

The background of the slide is a complex, abstract composition. It features a central white banner with a subtle geometric pattern of thin lines and small plus signs. This banner is flanked by two large, overlapping triangular shapes in a light gray color. The entire scene is set against a backdrop of a dark, reddish-brown color with a dense network of thin, light-colored lines forming a complex geometric pattern. Scattered throughout this background are numerous small, colorful dots in shades of green, blue, and yellow. In the top-left corner, there is a horizontal strip containing a series of small, stylized symbols and characters in a light purple color. In the bottom-left corner, there is a rectangular inset showing a cluster of small, orange and red dots on a light background, with a small grid of pink squares overlaid on it.

Pattern Evaluation

Pattern Evaluation

- ❑ Limitation of the Support-Confidence Framework
- ❑ Interestingness Measures: Lift and χ^2
- ❑ Null-Invariant Measures
- ❑ Comparison of Interestingness Measures

The background of the slide is a complex, abstract composition. It features a dark, reddish-brown base with a network of thin, light-colored lines forming a mesh or web-like structure. Scattered throughout are numerous small, colored dots in shades of green, blue, and orange. In the upper left, there is a horizontal band with a repeating pattern of small, stylized symbols. On the left side, there is a vertical strip containing a grid of small, colored squares. The overall aesthetic is technical and data-driven.

Limitation of the Support- Confidence Framework

How to Judge if a Rule/Pattern Is Interesting?

- ❑ Pattern-mining will generate a large set of patterns/rules
 - ❑ Not all the generated patterns/rules are interesting
- ❑ Interestingness measures: Objective vs. subjective
 - ❑ Objective interestingness measures
 - ❑ Support, confidence, correlation, ...
 - ❑ Subjective interestingness measures:
 - ❑ Different users may judge interestingness differently
 - ❑ Let a user specify
 - ❑ Query-based: Relevant to a user's particular request
 - ❑ Judge against one's knowledge-base
 - ❑ unexpected, freshness, timeliness

Limitation of the Support-Confidence Framework

- Are s and c interesting in association rules: " $A \Rightarrow B$ " [s, c]? **Be careful!**
- Example: Suppose one school may have the following statistics on # of students who may play basketball and/or eat cereal:

	play-basketball	not play-basketball	sum (row)
eat-cereal	400	350	750
not eat-cereal	200	50	250
sum(col.)	600	400	1000

2-way contingency table

- Association rule mining may generate the following:
 - $play\text{-}basketball \Rightarrow eat\text{-}cereal$ [40%, 66.7%] (higher s & c)
- But this strong association rule is misleading: The overall % of students eating cereal is 75% > 66.7%, a more telling rule:
 - $\neg play\text{-}basketball \Rightarrow eat\text{-}cereal$ [35%, 87.5%] (high s & c)

The background of the slide is a complex, abstract composition. It features a dark, reddish-brown base with a network of thin, light-colored lines forming a mesh or web-like structure. Scattered throughout are numerous small, colored dots in shades of green, blue, and orange. In the upper left, there's a horizontal band with a grid of small, light-colored plus signs. Below this, there's a section with a grid of small, light-colored plus signs and a series of small, light-colored plus signs arranged in a horizontal line. In the lower left, there's a small, light-colored plus sign. The overall effect is a sense of depth and complexity, suggesting a data-driven or geometric theme.

Interestingness Measures: Lift and χ^2

Interestingness Measure: Lift

- Measure of dependent/correlated events: **lift**

$$\text{lift}(B, C) = \frac{c(B \rightarrow C)}{s(C)} = \frac{s(B \cup C)}{s(B) \times s(C)}$$

- Lift(B, C) may tell how B and C are correlated

- Lift(B, C) = 1: B and C are independent
- > 1: positively correlated
- < 1: negatively correlated

- For our example, $\text{lift}(B, C) = \frac{400 / 1000}{600 / 1000 \times 750 / 1000} = 0.89$
 $\text{lift}(B, \neg C) = \frac{200 / 1000}{600 / 1000 \times 250 / 1000} = 1.33$

- Thus, B and C are negatively correlated since $\text{lift}(B, C) < 1$;
 - B and $\neg C$ are positively correlated since $\text{lift}(B, \neg C) > 1$

