Mining Strong Affinity Association Patterns in Data Sets with Skewed Support Distributions

Hui Xiong

Department of Computer Science & Engineering
University of Minnesota - Twin Cities

Overview

⇒ Introduction

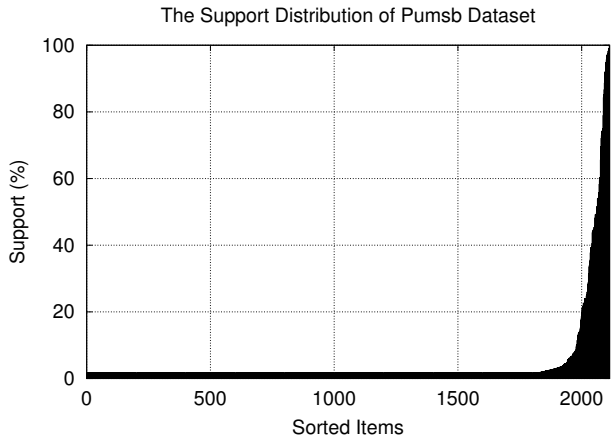
- ◇ General Problems
- ◇ Research Motivations
- ◇ Related Works

- Hyperclique Patterns
- Hyperclique Miner Algorithm
- Experimental Evaluation
- Conclusions and Future Work

General Problems: Cross-support Patterns

Definition 1 *Cross-support patterns are patterns which involve items with substantially different support levels.*

- Cross-support patterns tend to be poorly correlated and most of them are spurious patterns.
  - For instance: {TV, milk}, {bread, gold necklaces, earrings}
- in real world, many data sets have inherently skewed support distributions.



- Pumsb - a census data set from IBM (http://www.almaden.ibm.com/software).

Two major problems with frequent pattern mining framework

- If the minimum support threshold is low, a huge number of cross-support patterns can be generated.
  - Too many patterns (Pattern Jungle) and high computation cost, especially when data sets have skewed support distributions.

- If the minimum support threshold is high, many strong affinity patterns occurring at low levels of support cannot be identified.
  - miss interesting associations among rare but expensive items e.g.{earrings, gold ring, bracelet}, {TV, DVD players}.
  - miss interesting associations among rare anomalous events.

Research Motivations

- Ability to detect strong affinity patterns at low levels of support
  - LA1 Data Set

Hyperclique patterns	support
{najibullah, kabul, afghan}	0.2%
{arafat, yasser, PLO, Palestine}	0.4%
{amal, militia, hezbollah, syrian, beirut}	0.1%

  - Retail Data Set

Hyperclique patterns	support
{earrings, gold ring, bracelet}	0.019%
{coffee maker, can opener, toaster}	0.014%
{baby bumper pad, diaper stacker, baby crib sheet}	0.028%
{jar cookie, canisters 3pc, box bread, soup tureen, goblets 8pc}	0.012%

- Ability to remove cross-support patterns

Related Work

- Closed/Maximal Pattern Mining
  - Reduce the number of patterns generated.
  - Limitations
    - Do not remove cross-support patterns.
    - Algorithms may still break down at low levels of support, especially for data sets with skewed support distributions
- Constraint Pattern Mining
  - Finding interesting associations without support pruning. By Cohen et al. [ICDE'01 & TKDE]
  - Statistical χ^2 test to discover dependent patterns by Brin et al. [SIGMOD'97]
  - All-Confidence Measure by Omiecinski [TKDE'03]
    - The all-confidence measure for an itemset $P = \{i_1, i_2, \cdots, i_m\}$ is defined as

$$allconf(P) = \min[\{conf(A \rightarrow B | \forall A, B \subset P, A \cup B = P, A \cap B = \emptyset\}].$$

    - All-confidence measure has the anti-monotone property.

