SI 671/721:

Mining Itemset Data

Lecture 3

Fall 2021

Instructor: Prof. Paramveer Dhillon

dhillonp@umich.edu

University of Michigan

UMSI

Administrivia

- HW1 is out today on Canvas.
- It is due on 10/4 (2 weeks from today).
- It covers *itemsets.*
- Please get started early!

Pattern-based Itemset Mining:
Frequent Itemsets & Association Rules

What is a “pattern”?

- Pattern: A structure of attributes that represents the intrinsic and important properties of data objects.

- For itemset data, a “pattern” can be:
  - A frequent subset of items.
  - An association rule.
  - A ***correlation*** of two items.

Recap: What is a subset and a superset?

For two itemsets X_1 and X_2:

If every item in X_1 is also in X_2 , then
X_1 is a ***subset*** of X_2 (if X_1 appears in X_2)
X_2 is a ***superset*** of X_1 (if X_2 contains X_1)

{4,9} is a subset of {1,3,4,7,9,12}

Recall that an itemset is just an unordered list of items.

Frequent itemsets

A k-itemset: $X = \{x_1, x_2, \ldots, x_k\}$

Support: frequency of X in a database

- Absolute support: number of transactions that contain X.
- Relative support: fraction of transactions that contain X.

An itemset X is frequent if its support $sup(X)$ is no less than a threshold *min_sup*.

Example: shopping baskets

TID	Items Bought
1	🍺 🍼 🍉
2	🍺 🍭 🍼 🍋
3	🍼 🍺 🍋
4	🍼 🍭
5	🍷 🍪 🍺 🍞 🍋

{🍺}: support= 80%

{🍺, 🍼}: support=60%

{🍺, 🍼 ,🍋}: support= 40%

If *min_sup=50%,* then {🍺} & {🍺, 🍼} are frequent.

Association rules

$$X \longrightarrow Y$$

Both X and Y are itemsets

Support [P(x,y)]: probability that a transaction contains both X and Y.

Confidence [P(y|x)]: conditional probability that a transaction that contains X also contains Y.

Example: Shopping Basket

Calculating support and confidence for an association rule.

TID	Items Bought
1	🍺 🍼 🍉
2	🍺 🍭 🍼 🍋
3	🍼 🍺 🍋
4	🍼 🍭
5	🍷 🍪 🍺 🍞 🍋

{🍺, 🍼}: support=60%

{🍺} -> {🍼}

support [P(x,y)]=60%

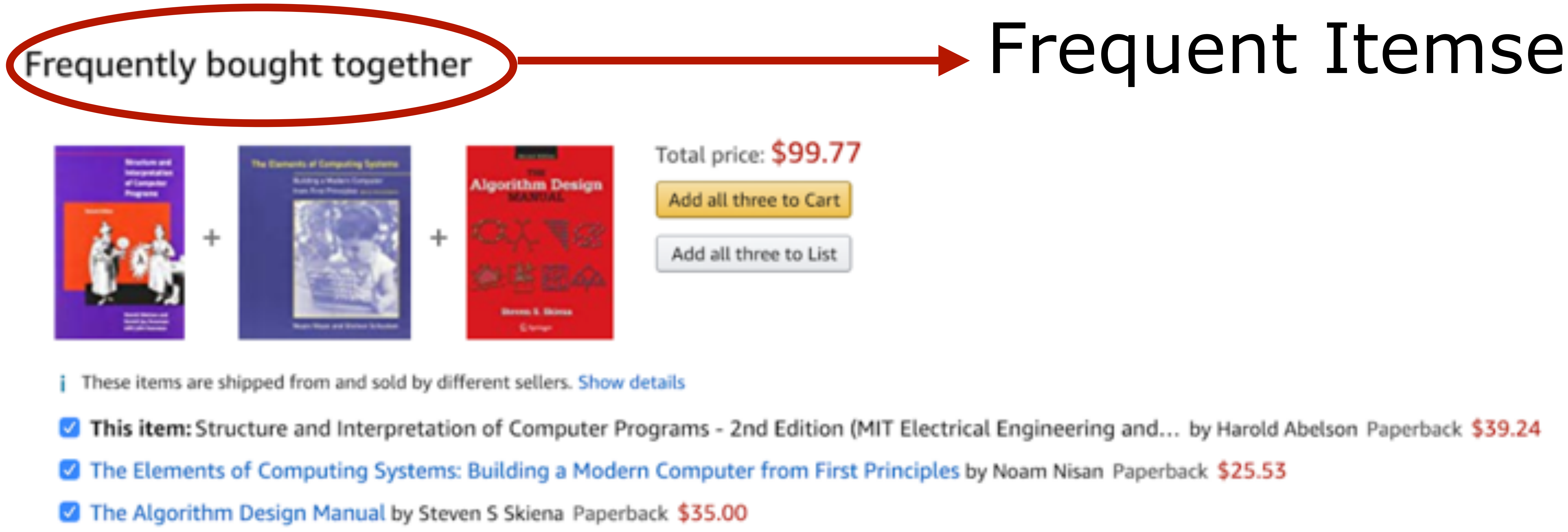
confidence [P(y|x)] =75%

[60%, 75%]

{🍺, 🍼} -> {🍋}

[40%, 66.7%]