Lift is more telling than s & c

	B	$\neg B$	Σ_{row}
C	400	350	750
$\neg C$	200	50	250
$\Sigma_{\text{col.}}$	600	400	1000

Interestingness Measure: χ^2

- Another measure to test correlated events: χ^2

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

- For the table on the right,

$$\chi^2 = \frac{(400 - 450)^2}{450} + \frac{(350 - 300)^2}{300} + \frac{(200 - 150)^2}{150} + \frac{(50 - 100)^2}{100} = 55.56$$

	B	$\neg B$	Σ_{row}
C	400 (450)	350 (300)	750
$\neg C$	200 (150)	50 (100)	250
Σ_{col}	600	400	1000

Expected value

Observed value

- By consulting a table of critical values of the χ^2 distribution, one can conclude that the chance for B and C to be independent is very low (< 0.01)
- χ^2 -test shows B and C are negatively correlated since the expected value is 450 but the observed is only 400
- Thus, χ^2 is also more telling than the support-confidence framework

Lift and χ^2 : Are They Always Good Measures?

- ❑ Null transactions: Transactions that contain neither B nor C
- ❑ Let's examine the new dataset D
 - ❑ BC (100) is much rarer than B¬C (1000) and ¬BC (1000), but there are many ¬B¬C (100000)
 - ❑ Unlikely B & C will happen together!
- ❑ But, $\text{Lift}(B, C) = 8.44 \gg 1$ (Lift shows B and C are strongly positively correlated!)
- ❑ $\chi^2 = 670$: Observed(BC) \gg expected value (11.85)
- ❑ *Too many null transactions may "spoil the soup"!*



	B	¬B	Σ_{row}
C	100	1000	1100
¬C	1000	100000	101000
$\Sigma_{\text{col.}}$	1100	101000	102100

null transactions

Contingency table with expected values added

	B	¬B	Σ_{row}
C	100 (11.85)	1000	1100
¬C	1000 (988.15)	100000	101000
$\Sigma_{\text{col.}}$	1100	101000	102100

The background of the slide is a complex, abstract composition. It features a central white banner with a subtle geometric pattern of thin lines and small plus signs. This banner is overlaid on a darker, textured background that includes a network of red and orange lines connecting various points, resembling a Voronoi diagram or a complex graph. There are also faint, repeating patterns of small plus signs and geometric shapes in the background. On the left side, there is a small, rectangular inset image showing a cluster of orange and red dots, possibly representing a galaxy or a specific data set, with a grid of plus signs overlaid on it.

Null Invariance Measures

Interestingness Measures & Null-Invariance

- ❑ **Null invariance**: Value does not change with the # of null-transactions
- ❑ A few interestingness measures: Some are null invariant

Measure	Definition	Range	Null-Invariant?
$\chi^2(A, B)$	$\sum_{i,j} \frac{(e(a_i, b_j) - o(a_i, b_j))^2}{e(a_i, b_j)}$	$[0, \infty]$	No
$Lift(A, B)$	$\frac{s(A \cup B)}{s(A) \times s(B)}$	$[0, \infty]$	No
$Allconf(A, B)$	$\frac{s(A \cup B)}{\max\{s(A), s(B)\}}$	$[0, 1]$	Yes
$Jaccard(A, B)$	$\frac{s(A \cup B)}{s(A) + s(B) - s(A \cup B)}$	$[0, 1]$	Yes
$Cosine(A, B)$	$\frac{s(A \cup B)}{\sqrt{s(A) \times s(B)}}$	$[0, 1]$	Yes
$Kulczynski(A, B)$	$\frac{1}{2} \left(\frac{s(A \cup B)}{s(A)} + \frac{s(A \cup B)}{s(B)} \right)$	$[0, 1]$	Yes
$MaxConf(A, B)$	$\max\left\{ \frac{s(A \cup B)}{s(A)}, \frac{s(A \cup B)}{s(B)} \right\}$	$[0, 1]$	Yes

χ^2 and lift are not null-invariant

Jaccard, Cosine, AllConf, MaxConf, and Kulczynski are null-invariant measures

Null Invariance: An Important Property

□ Why is null invariance crucial for the analysis of massive transaction data?