The h-confidence measure

Definition 2 *The **h-confidence** of an itemset* $P = \{i_1, i_2, \cdots, i_m\}$ *is defined as* $hconf(P) = \min[conf\{i_1 \rightarrow i_2, \ldots, i_m\}, conf\{i_2 \rightarrow i_1, i_3, \ldots, i_m\}, \ldots, conf\{i_m \rightarrow i_1, \ldots, i_{m-1}\}]$, *where* $conf$ *follows from the definition of association rule confidence.*

Lemma 1 *For an itemset* $P = \{i_1, i_2, \cdots, i_m\}$, $hconf(P)$ *is mathematically equivalent to* $allconf(P)$.

- $hconf(P) = allconf(P) = \frac{supp(\{i_1,i_2,\ldots,i_m\})}{max_{1 \leq k \leq m}\{supp(\{i_k\})\}}$

The h-confidence measure has the anti-monotone property. In other words, if $P \subseteq P'$, then $hconf(P) \geq hconf(P')$.

- For an itemset $P = \{A, B, C\}$, assume that:
  - $supp(\{A\}) = 0.1$, $supp(\{\mathsf{B}\}) = 0.1$, $supp(\{C\}) = 0.06$, $supp(\{A, B, C\}) = 0.06$.
  - $conf\{A \rightarrow B, C\} = supp(\{A, B, C\})/supp(\{A\}) = 0.6$.
  - $conf\{B \rightarrow A, C\} = 0.6; conf\{C \rightarrow A, B\} = 1$.
  - $hconf(P) = min\{conf\{B \rightarrow A, C\}, conf\{A \rightarrow B, C\}, conf\{C \rightarrow A, B\}\}$= 0.6.

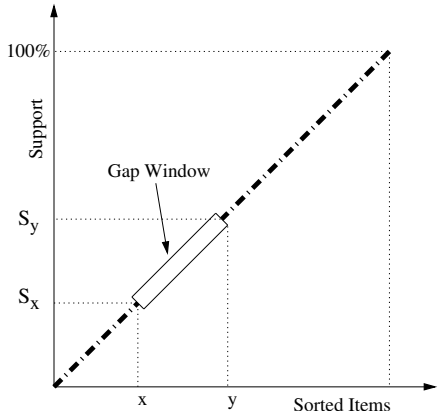
Overview

- Introduction
- ⇒ Hyperclique Patterns
  - ◇ Hyperclique Pattern Concepts
  - ◇ Properties of the H-confidence Measure
- Hyperclique Miner Algorithm
- Experimental Evaluation
- Future Work

Hyperclique Pattern Concepts

Definition 3 *Given a set of items $I = \{I_1, I_2, \ldots, I_n\}$, an itemset $P \subseteq I$ is a* **hyperclique pattern** *if and only if $|P| > 0$ and $hconf(P) \geq h_c$, where h_c is the minimum h-confidence threshold.*

- The cross-support property of the h-confidence measure.



Lemma 2 *Given: 1) Two items x and y with $supp(x) < supp(y)$, 2) Two item sets, $L(x) = \{x' | supp(\{x'\}) \leq supp(\{x\})\}$ and $U(y) = \{y' | supp(\{y'\}) \geq supp(\{y\})\}$, for any cross-support pattern $P = \{x_{i_1}, x_{i_2}, \ldots, x_{i_k}, y_{j_1}, y_{j_2}, \ldots, y_{j_l}\}$ from these two itemsets has an upper bound of the h-confidence given by $\frac{max_{1 \leq p \leq m}\{supp(\{x_p\})\}}{min_{1 \leq q \leq n}\{supp(\{y_q\})\}}$.*

The Cross-Support Property

- Consider the following set of items:

Item	Support
1	0.9
2	0.9
3	0.3
4	0.2
5	0.1

  - Let $L = \{3, 4, 5\}$ and $U = \{1, 2\}$
  - If P is a cross-support pattern, then $hconf(P) \leq \frac{max\{0.1,\ 0.2,\ 0.3\}}{min\{0.9,\ 0.9\}} = 1/3$
  - If $h_c > 1/3$, no cross-support pattern will be extracted.

