Using Extended Multidimensional Pattern Relation for Multidimensional On-line Mining

Ching-Yao Wang[1], Shian-Shyong Tseng[1], Tzung-Pei Hong[2] and Yian-Shu Chu[1]
[1]*Institute of Computer and Information Science, National Chiao-Tung University, Taiwan*
[2]*Department of Electrical Engineering, National University of Kaohsiung, Taiwan*
tphong@nuk.edu.tw, {cywang, sstseng, yschu}@cis.nctu.edu.tw

Abstract*-Although incremental data mining and online mining approaches are rather efficient, they usually can not flexibly obtain association rules or patterns from portions of data, diversely consider problems at different aspects, and provide on-line decision supports for users. In the past, we thus developed the* multidimensional on-line mining approach *to select and integrate related mining information for diverse mining requests. It uses the multidimensional pattern relation to structurally and systematically store the additional context information and mining information for each inserted block of data. It may, however, get loose upper-bound supports of candidate itemsets and thus cause datasets to be frequently re-processed for new mining requests. In this paper, we attempt to apply the concept of negative border to enlarge the mining information in the multidimensional pattern relation to help get tight upper-bound supports of candidate itemsets, and thus reduce the number of candidate itemsets to be considered. Based on the extended multidimensional pattern relation, a corresponding online mining approach is proposed.*

1. Introduction

Data mining attempts to discover non-trivial, implicit, previously unknown and potentially useful knowledge [14]. Recently, mining association rules from transaction databases has been one of the most interesting and popular research topics in data mining [2][3][8]. Since the process of mining association rules is rather costly and time-consuming, some famous mining approaches, such as Apriori [4], DIC [6], DHP [20], Partition [22], Sampling [18] and GSP [5], have been proposed to reduce the computation time and improve the performance. However, these approaches process the data in a batch way. They do not utilize previously mined patterns for later maintenance, and may require considerable computation time to obtain the updated set of association rules or patterns [9]. Incremental data mining [9][10][11][16][21][23] and online mining [1][15] have thus become important research topics for data mining in recent years.

Although incremental data mining and online mining approaches are rather efficient, there exist the following limitations for online decision support.

1. Users can not use them to obtain association rules or patterns from their only interested portion of data.
2. Some contexts such as region, time and branch have usually been ignored in mining requests. However, decision-makers usually diversely consider problems at different aspects [12][13][14].
3. Each selected portion from the entire database needs to be re-processed. This will incur inefficiency and delay in responding results.

Assume the data under consideration evolve in a systematic way. For example, data may be inserted or deleted in a block during an interval of a month. In our previous study, we proposed a *multidimensional on-line mining* approach for online generation of association rules under multidimensional consideration [24]. It uses the *multidimensional pattern relation* to structurally and systematically store the additional context information and mining information for each inserted block of data. The multidimensional on-line mining approach may, however, get loose upper-bound supports of candidate itemsets if the itemsets are originally small and not kept in the multidimensional pattern relation. Excessive I/O and computation costs are thus needed to re-process the underlying database. In this paper, we attempt to apply the concept of *negative border* to calculate tighter upper-bound supports of candidate itemsets and get a better pruning effect. The multidimensional pattern relation has to be extended for keeping the additional negative-border information. Based on the extended multidimensional pattern relation, we then develop an online mining approach called *Negative-Border Online Mining* (NOM) to efficiently and effectively utilize the information of *negative itemset* in the negative border. Experimental results show the good performance of the proposed NOM approach.

2. Review of Some Incremental Mining Approaches

In real-world applications, a database grows over time such that existing association rules may become invalid or new implicitly valid rules may appear. In these situations, conventional batch-mining algorithms may require considerable computation time to re-mine the entire updated database to get all up-to-date association rules [9]. Some researchers have then developed incremental mining algorithms to maintain association rules without re-processing the database. Considering an original database and newly inserted transactions, the following four cases may arise:

Case 1: An itemset is large in the original database and in the newly inserted transactions;

Case 2: An itemset is large in the original database, but is not large in the newly inserted transactions;

Case 3: An itemset is not large in the original database, but is large in the newly inserted transactions;

Case 4: An itemset is not large in the original database and in the newly inserted transactions.

Cases 1 and 4 will not affect the final association rules. However, Case 2 may remove existing association rules, and Case 3 may add new association rules.