□ Many transactions may contain neither milk nor coffee!

milk vs. coffee contingency table

	<i>milk</i>	$\neg milk$	Σ_{row}
<i>coffee</i>	<i>mc</i>	$\neg mc$	<i>c</i>
$\neg coffee$	<i>m</i> $\neg c$	$\neg m$ $\neg c$	$\neg c$
Σ_{col}	<i>m</i>	$\neg m$	Σ

□ Lift and χ^2 are not null-invariant: not good to evaluate data that contain too many or too few null transactions!

□ Many measures are not null-invariant!

Null-transactions
w.r.t. *m* and *c*

Data set	<i>mc</i>	$\neg mc$	<i>m</i> $\neg c$	$\neg m$ $\neg c$	χ^2	<i>Lift</i>
<i>D</i> ₁	10,000	1,000	1,000	100,000	90557	9.26
<i>D</i> ₂	10,000	1,000	1,000	100	0	1
<i>D</i> ₃	100	1,000	1,000	100,000	670	8.44
<i>D</i> ₄	1,000	1,000	1,000	100,000	24740	25.75
<i>D</i> ₅	1,000	100	10,000	100,000	8173	9.18
<i>D</i> ₆	1,000	10	100,000	100,000	965	1.97

The background features a complex geometric pattern of thin, light-colored lines forming a network of triangles and polygons. Overlaid on this are several semi-transparent rectangular panels. The top-left panel shows a grid of small, light-colored plus signs. The bottom-left panel displays a dense cluster of orange and red dots, with a horizontal band of pink and white squares. The right side of the image has a vertical strip of white lines. A large, white, semi-transparent trapezoidal shape serves as a backdrop for the title text.

Comparison of Null-Invariant Measures

Comparison of Null-Invariant Measures

- ❑ Not all null-invariant measures are created equal
- ❑ Which one is better?
 - ❑ D_4 — D_6 differentiate the null-invariant measures
 - ❑ Kulc (Kulczynski 1927) holds firm and is in balance of both directional implications

2-variable contingency table

	<i>milk</i>	$\neg milk$	Σ_{row}
<i>coffee</i>	<i>mc</i>	$\neg mc$	<i>c</i>
$\neg coffee$	$m\neg c$	$\neg m\neg c$	$\neg c$
Σ_{col}	<i>m</i>	$\neg m$	Σ

All 5 are null-invariant

Data set	<i>mc</i>	$\neg mc$	$m\neg c$	$\neg m\neg c$	<i>AllConf</i>	Jaccard	<i>Cosine</i>	<i>Kulc</i>	<i>MaxConf</i>
D_1	10,000	1,000	1,000	100,000	0.91	0.83	0.91	0.91	0.91
D_2	10,000	1,000	1,000	100	0.91	0.83	0.91	0.91	0.91
D_3	100	1,000	1,000	100,000	0.09	0.05	0.09	0.09	0.09
D_4	1,000	1,000	1,000	100,000	0.5	0.33	0.5	0.5	0.5
D_5	1,000	100	10,000	100,000	0.09	0.09	0.29	0.5	0.91
D_6	1,000	10	100,000	100,000	0.01	0.01	0.10	0.5	0.99

Subtle: They disagree on those cases

Analysis of DBLP Coauthor Relationships

- DBLP: Computer science research publication bibliographic database
 - > 3.8 million entries on authors, paper, venue, year, and other information

ID	Author <i>A</i>	Author <i>B</i>	$s(A \cup B)$	$s(A)$	$s(B)$	Jaccard	<i>Cosine</i>	<i>Kulc</i>
1	Hans-Peter Kriegel	Martin Ester	28	146	54	0.163 (2)	0.315 (7)	0.355 (9)
2	Michael Carey	Miron Livny	26	104	58	0.191 (1)	0.335 (4)	0.349 (10)
3	Hans-Peter Kriegel	Joerg Sander	24	146	36	0.152 (3)	0.331 (5)	0.416 (8)
4	Christos Faloutsos	Spiros Papadimitriou	20	162	26	0.119 (7)	0.308 (10)	0.446 (7)
5	Hans-Peter Kriegel	Martin Pfeifle	18	146	18	0.123 (6)	0.351 (2)	0.562 (2)
6	Hector Garcia-Molina	Wilburt Labio	16	144	18	0.110 (9)	0.314 (8)	0.500 (4)
7	Divyakant Agrawal	Wang Hsiung	16	120	16	0.133 (5)	0.365 (1)	0.567 (1)
8	Elke Rundensteiner	Murali Mani	16	104	20	0.148 (4)	0.351 (3)	0.477 (6)
9	Divyakant Agrawal	Oliver Po	12	120	12	0.100 (10)	0.316 (6)	0.550 (3)
10	Gerhard Weikum	Martin Theobald	12	106	14	0.111 (8)	0.312 (9)	0.485 (5)