Frequent itemsets and Recommendations



Frequently bought together

Frequent Itemsets

Total price: $99.77

Add all three to Cart

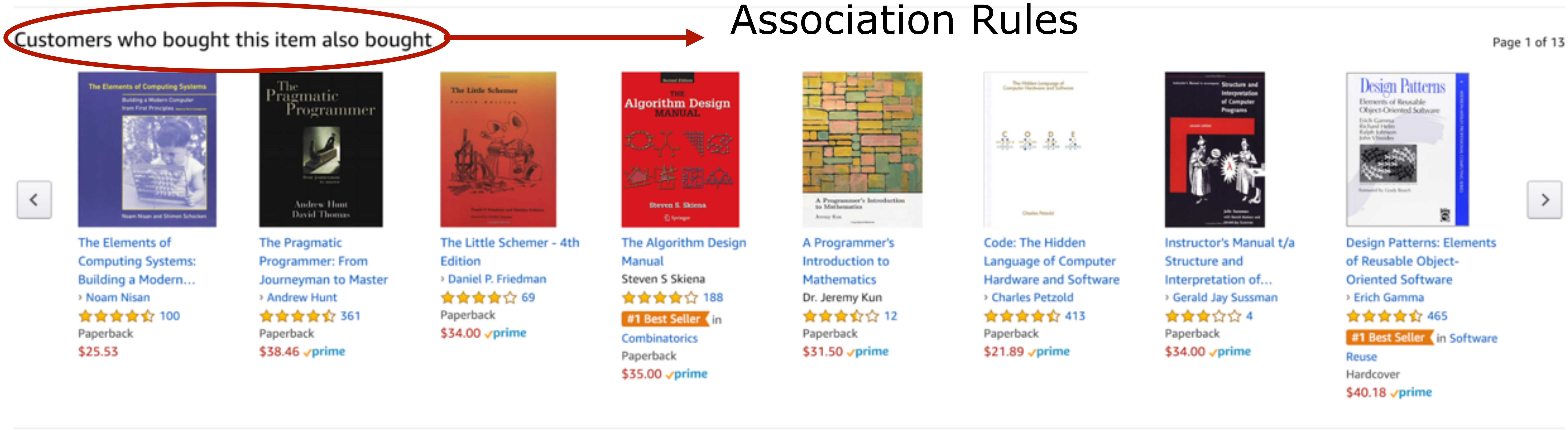
Add all three to List

i These items are shipped from and sold by different sellers. Show details

☑ **This item:** Structure and Interpretation of Computer Programs - 2nd Edition (MIT Electrical Engineering and... by Harold Abelson Paperback $39.24

☑ The Elements of Computing Systems: Building a Modern Computer from First Principles by Noam Nisan Paperback $25.53

☑ The Algorithm Design Manual by Steven S Skiena Paperback $35.00



Customers who bought this item also bought

Association Rules

Page 1 of 13

The Elements of Computing Systems: Building a Modern...
› Noam Nisan
★★★★☆ 100
Paperback
$25.53

The Pragmatic Programmer: From Journeyman to Master
› Andrew Hunt
★★★★☆ 361
Paperback
$38.46 ✓prime

The Little Schemer - 4th Edition
› Daniel P. Friedman
★★★★☆ 69
Paperback
$34.00 ✓prime

The Algorithm Design Manual
Steven S Skiena
★★★★☆ 188
#1 Best Seller in Combinatorics
Paperback
$35.00 ✓prime

A Programmer's Introduction to Mathematics
Dr. Jeremy Kun
★★★☆☆ 12
Paperback
$31.50 ✓prime

Code: The Hidden Language of Computer Hardware and Software
› Charles Petzold
★★★★☆ 413
Paperback
$21.89 ✓prime

Instructor's Manual t/a Structure and Interpretation of...
› Gerald Jay Sussman
★★★☆☆ 4
Paperback
$34.00 ✓prime

Design Patterns: Elements of Reusable Object-Oriented Software
› Erich Gamma
★★★★☆ 465
#1 Best Seller in Software Reuse
Hardcover
$40.18 ✓prime

How to find frequent itemsets?

- Scan every transaction in database.
- Enumerate the possible subsets.
- Check whether their frequency is above the minimal support.
- First find frequent itemsets, then calculate the confidence of associations.

A simple algorithm

To get frequency, we just need to count.

And counting is easy, right?

Except when there are too many possible candidates!

- 1,000 items: 1M possible 2-itemsets; 1B possible 3-itemsets

Scaling up frequent pattern mining

Intuition: the downward closure property

–Any subset of a frequent itemset must be frequent

–If {🍺,🍼,🍋} is frequent, so is {🍺,🍼}

–i.e., every transaction having {🍺,🍼,🍋} also contains {🍺,🍼}

What does this imply?

The Apriori Algorithm for Frequent Itemset Mining

Apriori: candidate generation & test

Apriori pruning principle: If any itemset is infrequent, none of its supersets need to be considered.

(Agrawal & Srikant @VLDB’94, Mannila et al. @KDD’ 94)

Intuition about the Apriori algorithm

- Initially, scan database once to get frequent 1-itemset (single items).

- Generate size (k+1) candidate itemsets from length k frequent itemsets.

- Test (count) the candidates against database.

- Terminate when no frequent or candidate set can be generated.

The Apriori algorithm - an example

TID	Items
1	🍺 🍼 🍉
2	🍺 🍭 🍼 🍋
3	🍼 🍺 🍋
4	🍼 🍭
5	🍷 🍺 🍋

→ First scan of DB

1-itemsets

TID	Count
{🍺}	4
{🍼}	4
{🍉}	1
{🍭}	2
{🍋}	2
{🍷}	1

→

Frequent
1-itemsets

TID	Count
{🍺}	4
{🍼}	4
{🍭}	2
{🍋}	2

Minimal Support = 2

The Apriori algorithm continued

TID	Items
1	🍺 🍼 🍉
2	🍺 🍭 🍼 🍋
3	🍼 🍺 🍋
4	🍼 🍭
5	🍷 🍺 🍋

Frequent 1-itemsets

TID	Count
{🍺}	4
{🍼}	4
{🍭}	2
{🍋}	2

Candidate 2-itemsets

TID
{🍺,🍼}
{🍺,🍭}
{🍺,🍋}
{🍼,🍭}
{🍼,🍋}
{🍭,🍋}

Candidate generation (self-join)