The Cross-Support Property

Theorem 1 *Given:*

1) A measure of association, f

2) Two items x and y with $supp(\{x\}) < supp(\{y\})$;

3) Two item sets $L(x) = \{x' | supp(\{x'\}) \leq supp(\{x\})\}$ and $U(y) = \{y' | supp(\{y'\}) \geq supp(\{y\})\}$;

If the following conditions hold,

1) There exists a non-trivial upper bound function, $upper(f)$, for the measure f;

2) $upper(f(\{x, y\}))$ can be computed by only $supp(\{x\})$ and $supp(\{y\})$;

3) If x is fixed, $upper(f(\{x, y\}))$ decreases monotonically with increasing $supp(\{y\})$;

4) If y is fixed, $upper(f(\{x, y\}))$ decreases monotonically with decreasing $supp(\{x\})$;

5) If the measure f is applied to patterns with three or more items, then f must have an anti-monotone property.

Then $f(p) \leq upper(f(\{x, y\}))$ if p is a cross-support pattern.

The Cross-Support Property

- Examples of measures of association which have the $cross\text{-}support$ property (Assume that $supp(\{x\}) < supp(\{y\})$).

Measure	Computation Formula	Upper Bound
Cosine	$\frac{supp(\{x,y\})}{\sqrt{supp(\{x\})supp(\{y\})}}$	$\sqrt{\frac{supp(\{x\})}{supp(\{y\})}}$
Jaccard	$\frac{supp(\{x,y\})}{supp(\{x\})+supp(\{y\})-supp(\{x,y\})}$	$\frac{supp(\{x\})}{supp(\{y\})}$
Interest	$\frac{supp(\{x,y\})}{supp(\{x\})supp(\{y\})}$	$\frac{1}{supp(\{y\})}$

The High-affinity Property

Given a pair of items $P = \{i_1, i_2\}$, the cosine measure for P can be computed as $\frac{supp(\{i_1,i_2\})}{\sqrt{supp(\{i_1\}) \times supp(\{i_2\})}}$, while the Jaccard measure for P is $\frac{supp(\{i_1,i_2\})}{supp(\{i_1\})+supp(\{i_2\})-supp(\{i_1,i_2\})}$.

Lemma 3 *If an item set $P = \{i_1, i_2\}$ is a size-2 hyperclique pattern, then we have $cosine(P) \geq h_c$.*

Lemma 4 *If an item set $P = \{i_1, i_2\}$ is a size-2 hyperclique pattern, then we have $jaccard(P) \geq h_c/2$.*

- If $P = \{A, B\}$ is a hyperclique pattern with the minimum h-confidence threshold h_c, then
  - $hconf(P) \geq h_c$, $cosine(P) \geq h_c$, $Jaccard(P) \geq h_c/2$
- For any hyperclique pattern $P = \{i_1, i_2, \ldots, i_k\}$ $(k > 2)$ at the h-confidence threshold h_c, we have
  - $hconf(P) \geq h_c$
  - $cosine(Q) \geq h_c$ and $Jaccard(Q) \geq h_c/2$, where $Q = \{i_l, i_m\}$ and $Q \subset P$.

Overview

- Introduction
- Hyperclique Patterns

⇒ Hyperclique Miner Algorithm

- Experimental Evaluation
- Conclusions and Future Work

Hyperclique Miner Algorithm

- Hyperclique Miner is an Apriori-type algorithm.
  - Method:

    1) Get size-1 prevalent items
    2) Partitioning items into different levels of support

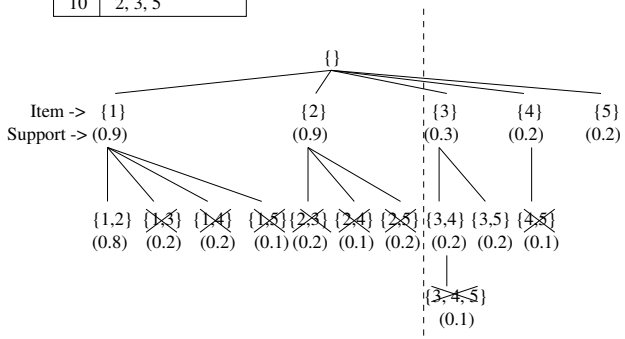
    # In each item partition
    3) for size of itemsets in $(2, 3, \ldots, k-1)$ do
    4) Generate candidate hyperclique patterns.
    5) Prune based on the support of candidate hyperclique patterns.
    6) Prune based on the h-confidence of candidate hyperclique patterns
    7) Generate hyperclique patterns.