Advisor-advisee relation: Kulc: high, Jaccard: low, cosine: middle

- Which pairs of authors are strongly related?
 - Use Kulc to find Advisor-advisee, close collaborators

Imbalance Ratio with Kulczynski Measure

- IR (Imbalance Ratio): measure the imbalance of two itemsets A and B in rule implications:

$$IR(A, B) = \frac{|s(A) - s(B)|}{s(A) + s(B) - s(A \cup B)}$$

- Kulczynski and Imbalance Ratio (IR) together present a clear picture for all the three datasets D_4 through D_6
 - D_4 is neutral & balanced; D_5 is neutral but imbalanced
 - D_6 is neutral but very imbalanced

Data set	mc	$\neg mc$	$m\neg c$	$\neg m\neg c$	Jaccard	Cosine	Kulc	IR
D_1	10,000	1,000	1,000	100,000	0.83	0.91	0.91	0
D_2	10,000	1,000	1,000	100	0.83	0.91	0.91	0
D_3	100	1,000	1,000	100,000	0.05	0.09	0.09	0
D_4	1,000	1,000	1,000	100,000	0.33	0.5	0.5	0
D_5	1,000	100	10,000	100,000	0.09	0.29	0.5	0.89
D_6	1,000	10	100,000	100,000	0.01	0.10	0.5	0.99

The background of the slide is a complex, abstract composition. It features a central white rectangular area where the word 'Summary' is written. This central area is flanked by two large, light gray triangular shapes that point towards the center. The background is further decorated with a grid of small gray plus signs and a network of thin, intersecting lines in shades of red, orange, and green, creating a web-like or geometric pattern. In the top-left corner, there is a small, semi-transparent inset image showing a cluster of orange and red dots, possibly representing a galaxy or a data visualization. The overall aesthetic is modern and scientific.

Summary

What Measures to Choose for Effective Pattern Evaluation?

- ❑ Null value cases are predominant in many large datasets
 - ❑ Neither milk nor coffee is in most of the baskets; neither Mike nor Jim is an author in most of the papers;
- ❑ *Null-invariance* is an important property
- ❑ Lift, χ^2 and cosine are good measures if null transactions are not predominant
 - ❑ Otherwise, *Kulczynski + Imbalance Ratio* should be used to judge the interestingness of a pattern
- ❑ Exercise: Mining research collaborations from research bibliographic data
 - ❑ Find a group of frequent collaborators from research bibliographic data (e.g., DBLP)
 - ❑ Can you find the likely advisor-advisee relationship and during which years such a relationship happened?
 - ❑ Ref.: C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo, "Mining Advisor-Advisee Relationships from Research Publication Networks", KDD'10

Summary: Pattern Evaluation

- ❑ Interestingness Measures in Pattern Mining
- ❑ Interestingness Measures: Lift and χ^2
- ❑ Null-Invariant Measures
- ❑ Comparison of Interestingness Measures

Recommended Readings

- ❑ C. C. Aggarwal and P. S. Yu. A New Framework for Itemset Generation. PODS'98
- ❑ S. Brin, R. Motwani, and C. Silverstein. Beyond market basket: Generalizing association rules to correlations. SIGMOD'97
- ❑ M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo. Finding interesting rules from large sets of discovered association rules. CIKM'94
- ❑ E. Omiecinski. Alternative Interest Measures for Mining Associations. TKDE'03
- ❑ P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the Right Interestingness Measure for Association Patterns. KDD'02
- ❑ T. Wu, Y. Chen and J. Han, Re-Examination of Interestingness Measures in Pattern Mining: A Unified Framework, Data Mining and Knowledge Discovery, 21(3):371-397, 2010

The background is a complex collage of abstract elements. It features a grid of small grey plus signs, a network of red lines connecting green dots, and a large white triangular shape in the center. On the left, there is a vertical strip with a blue and orange gradient and a series of small, colorful dots. The overall aesthetic is technical and data-driven.