TID	Count
{🍺,🍼}	3
{🍺,🍭}	1
{🍺,🍋}	3
{🍼,🍭}	2
{🍼,🍋}	2
{🍭,🍋}	1

2nd scan of DB

Minimal Support = 2

The Apriori algorithm continued

TID	Items
1	🍺 🍼 🍉
2	🍺 🍭 🍼 🍋
3	🍼 🍺 🍋
4	🍼 🍭
5	🍷 🍺 🍋

Frequent 2-itemsets

TID	Count
{🍺, 🍼}	3
{🍺, 🍋}	3
{🍼, 🍭}	2
{🍼, 🍋}	2

Candidate 3-itemsets

TID
{🍺, 🍼, 🍋}
{🍺, 🍼, 🍭}
{🍼, 🍭, 🍋}

3rd scan of DB

TID	Count
{🍺, 🍼, 🍋}	2
{🍺, 🍼, 🍭}	1
{🍼, 🍭, 🍋}	1

Candidate generation
(self-join)

The Apriori algorithm continued

TID	Items
1	🍺 🍼 🍉
2	🍺 🍭 🍼 🍋
3	🍼 🍺 🍋
4	🍼 🍭
5	🍷 🍺 🍋

Minimal Support = 2

Frequent 1-itemsets

TID	Count
{🍺}	4
{🍼}	4
{🍭}	2
{🍋}	2

Frequent 2-itemsets

TID	Count
{🍺, 🍼}	3
{🍺, 🍋}	3
{🍼, 🍭}	2
{🍼, 🍋}	2

Frequent 3-itemsets

TID	Count
{🍺, 🍼, 🍋}	2

The Apriori algorithm - pseudo code

Input: Database **D**, minimal support ***min_sup***
C_k: candidate itemsets of size k
L_k: frequent itemsets of size k

```
L_1 = {frequent single items};

for (k = 1; $L_k$ !=∅; k++) do
    $C_{k+1}$ = candidates of length k+1 generated from $L_k$;
    for each transaction t in D do
          increment the count of each candidate in $C_{k+1}$ that appear in t
          $L_{k+1}$  = candidates in $C_{k+1}$ with support >= min_sup

return {$L_1$, $L_2$, …, $L_{k-1}$}
```

Challenges of Apriori

- Needs multiple scans of database
- Huge number of candidates (most are not frequent)
- Tedious workload of counting the frequency of every candidate

Further improvements to Apriori

- Reduce passes of database scans
- Shrink number of candidates
- Sampling and approximation
- Distributed counting (e.g., Map-Reduce)
- Refer to [Han, Kamber, Pei] Chapter 6,7

How do we evaluate the Frequent itemsets?

Evaluation of frequent itemsets

- We derived all the frequent itemsets from our data.
- How do we use them?
- Are all frequent patterns interesting (i.e., support decisions)?
- Need a way to evaluate the true knowledge from frequent patterns.

Max and Closed Patterns

Closed pattern: An itemset *X* is closed if *X* is frequent and there exists no super-pattern with the same support as *X*.

Max pattern: An itemset *X* is a max-pattern if *X* is frequent and there exists no super-pattern that is also frequent.

Example

TID	Items Bought
1	🍺🍼🍉
2	🍺🍭🍼🍋
3	🍼🍺🍋
4	🍼🍭
5	🍷🍪🍺🍞🍋

min_sup=40%

{🍺, 🍼}: sup=60%

{🍺, 🍼, 🍋}: sup= 40%

{🍺 , 🍼 }: closed pattern

{ 🍺, 🍼, 🍋}: max-pattern and closed pattern

Association Rule Mining

Frequency → Association

Find X→Y that has a high support and a high confidence.