Hyperclique Miner Algorithm - An Illustration

- Pruning by the cross-support property (Let h_c=0.55 and support threshold = 0).
- Pruning by the anti-monotone property.

TID	Items
1	1, 2
2	1, 2
3	1, 3, 4
4	1, 2
5	1, 2
6	1, 2
7	1, 2, 3, 4, 5
8	1, 2
9	1, 2
10	2, 3, 5

Item	Support
1	0.9
2	0.9
3	0.3
4	0.2
5	0.2



- Let $L = \{3, 4, 5\}$ and $U = \{1, 2\}$. If P is a cross-support pattern, then $hconf(P) \leq \frac{max\{0.1,\ 0.2,\ 0.3\}}{min\{0.9,\ 0.9\}} = 1/3$

Overview

- Introduction
- Hyperclique Patterns
- Hyperclique Miner Algorithm

⇒ Experimental Evaluation

  - ◇ Experimental Setup
  - ◇ The Pruning Effect of Hyperclique Miner
  - ◇ Hyperclique Patterns - High Affinity Patterns
  - ◇ Hyperclique-based Clustering via Hypergraph Partition

- Conclusions and Future Work

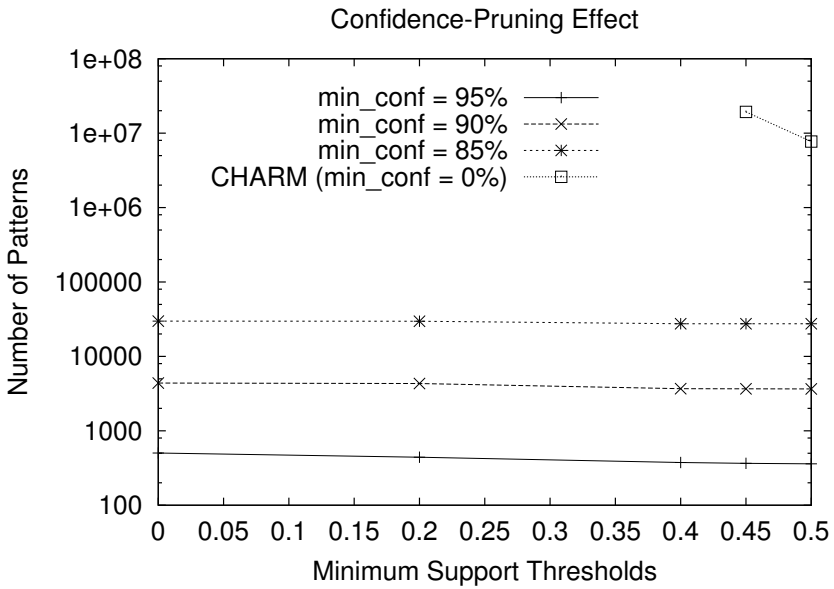
Experimental Setup

- Experimental Data Sets
  - Real Data Sets

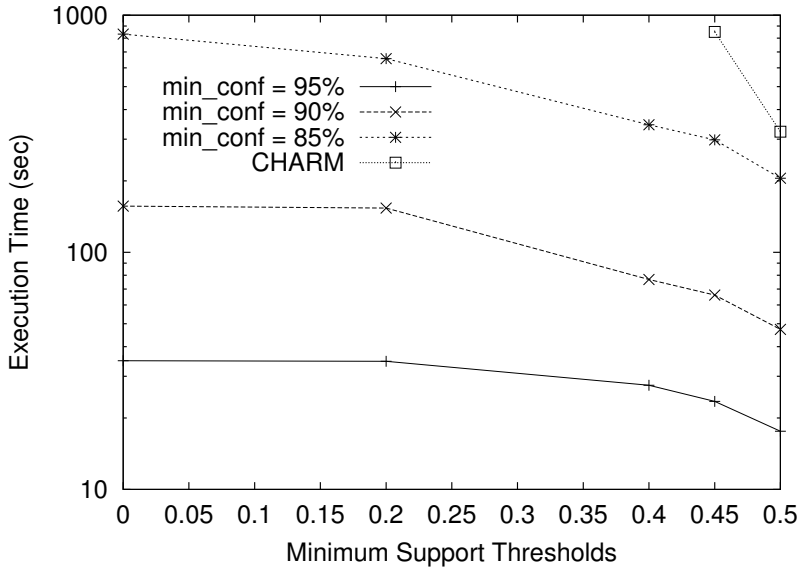
Data set	#Item	#Record	Avg. Length	Source
Pumsb	2113	49046	74	IBM Almaden
S&P 500	932	716	75	Stock Market
LA1	29704	3204	145	TREC-5
Retail	14462	57671	129	Retail Store