Cheung and his co-workers proposed an incremental mining algorithm, called FUP (Fast UPdate algorithm) [9][10], to efficiently cope with these four cases by pre-storing the previously mined large itemsets from the original database. It first calculates the large itemsets from the newly inserted transactions, and compares them with the pre-stored large itemsets. According to the comparison results, FUP can efficiently handle Cases 1, 2 and 4. It then re-processes the itemsets without sufficient information in Case 3 against the original database if necessary.

The performance of the FUP algorithm will get degraded if a lot of candidate itemsets from the newly inserted transactions belong to Case 3. As the result, the concept of *negative border* [19] was used to enlarge the amount of pre-stored mining patterns in incremental mining for further improving the maintenance performance at the expense of storage spaces [11][23]. A negative border is defined as follows.

Definition 1: Let R be a set of items, and L be a subset of the power set of R, which is closed with respect to the set inclusion relation. The *negative border* $NB(L)$ for L is a set that consists of the minimal itemsets $X \subseteq R$ and $X \notin L$.

Definition 2: A *negative itemset* for L is an itemset belonging to the negative border $NB(L)$.

3. The Extended Multidimensional Pattern Relation

The extended multidimensional pattern relation is conceptually similar to the construction of a data warehouse for OLAP [7][17][25]. Both of them systematically preprocess the underlying data in advance, integrate related information, and store the results in a centralized structural repository for later use and analysis. For providing ad-hoc, query-driven and online mining supports, the extended multidimensional pattern relation consists of two major types of information. One is the context information used to represent the contexts of each individual block of data which are gathered together from a specific business viewpoint. The other is the mining information used to record the available information mined from each individual block of data by a batch mining algorithm. Whenever a new block of data is inserted into the database, significant patterns are mined from this dataset as the mining information. The mining information along with the corresponding context information is then stored into the pre-defined extended multidimensional pattern relation. On the other hand, when an old block is deleted from the database, its corresponding context information and mining information are also removed from the extended multidimensional pattern relation. The definitions of the extended multidimensional pattern relation and its schema are given as follows.

Definition 3: An *extended multidimensional pattern relation schema EMPRS* with n_1 context attributes and n_2 content attributes can be represented as $EMPRS(ID, CX_1, CX_2, \ldots, CX_{n_1}, CN_1, CN_2, \ldots, CN_{n_2})$, where ID is an identification attribute, CX_i, $1 \le i \le n_1$, is a context attribute, and CN_i, $1 \le i \le n_2$, is a content attribute.

Definition 4: An *extended multidimensional pattern relation* including tuples $\{t_1, t_2, \ldots, t_m\}$ is an instance of the given $EMPRS(ID, CX_1, CX_2, \ldots, CX_{n_1}, CN_1, CN_2, \ldots, CN_{n_2})$. A tuple $t_i = (id_i, cx_{i1}, cx_{i2}, \ldots,$

cx_{in_1} , cn_{i1} , cn_{i2} , …, cn_{in_2}) in a extended multidimensional pattern relation indicates that for the block of data under the contexts of cx_{i1}, cx_{i2}, …, cx_{in_1}, the mining information contains cn_{i1}, cn_{i2}, …, cn_{in_2}.

The *frequent pattern se*t and the *negative pattern set* are two essential content attributes which are defined as follows.

Definition 5: A *frequent pattern set* (*fps*) for a block of data *D* is the set of all previously mined large itemsets with their supports for *D*. Assume the minimum support is *s* and the number of large itemsets discovered from *D* is *l*. A frequent pattern set can be represented as $fps = \{(x_i, s_i) \mid s_i \geq s$ and $1 \leq i \leq l\}$, where x_i is a large itemset and s_i is its support.

Definition 6: A *negative pattern set* (*nps*) for a block of data *D* is the set of previously mined negative itemsets with their supports from *NB* (*fps*) for *D*.

Therefore, whenever a new block of data is inserted, not only large itemsets but also negative itemsets are treated as the mining information and then stored into a pre-defined extended multidimensional pattern relation.

Example 1: Table 1 shows an extended multidimensional pattern relation with the initial minimum support set at 5%.