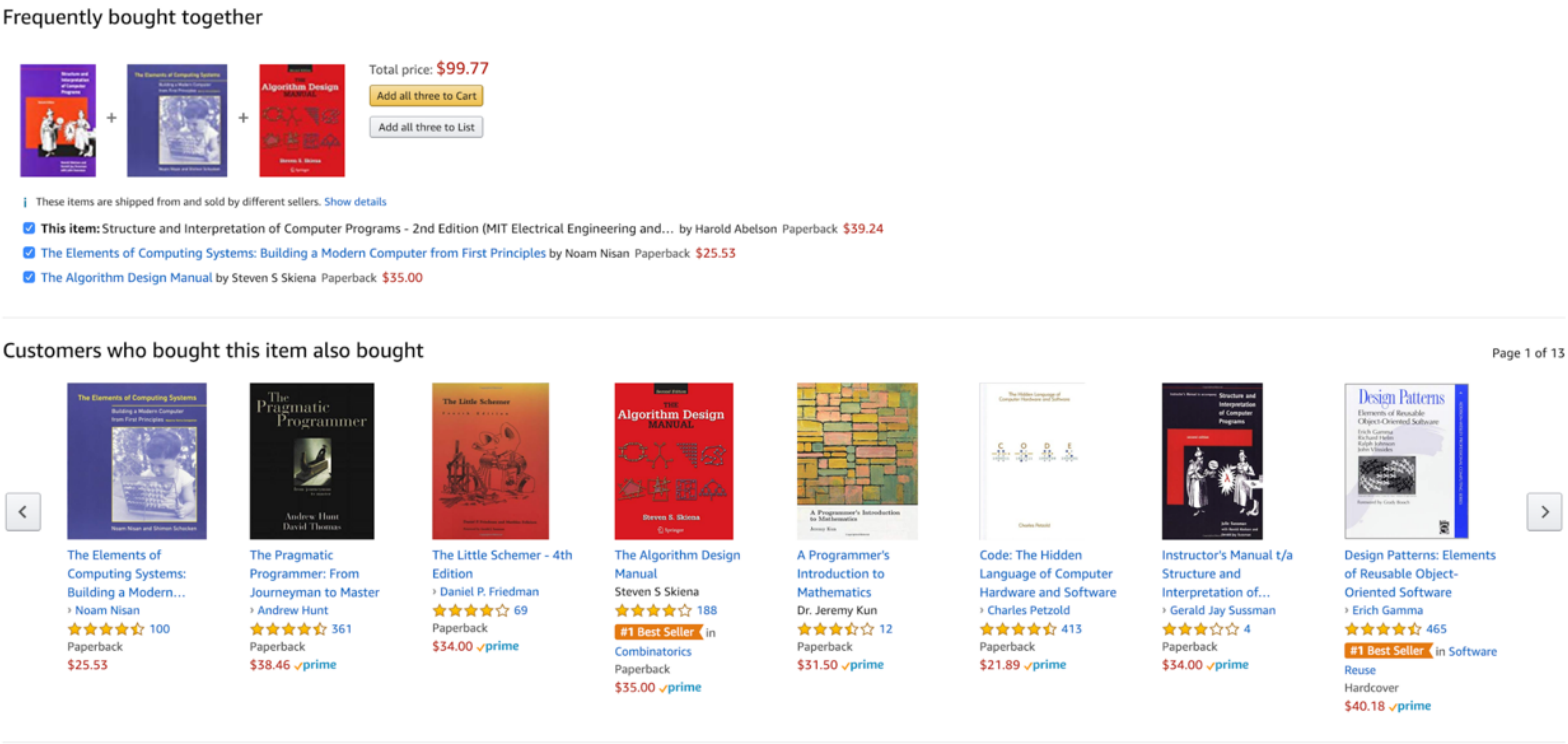
Support: $P(X, Y)$

Confidence: $P(Y|X)$

X → Y: [support, confidence]

Find frequent itemset (X, Y) first, then check the confidence (conditional probability)

Association Rules for Recommendation



Frequently bought together

Total price: $99.77

Add all three to Cart

Add all three to List

i These items are shipped from and sold by different sellers. Show details

☑ **This item:** Structure and Interpretation of Computer Programs - 2nd Edition (MIT Electrical Engineering and… by Harold Abelson Paperback $39.24

☑ The Elements of Computing Systems: Building a Modern Computer from First Principles by Noam Nisan Paperback $25.53

☑ The Algorithm Design Manual by Steven S Skiena Paperback $35.00

Customers who bought this item also bought

Page 1 of 13

The Elements of Computing Systems: Building a Modern…
› Noam Nisan
★★★★☆ 100
Paperback
$25.53

The Pragmatic Programmer: From Journeyman to Master
› Andrew Hunt
★★★★☆ 361
Paperback
$38.46 ✓prime

The Little Schemer - 4th Edition
› Daniel P. Friedman
★★★★☆ 69
Paperback
$34.00 ✓prime

The Algorithm Design Manual
Steven S Skiena
★★★★☆ 188
#1 Best Seller in Combinatorics
Paperback
$35.00 ✓prime

A Programmer's Introduction to Mathematics
Dr. Jeremy Kun
★★★☆☆ 12
Paperback
$31.50 ✓prime

Code: The Hidden Language of Computer Hardware and Software
› Charles Petzold
★★★★☆ 413
Paperback
$21.89 ✓prime

Instructor's Manual t/a Structure and Interpretation of…
› Gerald Jay Sussman
★★★☆☆ 4
Paperback
$34.00 ✓prime

Design Patterns: Elements of Reusable Object-Oriented Software
› Erich Gamma
★★★★☆ 465
#1 Best Seller in Software Reuse
Hardcover
$40.18 ✓prime

Association Rules for Classification

Y can be a class label (treated as an item)

e.g., X = {a URL, an image}; Y = {spam}

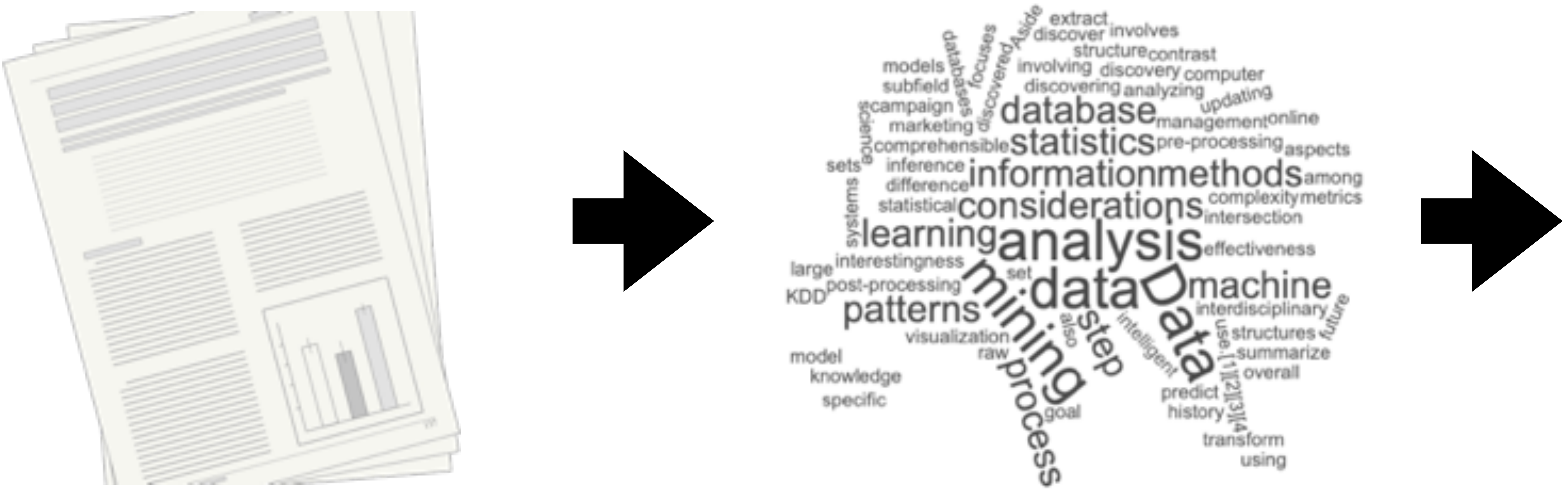
X → Y:

- Support: 1% of all emails
- Confidence: 90% of emails that have this URL and this image are spams.

Rule: classify as spam if X appears in email

Multiple rules (features) can be blended using machine learning!

Frequent patterns/association for classification



Data Science or Biology ?

- Represent data object as a set of features
- Features = items
- Features = frequent patterns / associations
- Then apply any classification algorithm

Are all associations “interesting”?