Mining Diverse Patterns

Mining Diverse Patterns

- ❑ Mining Multiple-Level Associations
- ❑ Mining Multi-Dimensional Associations
- ❑ Mining Quantitative Associations
- ❑ Mining Negative Correlations
- ❑ Mining Compressed and Redundancy-Aware Patterns

The background features a complex, abstract design. It includes a network of thin, light-colored lines forming a web-like structure. Overlaid on this are various data visualizations: a grid of small, light-colored plus signs, a series of small, colorful dots (green, blue, yellow) connected by lines, and a large, semi-transparent white triangle that serves as a backdrop for the title. The overall color palette is muted, with shades of brown, grey, and white, accented by the colors of the data points.

Mining Multiple-Level Associations

Mining Multiple-Level Frequent Patterns

- Items often form hierarchies

- Ex.: Dairyland 2% milk;
Wonder wheat bread

- How to set min-support thresholds?

- Uniform min-support across multiple levels (reasonable?)

- Level-reduced min-support: Items at the lower level are expected to have lower support

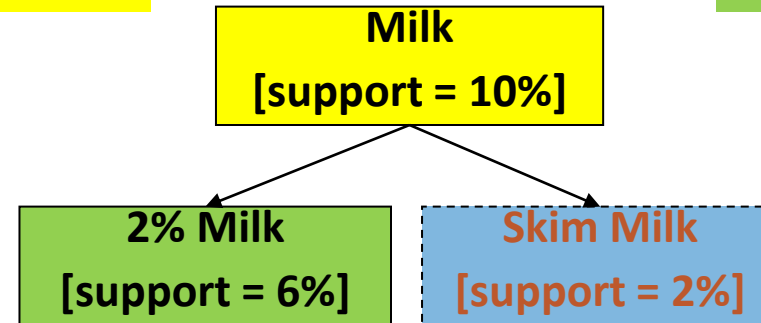
- Efficient mining: *Shared* multi-level mining

- Use the lowest min-support to pass down the set of candidates

Uniform support

Level 1
min_sup = 5%

Level 2
min_sup = 5%



Reduced support

Level 1
min_sup = 5%

Level 2
min_sup = 1%

Redundancy Filtering at Mining Multi-Level Associations

- ❑ Multi-level association mining may generate many redundant rules
- ❑ Redundancy filtering: Some rules may be redundant due to “ancestor” relationships between items
 - ❑ milk \Rightarrow wheat bread [support = 8%, confidence = 70%] (1)
 - ❑ 2% milk \Rightarrow wheat bread [support = 2%, confidence = 72%] (2)
 - ❑ Suppose the 2% milk sold is about $\frac{1}{4}$ of milk sold in gallons
 - ❑ (2) should be able to be “derived” from (1)
- ❑ A rule is *redundant* if its support is close to the “expected” value, according to its “ancestor” rule, and it has a similar confidence as its “ancestor”
 - ❑ Rule (1) is an ancestor of rule (2), which one to prune?

Customized Min-Supports for Different Kinds of Items


- ❑ We have used the same min-support threshold for all the items or item sets to be mined in each association mining
- ❑ In reality, some items (e.g., diamond, watch, ...) are valuable but less frequent
- ❑ It is necessary to have customized min-support settings for different kinds of items
- ❑ One Method: Use **group-based “individualized” min-support**
 - ❑ E.g., {diamond, watch}: 0.05%; {bread, milk}: 5%; ...
 - ❑ How to mine such rules efficiently?
 - ❑ Existing scalable mining algorithms can be easily extended to cover such cases

The background of the slide is a complex, abstract composition. It features a dark, reddish-brown base with a network of thin, light-colored lines forming a mesh or web-like structure. Scattered throughout are numerous small, colored dots in shades of green, blue, and orange. In the upper left, there is a horizontal band with a repeating pattern of small, stylized symbols. On the left side, there is a vertical strip containing a cluster of orange and red dots, with a horizontal bar of pink and white squares overlaid. The overall aesthetic is technical and data-driven.