Table 1: An extended multidimensional pattern relation with minimum support = 5%

ID	*Region*	*Branch*	*Time*	*No_ Trans*	*No_ Patterns*	*Frequent_ Pattern_Set (Itemset, Support)*	*Negative_ Pattern_Set (Itemset, Support)*
1	CA	San Francisco	2003/10	10000	8	(A,10%),(B,11%),(C,9%),(AB,8%),(AC,7%),(BC,6%),(ABC,6%)	(D,2%)
2	CA	San Francisco	2003/11	15000	7	(A,5%),(B,7%),(C,5%)	(D,1%),(AB,2%),(AC,2%),(BC,1%)
3	CA	San Francisco	2003/12	12000	5	(A,5%),(C,9%)	(B,4%),(D,1%),(AC,4%)
4	CA	Los Angeles	2003/10	20000	5	(A,8%),(B,6%)	(C,2%),(D,3%),(AB,3%)
5	CA	Los Angeles	2003/11	25000	5	(A,5%),(C,6%)	(B,3%),(D,4%),(AC,3%)
6	CA	Los Angeles	2003/12	30000	7	(A,6%),(B,6%),(C,9%),(AB,6%)	(D,3%),(AC,4%),(BC,3%)
7	NY	New York	2003/10	18000	5	(B,8%),(C,7%),(BC,6%)	(A,2%),(D,2%)
8	NY	New York	2003/11	18500	5	(B,8%),(C,6%)	(A,4%),(D,2%),(BC,3%)
9	NY	New York	2003/12	19000	10	(A,5%),(B,9%),(C,8%),(D,6%),(BC,6%)	(AB,4%),(AC,4%),(AD,2%),(BD,4%)(CD,4%)

The tuple with *ID* = 1 shows that there are seven large itemsets with supports, {(A, 10%), (B, 11%), (C, 9%), (AB, 8%), (AC, 7%), (BC, 6%), (ABC, 6%)}, and one negative itemset with support, (D, 2%), are discovered from 10000 transactions and under the contexts of *Region* = **CA**, *Branch* = **San Francisco** and *Time* = **2003/10**. The other tuples have similar meaning.

4. Online Generation of Association Rules

The goal of online generation of association rules is to find the association rules satisfying the constraints in a mining request on line. The flexibility of mining requests allowed can increase through the usage of the proposed extended multidimensional pattern relation. Assume an extended multidimensional pattern relation based on an initial minimum support *s* includes tuples $\{t_1, t_2, \ldots, t_m\}$. Given a mining request *q* with a set of contexts cx_q, a new minimum support s_q ($s_q \geq s$), and a new minimum confidence $conf_q$, the online generation of association rules can select the tuples from the relation satisfying cx_q, integrate the mining information in these tuples, and then derive the association rules simultaneously satisfying s_q and $conf_q$. However, unlike the summarized information of *fact attributes* in a data warehouse, the mined patterns in the mining information can not be directly aggregated to satisfy users' mining requests. In this paper, an online mining approach called *Negative-Border Online Mining* (NOM) is proposed to achieve the mining task for online decision support. Let s_x denote the support of an itemset *x*, t_i denote the *i*-th tuple in an extended multidimensional pattern relation, $t_i.trans$ denote the number of transactions kept in t_i, $t_i.fps$ denote the *frequent pattern set* for t_i, $t_i.nps$ denote the *negative pattern set* for t_i, and $t_i.s_x$ denote the support of *x* in t_i. Also, a tuple in an extended multidimensional pattern relation is called a *matched tuple* (*mt*) if it satisfies the given context constraints. The following lemma can be derived (The detailed proofs are omitted here).

Lemma 1: If an itemset *x* satisfies a mining request *q*, there must exist at least a matched tuple *t*, such that $t.s_x$ satisfies s_q.

Lemma 2: If an itemset *x* satisfies a mining request *q*, it must belong to the candidate itemsets obtained by collecting the ones whose supports are larger than or equal to s_q in a matched tuple.

Lemma 3: If *x* is a candidate itemset, then $\forall x' \subset x$, *x'* is also a candidate itemset.