Association → Correlation

Customers who buy “computer games” → buy “videos”

- [40%, 66.7%]
- Support and confidence are misleading
- What if the overall probability of buying videos is 75%?

Association → Correlation

Customers who buy “computer games” → buy “videos”

- [40%, 66.7%]
- Support and confidence are misleading
- What if the overall probability of buying videos is 75%?

Customers who buy “computer games” → **not** buy “videos”

- [20%, 33.3%]
- More accurate, although with lower support and confidence

Solution: Measure **correlation** instead of conditional probabilities

Measuring correlation

Support and confidence are not sufficient to indicate true interestingness of patterns.

Measure correlation through the 2-way contingency table:

	Games	Not Games	Sum (row)
Videos	4,000	3,500	7,500
Not Videos	2,000	500	2,500
Sum (col.)	6,000	4,000	10,000

Interestingness Measure: Lift

Ratio of conditional probability (X → Y) and the marginal probability (Y).

$$\text{lift} = \frac{P(Y|X)}{P(X)} = \frac{P(X,Y)}{P(X)P(Y)}$$

	Games	¬ Games	Sum (row)
Videos	4,000	3,500	7,500
¬ Videos	2,000	500	2,500
Sum (col.)	6,000	4,000	10,000

$$\text{lift(Games,Videos)} = \frac{4000/10000}{6000/10000 * 7500/10000} = 0.89$$

$$\text{lift(Games,}\neg\text{Videos)} = \frac{2000/10000}{6000/10000 * 2500/10000} = 1.33$$

Interestingness Measure: χ^2

A hypothesis test: whether two variables are independent

Observed Count of a cell

Expected Count of a cell

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - \mathbb{E}(n_{ij}))^2}{\mathbb{E}(n_{ij})}$$

	i = 1	i = 0	Sum (row)
j = 1	n_{11}	n_{01}	$n_{11}+n_{01}$
j = 0	n_{10}	n_{00}	$n_{10}+n_{00}$
Sum (col.)	$n_{11}+n_{10}$	$n_{01}+n_{00}$	$n_{11}+n_{01}+n_{10}+n_{00}$

Interestingness Measure: χ^2

A hypothesis test: whether two variables are independent

Observed Count of a cell

Expected Count of a cell

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \mathbb{E}(n_{ij}))^2}{\mathbb{E}(n_{ij})}$$

	i = 1	i = 0	Sum (row)
j = 1	n_{11}	n_{01}	$n_{11}+n_{01}$
j = 0	n_{10}	n_{00}	$n_{10}+n_{00}$
Sum (col.)	$n_{11}+n_{10}$	$n_{01}+n_{00}$	$n_{11}+n_{01}+n_{10}+n_{00}$

$$\text{Expected Count} = \frac{\text{(Row Total)} \cdot \text{(Column Total)}}{\text{(Sample Size)}}$$

Compare this test statistic with the critical value of a given confidence level

Calculating χ^2

	Games	¬ Games	Sum (row)
Videos	4,000 (4,500)	3,500 (3,000)	7,500
¬ Videos	2,000 (1,500)	500 (1,000)	2,500
Sum (col.)	6,000	4,000	10,000

(): expected values

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$= \frac{(4000-4500)^2}{4500} + \frac{(3500-3000)^2}{3000} + \frac{(2000-1500)^2}{1500} + \frac{(500-1000)^2}{1000} = 555.6$$

