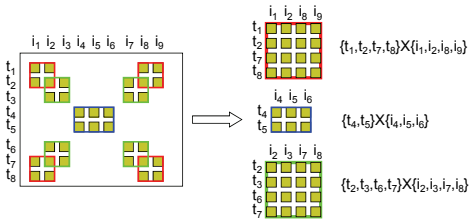
Succinct Summarization of Transactional Databases: An Overlapped Hyperrectangle Scheme

Yang Xiang, Ruoming Jin, David Fuhry, Feodor F. Dragan
Kent State University

Presented by: Yang Xiang

Introduction

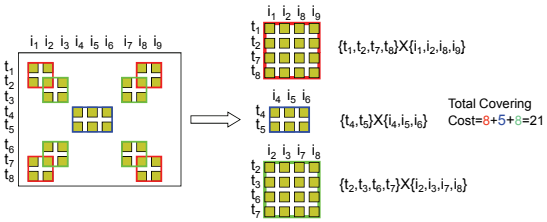
- How to succinctly describe a transactional database?



- Summarization ⇔ a set of hyperrectangles (Cartesian products) with minimal cost to cover ALL the cells (transaction-item pairs) of the transactional database

Problem Formulation

- Example: cost ($\{t_1,t_2,t_7,t_8\}X\{i_1,i_2,i_8,i_9\}$)
- For a hyperrectangle, $T_i X I_i$, we define its cost to be $|T_i|+|I_i|$



Related Work

- Data Descriptive Mining and Rectangle Covering [Agrawal94] [Lakshmanan02] [Gao06]
- Summarization for categorical databases [Wang06] [Chandola07]
- Data Categorization and Comparison [Siebes06] [Leeuwen06] [Vreeken07]
- Others: Co-clustering [Li05], Approximate Frequent Itemset Mining [Afrati04] [Pei04] [Steinbach04], Data Compression [Johnson04]…

Hardness Results

- Unfortunately, this problem and several variations are proved to be NP-Hard!
  (Proof hint: Reduce minimum set cover problem to this problem.)

Weighted Set Cover Problem

- The summarization problem is closely related to the weighted set covering problem
  - Ground set → All cells of the database
  - Candidate sets (each set has a weight)
    → All possible hyperrectangles (each hyperrectangle has a cost)
  - Weighted set cover problem:
    Use a subset of candidate sets to cover the ground set such that the total weight is minimum
    → Use a subset of all possible hyperrectangles to cover the database such that the total cost is minimum

Naïve Greedy Algorithm

- Greedy algorithm:
  - Each time choose a hyperrectangle ($H_i = T_i \times I_i$) with lowest price $\gamma(H_i) = \frac{|T_i| + |I_i|}{|T_i \times I_i \setminus R|}$.
  - $|T_i|+|I_i|$ is hyperrectangle cost. R is the set of covered cells
  - Approximation ratio is $ln(k)+1$ [V.Chvátal 1979]. k is the number of selected hyperrectangles.
- The problem?
  - The number of candidate hyperrectangles are $2^{|T|+|I|}$!!!

Basic Idea-1

- Restricted number of candidates
  - A candidate is a hyperrectangle whose itemset is either frequent, or a single item. $C_\alpha = \{T_i \times I_i \mid I_i \in F_\alpha \cup I_s\}$ is the set of candidates.
- Given an itemset either frequent or singleton, it corresponds to an exponential number of hyperrectangles. For example: $\{1,2,3\} \times \{a\}$

  It *corresponds to the following hyperrectangles:* $\{1\} \times \{a\}$, $\{2\} \times \{a\}$, $\{3\} \times \{a\}$, $\{1,2\} \times \{a\}$, $\{1,3\} \times \{a\}$, $\{2,3\} \times \{a\}$, $\{1,2,3\} \times \{a\}$
- *The number of hyperrectangle is still exponential*

$$|C_\alpha| = \sum_{I_i \in F_\alpha \cup I_s} 2^{|T(I_i)|}$$

transaction set where I_i appear

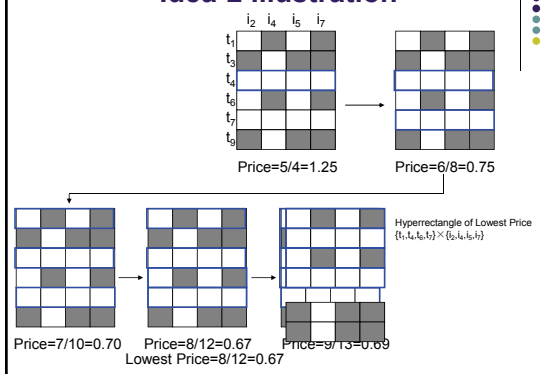
set of all singleton items

frequent itemsets with support α

Basic Idea-2

- Given an itemset, we do NOT try to enumerate the exponential number of hyperrectangles sharing this itemset.
- A linear algorithm to find the hyperrectangle with the lowest price among all the hyperrectanges sharing the same itemset.

Idea-2 Illustration



HYPER Algorithm

While there are some cells not covered

[STEP1] **Calculate** lowest price hyperrectangle for each frequent or single itemset. (basic idea-2)

[STEP2] **Find** the frequent or single itemset whose corresponding lowest price hyperrectangle is the lowest among all.

[STEP3] **Output** this hyperrectangle.

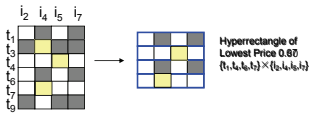
[STEP4] **Update** Coverage of the database.

