.Lu, and P.Tzvetkov: Tfp: an efficient algorithm for mining top-k frequent closed itemsets, *IEEE Transactions on Knowledge and Data Engineering*, 17(5), pp. 652–663, 2005.
18. M.Zaki and C.-J. Hsiao: Efficient algorithms for mining closed itemsets and their lattice structure, *IEEE Transactions on Knowledge and Data Engineering*, 17(4), pp. 462–478, 2005.
19. Y.Chen, G.-R. Xue, and Y.Yu: Advertising keyword suggestion based on concept hierarchy, WSDM '08, New York, USA, ACM, pp. 251–260, 2008.
20. L.Geng and H.J. Hamilton: Interestingness measures for data mining: A survey, *ACM Comput. Surv.*, 38, September 2006.
21. R.J. Hilderman and H.J. Hamilton: Knowledge Discovery and Measures of Interest, Kluwer Academic Publishers, Norwell, USA, 2001.
22. R.A. Huebner: Diversity-based interestingness measures for association rule mining, 2009.
23. R.Srikant and R.Agrawal: Mining generalized association rules, pp. 407–419, 1995.
24. S.Thomas and S.Sarawagi: Mining generalized association rules and sequential patterns using sql queries, In *Proc. of 4th Intl. Conf. on Knowledge Discovery and Data Mining KDD98*, AAAI Press, pp. 344–348, 1998.
25. N.Zbidi, S.Faiz, and M.Limam: On mining summaries by objective measures of interestingness, *Mach. Learn.*, 62, pp. 175–198, 2006.
26. Man Lung Yiu and Nikos Mamoulis: Frequent-Pattern based Iterative Projected Clustering, In *ICDM*, pp.689–692, 2003.
27. Man Lung Yiu and Nikos Mamoulis: Iterative Projected Clustering by Subspace Mining, *IEEE Trans. Knowl. Data Eng.*, (17) 2, pp.176-189, 2005.
28. C. Blake and C. Merz: UCI Repository of Machine Learning Databases, Univ. of Calif., Irvine, Dept. of Information and Computer Sciences, http://www.ics.uci.edu/ mlearnMLRepository.html, 1998.