Mining Multi-Dimensional Associations

Mining Multi-Dimensional Associations

- ❑ Single-dimensional rules (e.g., items are all in “product” dimension)
 - ❑ $\text{buys}(X, \text{“milk”}) \Rightarrow \text{buys}(X, \text{“bread”})$
- ❑ Multi-dimensional rules (i.e., items in ≥ 2 dimensions or predicates)
 - ❑ Inter-dimension association rules (*no repeated predicates*)
 - ❑ $\text{age}(X, \text{“18-25”}) \wedge \text{occupation}(X, \text{“student”}) \Rightarrow \text{buys}(X, \text{“coke”})$
 - ❑ Hybrid-dimension association rules (*repeated predicates*)
 - ❑ $\text{age}(X, \text{“18-25”}) \wedge \text{buys}(X, \text{“popcorn”}) \Rightarrow \text{buys}(X, \text{“coke”})$
- ❑ Attributes can be categorical or numerical
 - ❑ Categorical Attributes (e.g., *profession*, *product*: no ordering among values): Data cube for inter-dimension association
 - ❑ Quantitative Attributes: Numeric, implicit ordering among values—discretization, clustering, and gradient approaches

The background of the slide is a complex, abstract composition. It features a dark, reddish-brown base with a network of thin, light-colored lines forming a mesh or web-like structure. Overlaid on this are various data visualization elements: a horizontal band of small, colorful dots (green, blue, yellow) and lines; a vertical strip of orange and red dots on the left; and a central white area containing the title. The overall aesthetic is technical and data-driven.

Mining Quantitative Associations

Mining Quantitative Associations

- ❑ Mining associations with numerical attributes
 - ❑ Ex.: Numerical attributes: **age** and **salary**
- ❑ Methods
 - ❑ Static discretization based on predefined concept hierarchies
 - ❑ Discretization on each dimension with hierarchy
 - ❑ age: {0-10, 10-20, ..., 90-100} \rightarrow {young, mid-aged, old}
 - ❑ Dynamic discretization based on data distribution
 - ❑ Clustering: Distance-based association
 - ❑ First one-dimensional clustering, then association
 - ❑ Deviation analysis:
 - ❑ Gender = female \Rightarrow Wage: mean=\$7/hr (overall mean = \$9)

Mining Extraordinary Phenomena in Quantitative Association Mining

- ❑ Mining extraordinary (i.e., interesting) phenomena
 - ❑ Ex.: Gender = female \Rightarrow Wage: mean=\$7/hr (overall mean = \$9)
 - ❑ LHS: a subset of the population
 - ❑ RHS: an extraordinary behavior of this subset
- ❑ The rule is accepted only if a statistical test (e.g., Z-test) confirms the inference with high confidence
- ❑ Subrule: Highlights the extraordinary behavior of a subset of the population of the super rule
 - ❑ Ex.: (Gender = female) \wedge (South = yes) \Rightarrow mean wage = \$6.3/hr
- ❑ Rule condition can be categorical or numerical (quantitative rules)
 - ❑ Ex.: Education in [14-18] (yrs) \Rightarrow mean wage = \$11.64/hr
- ❑ Efficient methods have been developed for mining such extraordinary rules (e.g., Aumann and Lindell@KDD'99)

The background features a complex network of thin, light-colored lines forming a mesh-like structure. Overlaid on this are various data points: small green dots, larger blue dots, and a cluster of orange dots on the left. A horizontal band of semi-transparent white and light blue shapes runs across the middle, serving as a backdrop for the title. The overall aesthetic is technical and data-driven.

Mining Negative Correlations

Rare Patterns vs. Negative Patterns

❑ Rare patterns

- ❑ Very low support but interesting (e.g., buying Rolex watches)
- ❑ How to mine them? Setting individualized, group-based min-support thresholds for different groups of items

❑ Negative patterns

- ❑ Negatively correlated: Unlikely to happen together
- ❑ Ex.: Since it is unlikely that the same customer buys both a **Ford Expedition** (an SUV car) and a **Ford Fusion** (a hybrid car), buying a **Ford Expedition** and buying a **Ford Fusion** are likely negatively correlated patterns
- ❑ How to define negative patterns?