The variety of interestingness measures

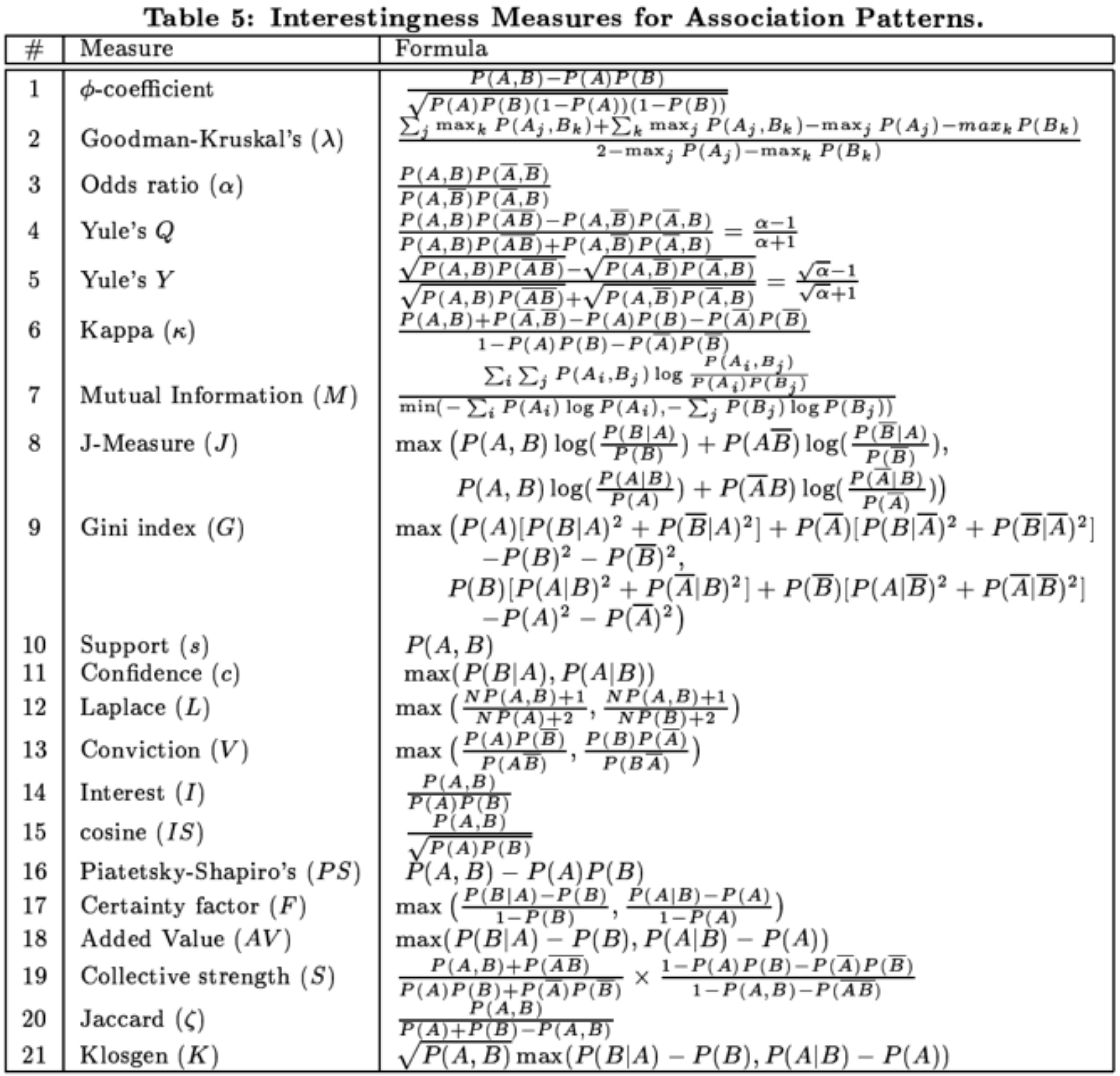


Table 5: Interestingness Measures for Association Patterns.

#	Measure	Formula
1	ϕ-coefficient	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\overline{AB})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{AB})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{AB})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{AB})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i)\log P(A_i),-\sum_j P(B_j)\log P(B_j))}$
8	J-Measure (J)	$\max\big(P(A,B)\log(\frac{P(B\|A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),$ $P(A,B)\log(\frac{P(A\|B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\big)$
9	Gini index (G)	$\max\big(P(A)[P(B\|A)^2 + P(\overline{B}\|A)^2] + P(\overline{A})[P(B\|\overline{A})^2 + P(\overline{B}\|\overline{A})^2]$ $-P(B)^2 - P(\overline{B})^2,$ $P(B)[P(A\|B)^2 + P(\overline{A}\|B)^2] + P(\overline{B})[P(A\|\overline{B})^2 + P(\overline{A}\|\overline{B})^2]$ $-P(A)^2 - P(\overline{A})^2\big)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\|A), P(A\|B))$
12	Laplace (L)	$\max\big(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2}\big)$
13	Conviction (V)	$\max\big(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})}\big)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B) - P(A)P(B)$
17	Certainty factor (F)	$\max\big(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)}\big)$
18	Added Value (AV)	$\max(P(B\|A) - P(B), P(A\|B) - P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})} \times \frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\|A) - P(B), P(A\|B) - P(A))$

Tan, Kumar, Srivastava @KDD'02

Mutual Information

Measures mutual dependence between two random variables (X, Y)

Classical concept in information theory: Amount of information obtained about X through observing Y.

Mutual Information

(Full) mutual information

$$I(X;Y) = \sum_{x \in \mathscr{X}} \sum_{y \in \mathscr{Y}} P(X=x, Y=y) \log \frac{P(X=x, Y=y)}{P(X=x)P(Y=y)}$$

x, y are possible values of X and Y; in the case of appearance of an item, 1 or 0.

Calculating Mutual Information

$$I(Games; Video) = P(G,V)\log\frac{P(G,V)}{P(G)P(V)} + P(G,\neg V)\log\frac{P(G,\neg V)}{P(G)P(\neg V)}$$
$$+P(\neg G,V)\log\frac{P(\neg G,V)}{P(\neg G)P(V)} + P(\neg G,\neg V)\log\frac{P(\neg G,\neg V)}{P(\neg G)P(\neg V)}$$

Using base-2 logarithms:

	Games	¬ Games	Sum
Videos	4,000	3,500	7,500
¬ Videos	2,000	500	2,500
Sum	6,000	4,000	10,000

$$I(Games; Videos) = 0.4 \times \log\frac{0.4}{0.6*0.75}$$
$$+0.2 \times \log\frac{0.2}{0.6*0.25} + 0.35 \times \log\frac{0.35}{0.4*0.75}$$
$$+0.05 \times \log\frac{0.05}{0.4*0.25}$$
$$= 0.04$$

Pointwise Mutual Information

If we only care about one configuration

$$\text{PMI}(X{=}x;Y{=}y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

	Games	¬ Games	Sum
Videos	4,000	3,500	7,500
¬ Videos	2,000	500	2,500
Sum	6,000	4,000	10,000

Pointwise Mutual Information

If we only care about one configuration

$$\text{PMI}(X{=}x;Y{=}y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

	Games	¬ Games	Sum
Videos	4,000	3,500	7,500
¬ Videos	2,000	500	2,500
Sum	6,000	4,000	10,000

$$\text{PMI}(G;V) = \log_2 \frac{0.4}{0.6 * 0.75} = -0.17$$

Pointwise Mutual Information

If we only care about one configuration

$$\text{PMI}(X=x;Y=y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

	Games	¬ Games	Sum
Videos	4,000	3,500	7,500
¬ Videos	2,000	500	2,500
Sum	6,000	4,000	10,000