HYPER

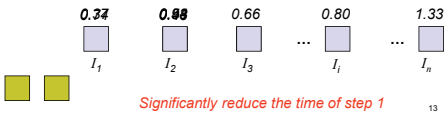
- *We assume Apriori algorithm provides F_α .*
- *HYPER* is able to find the best cover which utilizes the exponential number of hyperrectangles, described by candidate sets $C_\alpha = \{T_i \times I_i \mid I_i \in F_\alpha \cup I_s\}$ ($|C_\alpha| = \sum_{I_i \in F_\alpha \cup I_s} 2^{|T(I_i)|}$).
- Properties:
  - Approximation ratio is still $ln(k)+1$ *w.r.t.* C_α,
  - Running time is $O(|T|(|I_s| + \log|T|)(|F_\alpha| + |I_s|)k)$ polynomial to $F_\alpha \cup I_s$.

Pruning Technique for HYPER

- One important observation: For each frequent or single itemset, the price of lowest price hyperrectangle will only increase!



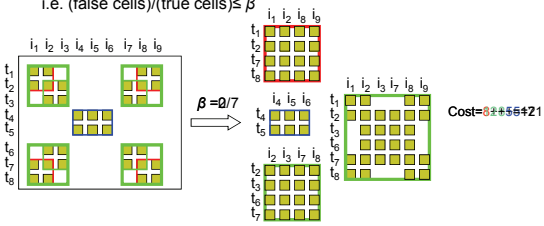
$I_i \in F_d \cup I_s$



Significantly reduce the time of step 1

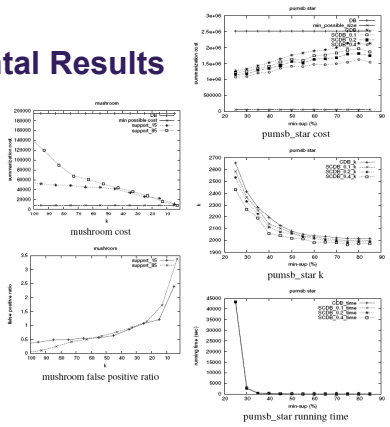
Further Summarization: HYPER+

- The number of hyperrectangles returned by *HYPER* may be too large or the cost is too high.
- We can do further summarization by allowing false positive budget β, i.e. (false cells)/(true cells)≤ β



Experimental Results

- Two important observations:
  - Convergence behavior
  - Threshold behavior
- Two important conclusions:
  - min-sup doesn't need to be too low.
  - We can reduce *k* to a relatively small number without increasing false positive ratio too much.



Conclusion and Future Work

- Conclusion
  - *HYPER* can utilize exponential number of candidates to achieve a *ln(k)+1* approximate bound but works in polynomial time.
  - We can speed up *HYPER* by pruning technique.
  - *HYPER* and *HYPER+* works effectively and we find threshold behavior and convergence behavior in the experiments.

Discussion

- How to choose of \alpha, the support threshold?
- Overlap hypercube sampling vs co-clustering
- Frequent set on transactions?
- High false positive rates, small k
- Other real applications?

Discussion for Suggestion Sampling paper

- Applicability of the work? Currently, we see the application providers are the ones who owns the logs.
- May not appeal to statisticians.
- How to compute the trial distribution, p? Of course, not feasible to query nodes of the tree. What if the p is too far from \phi?
- Tuning values for power law distribution, no guidelines
- Comparison with Wikipedia?

Thank you!

Questions?

References

[Agrawal94] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In SIGMOD Conference, pp 94-105, 1998.

[Lakshmanan02] Laks V. S. Lakshmanan, Raymond T. Ng, Christine Xing Wang, Xiaodong Zhou, and Theodore J. Johnson. The generalized mdl approach for summarization. In *VLDB '02*, pp 766–777, 2002.

[Gao06] Byron J. Gao and Martin Ester. Turning clusters into patterns: Rectangle-based discriminative data description. In *ICDM*, pages 200–211, 2006.

[Wang06] Jianyong Wang and George Karypis. On efficiently summarizing categorical databases. *Knowl. Inf. Syst.*, 9(1):19–37, 2006.

[Chandola07] Varun Chandola and Vipin Kumar. Summarization -compressing data into an informative representation. *Knowl. Inf. Syst.*, 12(3):355–378, 2007.

[Siebes06] Arno Siebes, Jilles Vreeken, and Matthijs van Leeuwen. Itemsets that compress. In *SDM*, 2006.

[Leeuwen06] Matthijs van Leeuwen, Jilles Vreeken, and Arno Siebes. Compression picks item sets that matter. In *PKDD*, pp 585–592, 2006.

[Vreeken07] Jilles Vreeken, Matthijs van Leeuwen, and Arno Siebes. Characterising the difference. In *KDD '07*, pages 765–774, 2007.

[Li05] Tao Li. A general model for clustering binary data. In *KDD*, pp 188–197, 2005.

[Afrati04] Foto N. Afrati, Aristides Gionis, and Heikki Mannila. Approximating a collection of frequent sets. In *KDD*, pp 12–19, 2004.

[Pei04] Jian Pei, Guozhu Dong, Wei Zou, and Jiawei Han. Mining condensed frequent-pattern bases. *Knowl. Inf. Syst.*, 6(5):570–594, 2004.

[Steinbach04] Michael Steinbach, Pang-Ning Tan, and Vipin Kumar. Support envelopes: a technique for exploring the structure of association patterns. In *KDD '04*, pages 296–305, New York, NY, USA, 2004. ACM.

[Johnson04] David Johnson, Shankar Krishnan, Jatin Chhugani, Subodh Kumar, and Suresh Venkatasubramanian. Compressing large boolean matrices using reordering techniques. In *VLDB 2004*, pages 13–23. VLDB Endowment, 2004.

[V.Chvátal 1979] V. Chvátal. A greedy heuristic for the set-covering problem. *Math. Oper. Res*, 4:233–235, 1979.