Definition 7: The *appearing count* $Count_x^{appearing}$ of a candidate itemset *x* is the sum of the counts of *x* appearing in the frequent pattern sets or negative pattern sets of matched tuples. Thus:

$$Count_x^{appearing} = \sum_{t_i \in mt \,\&\, x \in t_i.fps \cup t_i.nps} t_i.trans * t_i.s_x .$$

Definition 8: The not-appearing *upper-bound count* $Count_{\bar{x}}^{UB}$ of a candidate itemset *x* is the sum of the upper-bound counts of x not appearing in the frequent pattern sets and negative pattern sets of matched tuples. Thus:

$$Count_{\bar{x}}^{UB} = \sum_{t_i \in mt \,\&\, x \notin t_i.fps \cup t_i.nps} min(t_i.trans * s - 1, t_i.trans * \min_{\forall x' \subset x}(t_i.s_{x'})).$$

Definition 9: The *new upper-bound support* s_x^{UB} of a candidate itemset x is defined as follows:

$$s_x^{N\_UB} = \frac{Count_x^{appearing} + Count_{\bar{x}}^{UB}}{Match\_Trans},$$

where $Match\_Trans = \sum_{t_i \in mt} t_i.trans$ is the number of transactions in matched tuples.

Lemma 4: If x is a candidate itemset, then $s_x \le s_x^{UB}$.

Lemma 5: If x is a candidate itemset, then $\forall x' \subset x$, $s_{x'}^{UB} \ge s_x^{UB}$.

Lemma 6: If a candidate itemset x is contained in all the matched tuples, then $s_x^{UB} = s_x$.

5. Negative-Border Online Mining (NOM)

The NOM approach consists of the following three phases, *generation of candidate itemsets*, *reduction of candidate itemsets* and *generation of association rules*. The phase for *generation of candidate itemsets* first selects the tuples in the extended multidimensional pattern relation satisfying the context constraints in a mining request, and generates the candidate itemsets from the mining information of these matched tuples. After that, the phase for *reduction of candidate itemsets* calculates the upper-bound supports of the candidate itemsets and adopts two pruning strategies to reduce the number of candidate itemsets. Finally, the phase for *generation of association rules* re-processes the remaining candidate itemsets without sufficient information against the underlying database if necessary, and then derives the association rules from the found large itemsets. The proposed three-phased online mining approach is described below.

The Negative-Border Online Mining (NOM) approach:

INPUT: An extended multidimensional pattern relation based on an initial minimum support s and a mining request q with a set of contexts cx_q, a minimum support s_q ($s_q \ge s$) and a minimum confidence $conf_q$.

OUTPUT: A set of association rules satisfying the mining request q.

***Phase 1*: *Generation of candidate itemsets*:**

(a) Select the tuples satisfying cx_q from the extended multidimensional pattern relation.

(b) Collect the itemsets appearing in the matched tuples and satisfying s_q as the candidate itemsets for q.

(c) Calculate $Count_x^{appearing}$ and $Count_{\bar{x}}^{UB}$ for each candidate itemset x.

Phase 2*: *Reduction of candidate itemsets:

(a) Calculate $s_x^{N\_UB}$ for each candidate itemset x by the formula:

$$s_x^{N\_UB} = \frac{Count_x^{appearing} + Count_{\bar{x}}^{UB}}{Match\_Trans}.$$

(b) Discard the candidate itemset x and its proper supersets from the candidate set if $s_x^{N\_UB} \le s_q$.

(c) Put x into the set of final large itemsets if $s_x^{N\_UB} = \frac{Count_x^{appearing}}{Match\_Trans}$ and $s_x^{N\_UB} \ge s_q$.

Phase 3*: *Generation of association rules:

(a) Check whether each remaining candidate itemset x is large by scanning the underlying blocks of data for the matched tuples in which x does not appear.

(b) Generate the association rules satisfying $conf_q$ from the set of large itemsets.

Example 2: For the extended multidimensional pattern relation given in Table 1, assume a mining request q wants to get the patterns with the contexts cx_q of *Region* = **CA** and *Time* = **2003/10** and satisfying the minimum support s_q = 5.5%. According to Lemma 2, the set of candidate itemsets is {{A}, {B}, {C}, {AB}, {AC}, {BC}, {ABC}}, which is the union of the itemsets appearing in the frequent pattern sets and with their supports larger than 5.5%. Among these candidate itemsets, in Phase 2, the NOM approach can remove the candidate itemsets {C}, {AB}, {AC}, {BC} and {ABC}according to the negative itemset with support information of (C, 2%), and put the candidate itemsets {A} and {B} into the set of large itemsets for q. No candidate itemsets need to be processed in Phase 3.