$$\text{PMI}(G;V) = \log_2 \frac{0.4}{0.6 * 0.75}$$

$$= -0.17$$

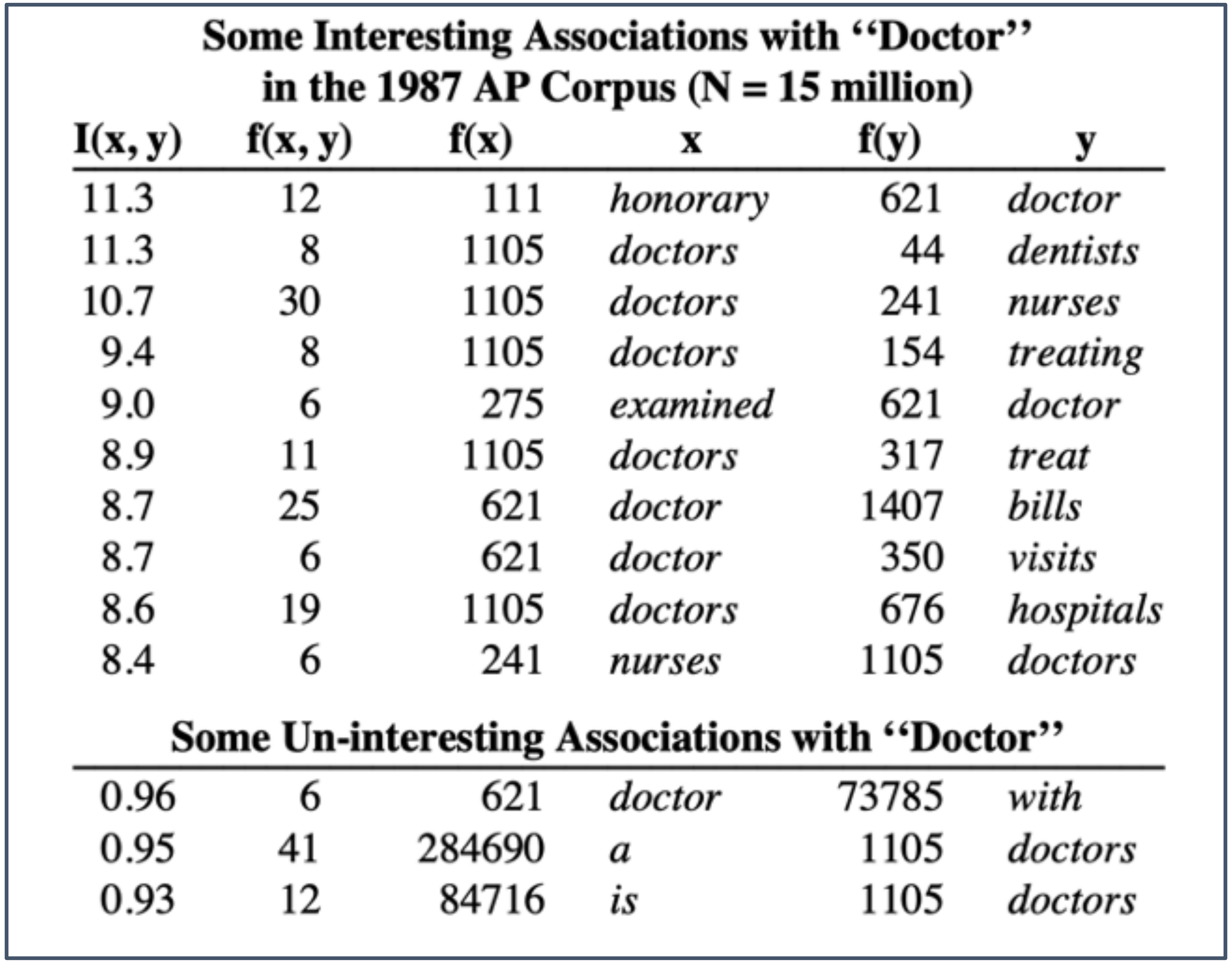
$$\text{PMI}(G;\neg V) = \log_2 \frac{0.2}{0.6 * 0.25}$$

$$= 0.42$$

Application: Word Collocations

- Determine the likelihood that two words will be used together
- Words as items (bag of words)
- Co-occurrences of words indicate semantic relationship

Application: Word Collocations



Some Interesting Associations with "Doctor" in the 1987 AP Corpus (N = 15 million)

I(x, y)	f(x, y)	f(x)	x	f(y)	y
11.3	12	111	*honorary*	621	*doctor*
11.3	8	1105	*doctors*	44	*dentists*
10.7	30	1105	*doctors*	241	*nurses*
9.4	8	1105	*doctors*	154	*treating*
9.0	6	275	*examined*	621	*doctor*
8.9	11	1105	*doctors*	317	*treat*
8.7	25	621	*doctor*	1407	*bills*
8.7	6	621	*doctor*	350	*visits*
8.6	19	1105	*doctors*	676	*hospitals*
8.4	6	241	*nurses*	1105	*doctors*

Some Un-interesting Associations with "Doctor"

I(x, y)	f(x, y)	f(x)	x	f(y)	y
0.96	6	621	*doctor*	73785	*with*
0.95	41	284690	*a*	1105	*doctors*
0.93	12	84716	*is*	1105	*doctors*

(Church and Hanks, 1990) *"Word Association Norms, Mutual Information, and Lexicography", Computational Linguistics.*

More applications in text mining

- Spell check
- Polysemy – one word with multiple meanings
- Disambiguation
- Synonymy – multiple words with the same meaning
- Phrase detection, entity extraction
- Word clustering, concept extraction, topic extraction
- Ontology, taxonomy

Correlation != Causation

Ice cream sales is correlated with rate of drowning deaths, but no real causation between the two!

Causal inference techniques are needed to discover real causality; but frequency and correlation are still the prerequisites.

Itemset Mining: Open Questions

How do we extract frequent itemsets with **efficiency and scalability** in mind?

How do we extract **patterns with certain constraints** (e.g., significant patterns instead of frequent patterns)?

How do we **interpret patterns**?

How do we **evaluate interestingness measures**?

How do we use **itemset patterns for classification**?

Applications! Applications!!

Tools for frequent itemset mining

Weka:
http://sourceforge.net/projects/weka/?source=typ_redirect

- Java package.
- Well maintained/documented; many data mining algorithms implemented.

Mlxtend:
http://rasbt.github.io/mlxtend/

- Python package with limited but useful frequent pattern mining functionalities.

What you should know

- Basic methods of frequent pattern mining
- How Apriori is able to scale up frequent pattern mining
- How to measure the interestingness of patterns
- Applications of itemset mining
- Co-occurrence is the key in text mining

Similarity-based Itemset Mining

What is “similarity”?