- Experimental Platform
  - Sun Ultra 10 Work Station with a 440 MHz CPU and 128 Mbytes of memory running the SunOS 5.7 operating system.
- CHARM as the base line to show the relative performance.
  - CHARM has better performance than Apriori and MAFIA (for maximal frequent patterns) at low levels of support.

The Pruning Effect on Pumsb data set



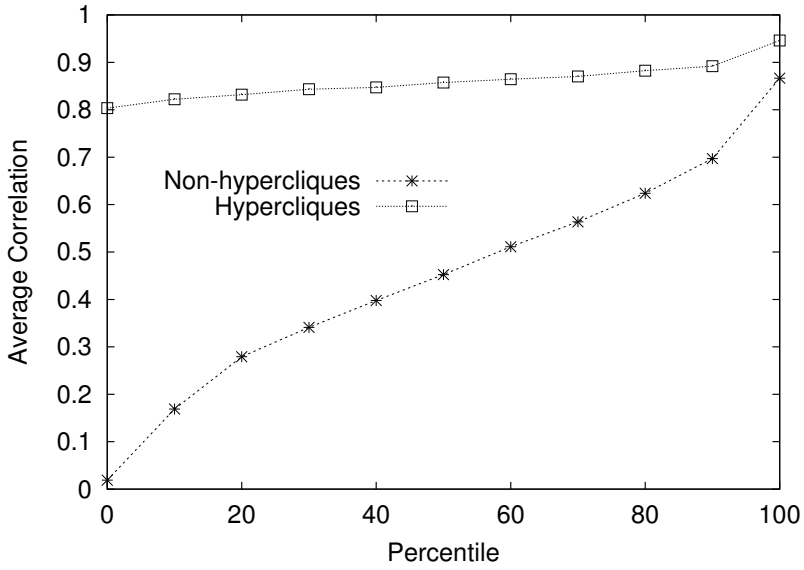
(a)



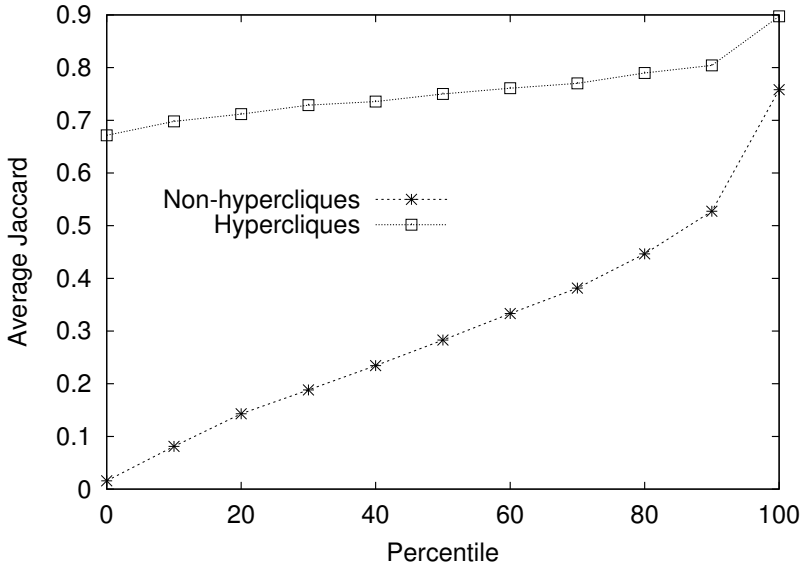
(b)

- CHARM has difficulties in identifying patterns when $minsupp \leq 40\%$.
- For instance, hyperclique miner finds one long pattern containing 9 items with the support 0.23 and h-confidence 94.2%.