6. Experimental Results

The experiments were implemented in Java on a workstation with dual XEON 2.8GHz processors and 2048MB main memory, running RedHat 9.0 operation system. The datasets were generated by a generator similar to that used in [4]. The generator first generated L maximal potentially large itemsets, each with an average size of I items. The items in a potentially large itemset were randomly chosen from the total N items according to its actual size. The generator then generated D transactions, each with an average size of T items. The items in a transaction were generated according to the *L* maximal potentially large itemsets in a probabilistic way.

A group of datasets generated in the above way and used in our experiments are listed in Table 2, where the datasets in the group had the same *D*, *T*, *L* and *I* values but different *N* value. Each dataset was treated as a block of data in the database.

Table 2: The group of datasets generated for the experiments

Size	*Datasets*	*D*	*T*	*I*	*L*	*N*
10	*T20I8D100KN*^{1} *to* *T20I8D100KN*^{10}	100000	20	8	400	200 to 290

The NOM and the Apriori algorithms were run for the group along with different minimum supports ranging from 0.022 to 0.04 in the mining requests. The execution times spent by the two algorithms for each group are respectively shown in Figure 1. It is easily seen that the execution time by the NOM algorithm was always much less than that by the Apriori algorithm.

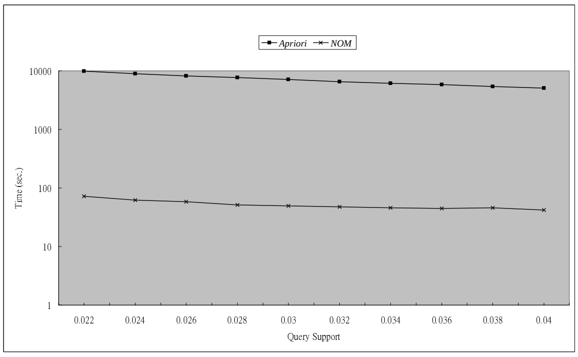

Figure 1: The execution time spent by the NOM and the Apriori algorithms

Next, we compare the NOM algorithm with our previous multidimensional on-line mining algorithm [24]. The execution times spent by the two algorithms for the group are shown in Figure 2. It is also easily seen that the execution time by the NOM algorithm was much less than that by the multidimensional on-line mining algorithm.

7. Conclusion

In this paper, the concept of *negative border* has been used to enlarge the mining information in the multidimensional pattern relation [24] to help get tight upper-bound supports of candidate itemsets, and thus reduce the number of candidate itemsets to be considered. Based on the extended multidimensional pattern relation, a corresponding online mining approach called *Negative-Border Online Mining* (NOM) has been proposed to efficiently and effectively utilize the information of *negative itemset* in the negative border. The experimental results have been shown that the proposed NOM approach is more efficient than the well-known Apriori approach and better than the multidimensional on-line mining algorithm.

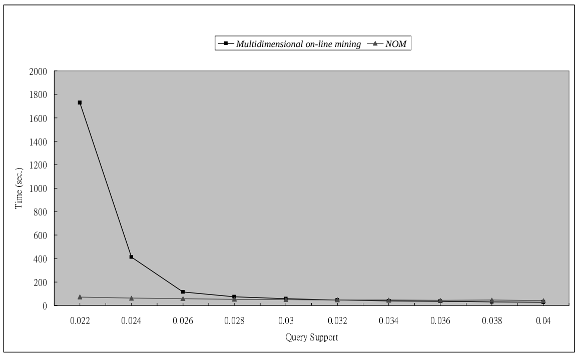

Figure 2: The execution time spent by the NOM algorithm and the multidimensional on-line mining algorithms

Acknowledgement

This research was supported by the National Science Council of the Republic of China under Grand No. NSC93-2752-E-009-006-PAE.