Defining Negative Correlated Patterns

- A support-based definition
 - If itemsets A and B are both frequent but rarely occur together, i.e.,
 $\text{sup}(A \cup B) \ll \text{sup}(A) \times \text{sup}(B)$
 - Then A and B are negatively correlated
- Is this a good definition for large transaction datasets?
- Ex.: Suppose a store sold two needle packages A and B 100 times each, but only one transaction contained both A and B
 - When there are in total 200 transactions, we have
 - $s(A \cup B) = 0.005, s(A) \times s(B) = 0.25, s(A \cup B) \ll s(A) \times s(B)$
 - But when there are 10^5 transactions, we have
 - $s(A \cup B) = 1/10^5, s(A) \times s(B) = 1/10^3 \times 1/10^3, s(A \cup B) > s(A) \times s(B)$
 - What is the problem?—Null transactions: The support-based definition is not null-invariant!

Does this remind you the definition of *lift*?

Defining Negative Correlation: Need Null-Invariance in Definition

- ❑ A good definition on negative correlation should take care of the null-invariance problem
 - ❑ Whether two itemsets A and B are negatively correlated should not be influenced by the number of null-transactions
- ❑ A Kulczynski measure-based definition
 - ❑ If itemsets A and B are frequent but
$$(s(A \cup B)/s(A) + s(A \cup B)/s(B))/2 < \epsilon,$$
where ϵ is a negative pattern threshold, then A and B are negatively correlated
- ❑ For the same needle package problem:
 - ❑ No matter there are in total 200 or 10^5 transactions
 - ❑ If $\epsilon = 0.01$, we have
$$(s(A \cup B)/s(A) + s(A \cup B)/s(B))/2 = (0.01 + 0.01)/2 < \epsilon$$

The background is a complex collage of abstract elements. It features a grid of small grey plus signs, a network of red lines connecting green dots, and a large white triangular shape in the center. On the left, there is a vertical strip with a blue and orange gradient, and a horizontal strip with a purple and white gradient. The overall aesthetic is technical and data-driven.

Mining Compressed Patterns

Mining Compressed Patterns

Pat-ID	Item-Sets	Support
P1	{38,16,18,12}	205227
P2	{38,16,18,12,17}	205211
P3	{39,38,16,18,12,17}	101758
P4	{39,16,18,12,17}	161563
P5	{39,16,18,12}	161576

- ❑ Closed patterns
 - ❑ P1, P2, P3, P4, P5
 - ❑ Emphasizes too much on support
 - ❑ There is no compression
- ❑ Max-patterns
 - ❑ P3: information loss
- ❑ Desired output (a good balance):
 - ❑ **P2, P3, P4**

❑ Why mining compressed patterns?