Hyperclique Patterns - High Affinity Patterns



(c)



(d)

- $minsupp = 0.0005$ and $h_c = 80\%$ on Retail data set.
- Hyperclique patterns have extremely high average pair wise correlation compared to the non-hyperclique patterns.

Hyperclique Based Clustering via Hypergraph Partition

- Frequent pattern based clustering via hypergraph partition [Han et.al.98]
  - Limitation 1: The frequent pattern is not a good representative to capture the overall affinity among items.
  - Limitation 2: To cover more items, we have to set a low support threshold and get many frequent patterns.

	Vertices	Hyperedges	Partitions	#Clean Clusters
Frequent Pattern	440	19,602	40	16
Hyperclique Pattern	861	11,207	80	41

- Frequent patterns: $minsupp = 3\%$
- Hyperclique patterns: $minsupp = 0\%$ and $h_c = 20\%$.
- Noise Smoothing
  - When $minsupp = 0\%$ and $h_c = 20\%$, the discovered hyperclique patterns cover 861 items. 71 items have been eliminated by hyperclique miner.
  - 68 out of 71 items are assigned to wrong clusters.

Hyperclique Based Clustering via Hypergraph Partition

- Six clusters at low levels of support (around 1%) from S&P 500 Stock Data Set.

No	Discovered Clusters	Industry Group
1	Baltimore Gas↓, CINergy Corp↓, Amer Electric Power↓, Duke Power↓, Consolidated Edi↓, Entergy Corp↓, Genl Public Util↓, Houston Indus↓, PECO Energy↓, Texas Utilities↓	Power
2	Becton Dickinso↓, Emerson Electric↓, Amer Home Product↓, Johnson & Johnson↓, Merck Co↓, Pfizer Inc↓, Schering-Plough↓, Warner-Lambert↓	health product
3	Bank of New York↓, Bank of Boston↓, CoreStates Financial↓, CIGNA Corp↓, Comerica Inc↓, Aetna Life & Cas↓, Amer General↓, Fleet financial↓, Morgan (J.P.↓), KeyCorp↓, Mellon Bank Corp↓, NationsBank Corp↓, Natl City Corp↓, Wells Fargo↓, BankAmerica Corp↓	Financial
4	Bell Atlantic Co↑, BellSouth Corp↑, CPC Intl↑, GTE Corp↑, Ameritech Corp↑, NYNEX Corp↑, Pacific Telesis↑, SBC Communication↑, US West Communication↑	Comm.
5	duPont (EI) deNemo↑, Goodrich (B.F.)↑, Nalco Chemical↑, Rohm & Haas↑, Avon Products↑	chemical
6	Federated Dept↑, Gap Inc↑, Nordstrom Inc↑, Pep Boys-Man↑, Sears↑, TJX companies↑, Walmart↑	Retail

Overview

- Introduction
- Hyperclique Patterns
- Hyperclique Miner Algorithm
- Experimental Evaluation

$\Rightarrow$ Conclusions and Future Work

Conclusions and Future Work

- Conclusions
  - Introduce hyperclique patterns
  - Present the h-confidence measure
    - The Anti-monotone Property
    - The Cross-support Property
    - The High-affinity Property
  - Design hyperclique miner algorithm.
  - Conduct experiments to show the performance of hyperclique miner and the application of hyperclique patterns for clustering via hypergraph partition.
- Future Works
  - Extending the Notion of Hyperclique Patterns.
  - Hyperclique Pattern based Clustering.
  - Comprehensive Understanding of Hyperclique Patterns.

Thank You!

- Personal Homepage - http://www.cs.umn.edu/∼huix