References

1. C. C. Aggarwal and P. S. Yu, "A New Approach to Online Generation of Association Rules," IEEE Transactions on Knowledge and Data Engineering, Vol. 13, No. 4, pp. 527-540, 2001.
2. R. Agrawal, T. Imielinksi and A. Swami, "Mining Association Rules between Sets of Items in Large Database,“ ACM SIGMOD Conference, pp. 207-216, Washington DC, USA, 1993.
3. R. Agrawal, T. Imielinksi and A. Swami, "Database Mining: A Performance Perspective," IEEE Transactions on Knowledge and Data Engineering, Vol. 5, No. 6, pp. 914-925, 1993.
4. R. Agrawal and R. Srikant, "Fast Algorithm for Mining Association Rules," ACM International Conference on Very Large Data Bases, pp. 487-499, 1994.
5. R. Agrawal and R. Srikant, "Mining Sequential Patterns," IEEE International Conference on Data Engineering, pp. 3-14, 1995.
6. S. Brin, R. Motwani, J. D. Ullman and S. Tsur, "Dynamic Itemset Counting and Implication Rules for Market Basket Data," ACM SIGMOD Conference, pp. 255-264, Tucson, Arizona, USA, 1997.
7. S. Chaudhuri and U. Dayal, "An Overview of Data

Warehousing and OLAP Technology," ACM SIGMOD Record, 26:65-74, 1997.

8. M. S. Chen, J. Han and P. S. Yu, "Data mining: An Overview from A Database Perspective," IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, pp. 866-883, 1996.
9. D. W. Cheung, J. Han, V. T. Ng and C. Y. Wong, "Maintenance of Discovered Association Rules in Large Databases: An Incremental Updating Approach," IEEE International Conference on Data Engineering, pp. 106-114, 1996.
10. D. W. Cheung, S. D. Lee, and B. Kao, "A General Incremental Technique for Maintaining Discovered Association Rules," In Proceedings of Database Systems for Advanced Applications, pp. 185-194, Melbourne, Australia, 1997.
11. R. Feldman, Y. Aumann, A. Amir, and H. Mannila, "Efficient Algorithms for Discovering Frequent Sets in Incremental Databases," ACM SIGMOD Workshop on DMKD, pp. 59-66, USA, 1997.
12. G. Grahne, L. V. S. Lakshmanan, X. Wang and M. H. Xie, "On Dual Mining: From Patterns to Circumstances, and Back," IEEE International Conference on Data Engineering, pp. 195-204, 2001.
13. J. Han, L. V. S. Lakshmanan and R. Ng, " Constraint-based, Multidimensional Data Mining," IEEE Computer Magazine, pp.2-6, 1999.
14. J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2001.
15. C. Hidber, "Online Association Rule Mining," ACM SIGMOD Conference, pp. 145-156, USA, 1999.
16. T. P. Hong, C. Y. Wang and Y. H. Tao, "A New Incremental Data Mining Algorithm Using Pre-large Itemsets," International Journal on Intelligent Data Analysis, 2001.
17. W. H. Immon, Building the Data Warehouse, Wiley Computer Publishing, 1996.
18. H. Mannila, H. Toivonen and A.I. Verkamo, "Efficient Algorithm for Discovering Association Rules," The AAAI Workshop on Knowledge Discovery in Databases, pp. 181-192, 1994.
19. H. Mannila and H. Toivonen, "On an Algorithm for Finding all Interesting Sentences," The European Meeting on Cybernetics and Systems Research, Vol. II, 1996.
20. J. S. Park, M. S. Chen and P. S. Yu, "Using a Hash-based Method with Transaction Trimming for Mining Association Rules," IEEE Transactions on Knowledge and Data Engineering, Vol. 9, No. 5, pp. 812-825, 1997.
21. N. L. Sarda and N. V. Srinivas, "An Adaptive Algorithm for Incremental Mining of Association Rules," IEEE International Workshop on Database and Expert Systems, pp. 240-245, 1998.
22. A. Savasere, E. Omiecinski and S. Navathe, "An Efficient Algorithm for Mining Association Rules in Large Database," ACM International Conference on Very Large Data Bases, pp. 432-444, 1995.
23. S. Thomas, S. Bodagala, K. Alsabti and S. Ranka "An Efficient Algorithm for the Incremental Update of Association Rules in Large Databases," The International Conference on Knowledge Discovery and Data Mining, pp. 263-266, 1997.
24. C. Y. Wang, T. P. Hong and S. S. Tseng, "Multidimensional On-line Mining," IEEE ICDM Foundation of Data Mining Workshop, pp. 196-202, 2003.
25. J. Widom, "Research Problems in Data Warehousing," ACM International Conference on Information and Knowledge Management, 1995.