- ❑ Too many scattered patterns but not so meaningful

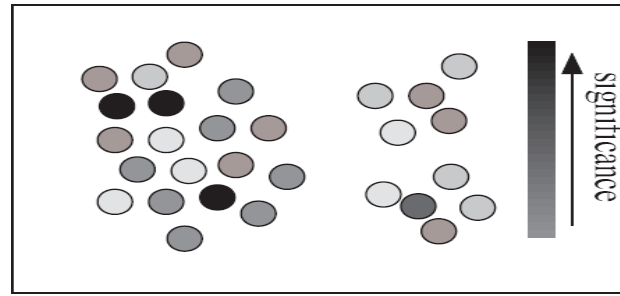
❑ Pattern distance measure

$$Dist(P_1, P_2) = 1 - \frac{|T(P_1) \cap T(P_2)|}{|T(P_1) \cup T(P_2)|}$$

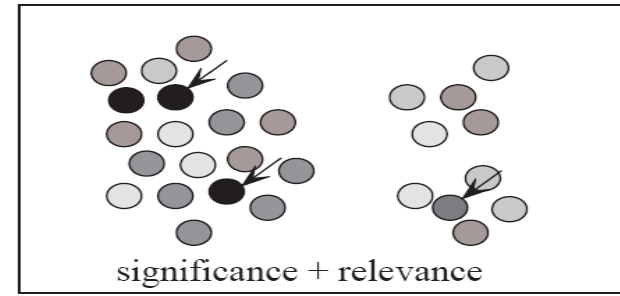
- ❑ δ -clustering: For each pattern P, find all patterns which can be expressed by P and whose distance to P is within δ (δ -cover)
- ❑ All patterns in the cluster can be represented by P
- ❑ Method for efficient, direct mining of compressed frequent patterns (e.g., D. Xin, J. Han, X. Yan, H. Cheng, "On Compressing Frequent Patterns", Knowledge and Data Engineering, 60:5-29, 2007)

Redundancy-Aware Top-k Patterns

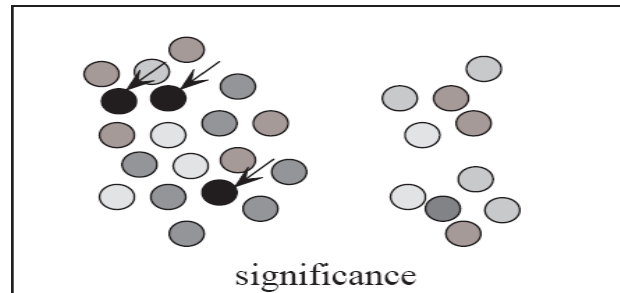
- Desired patterns: high significance & low redundancy



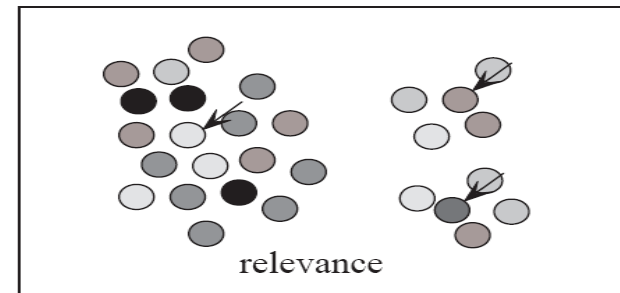
(a) a set of patterns



(b) redundancy-aware top- k



(c) traditional top- k



(d) summarization

- Method: Use MMS (Maximal Marginal Significance) for measuring the combined significance of a pattern set
- Xin et al., Extracting Redundancy-Aware Top-K Patterns, KDD'06

The background of the slide is a complex, abstract composition. It features a central white rectangular area where the word 'Summary' is written. This central area is flanked by two large, light gray triangular shapes that point towards the center. The background is further decorated with a grid of small gray plus signs and a network of thin, intersecting lines in shades of brown and orange. Scattered throughout are small, colorful dots in green, blue, and orange. In the top-left corner, there is a horizontal band with a repeating pattern of small, stylized symbols. In the bottom-left corner, there is a small, square inset image showing a cluster of orange and brown dots on a light background, with a grid of plus signs overlaid on it.

Summary

Summary: Mining Diverse Patterns

- ❑ Efficient methods have been developed for mining various kinds of patterns
 - ❑ Mining Multiple-Level Associations
 - ❑ Mining Multi-Dimensional Associations
 - ❑ Mining Quantitative Associations
 - ❑ Mining Negative Correlations
 - ❑ Mining Compressed and Redundancy-Aware Patterns

Recommended Readings

- ❑ R. Srikant and R. Agrawal, “Mining generalized association rules”, VLDB'95
- ❑ Y. Aumann and Y. Lindell, “A Statistical Theory for Quantitative Association Rules”, KDD'99
- ❑ K. Wang, Y. He, J. Han, “Pushing Support Constraints Into Association Rules Mining”, IEEE Trans. Knowledge and Data Eng. 15(3): 642-658, 2003
- ❑ D. Xin, J. Han, X. Yan and H. Cheng, "On Compressing Frequent Patterns", Knowledge and Data Engineering, 60(1): 5-29, 2007
- ❑ D. Xin, H. Cheng, X. Yan, and J. Han, "Extracting Redundancy-Aware Top-K Patterns", KDD'06
- ❑ J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent Pattern Mining: Current Status and Future Directions", Data Mining and Knowledge Discovery, 15(1): 55-86, 2007