Similarity is a measure of how much two data objects are alike

Distance measures the opposite: how much they are dissimilar

How similar are two itemsets?

Are T2 & T3 more similar than T2 & T4?

Intuition:

- Two sets are similar if they share a lot of items
- But larger sets are likely to share more items with others.

TID	Items Bought
T1	🍺 🍼 🍉
T2	🍺 🍭 🍼 🍋
T3	🍼 🍺 🍋
T4	🍼 🍭
T5	🍷 🍪 🍺 🍞 🍋

Intersection and Union

Intersection (**A**∩**B**): largest common subset of A and B

Union (**A**∪**B**): smallest common superset of A and B

Intersection and Union

Intersection (**A**∩**B**): largest common subset of A and B

Union (**A**∪**B**): smallest common superset of A and B

{🍺, 🍼, 🍉} ∩ {🍼, 🍺, 🍋} = {🍺, 🍼}

{🍺, 🍼, 🍉} ∪ {🍼, 🍺, 🍋} = {🍺, 🍼, 🍉, 🍋}

The Jaccard Similarity

A simple (but powerful) measure of similarity of two **sets**

also known as Jaccard coefficient and Jaccard index

Number of items in the intersection of two sets

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Number of items in the union of two sets

Jaccard distance: 1 - J(A, B)

Properties of Jaccard Similarity

- $J(A, B) = J(B, A)$
- $0 \leq J(A, B) \leq 1$
- $J(A, B) = 0$ if two sets share no items
- $J(A, B) = 1$ if two sets are identical

Calculating Jaccard Similarity

TID	Items Bought
T1	🍺 🍼 🍉
T2	🍺 🍭 🍼 🍋
T3	🍼 🍺 🍋
T4	🍼 🍭
T5	🍷 🍪 🍺 🍞 🍋

J(T1,T2)= |{🍺,🍼}|/|{🍋,🍺,🍼,🍉,🍭}|
= 2/5 = 0.4

J(T2,T3)= |{🍺,🍼,🍋}|/|{🍋,🍺,🍼,🍭}|
= 3/4 = 0.75

Calculating Jaccard Similarity

TID	Items Bought
T1	🍺 🍼 🍉
T2	🍺 🍭 🍼 🍋
T3	🍼 🍺 🍋
T4	🍼 🍭
T5	🍷 🍪 🍺 🍞 🍋

J(T1,T2)= |{🍺,🍼}|/|{🍋,🍺,🍼,🍉,🍭}|
= 2/5 = 0.4

J(T2,T3)= |{🍺,🍼,🍋}|/|{🍋,🍺,🍼,🍭}|
= 3/4 = 0.75

Can you calculate J(T2, T4)?

Measure the similarity of items

TID	Items Bought
T1	🍺 🍼 🍉
T2	🍺 🍭 🍼 🍋
T3	🍼 🍺 🍋
T4	🍼 🍭
T5	🍷 🍺 🍋

Transpose →

Item	Transactions
🍺	T1, T2, T3, T5
🍼	T1, T2, T3, T4
🍉	T1
🍭	T2, T4
🍋	T2, T3, T5
🍷	T5

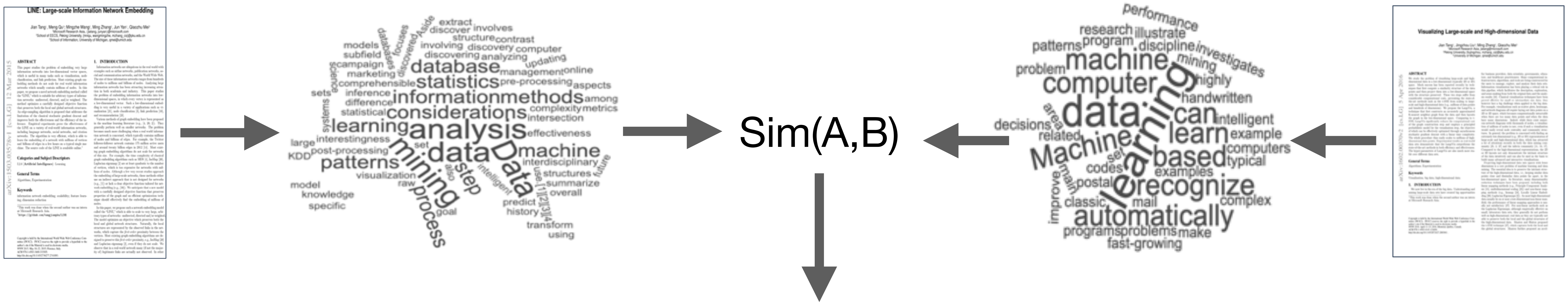
Measure the similarity of items

J(🍺,🍼)=|{T1, T2, T3}|/|{T1,T2,T3,T4,T5}|
= 3/5 = 0.6

J(🍺,🍋)=|{T2, T3, T5}|/|{T1,T2,T3,T5}|
= 3/4 = 0.75

Item	Transactions
🍺	T1, T2, T3, T5
🍼	T1, T2, T3, T4
🍉	T1
🍭	T2, T4
🍋	T2, T3, T5
🍷	T5

Use itemset similarity for complex data mining tasks


LINE: Large-scale Information Network Embedding
extract involves discover structure contrast involving discovery computer discovering analyzing updating models subfield databases focuses discovered aside campaign marketing database management online science comprehensible statistics pre-processing aspects sets inference information methods among difference statistical considerations complexity metrics intersection systems learning analysis effectiveness large interestingness set KDD post-processing data Data machine patterns mining step interdisciplinary visualization also intelligent use structures future raw summarize model overall knowledge specific process goal predict history transform using
performance research illustrate patterns program discipline investigates problem machine mining computer highly handwritten data decisions area can intelligent related learn example Machine learning based computers set codes typical examples postal recognize classic domain mail complex improve automatically programs problems make fast-growing
Visualizing Large-scale and High-dimensional Data

Sim(A,B)

Classification

Clustering

Ranking

Recommendation

Applications of Itemset Similarity

- Similarity of text
  - ✓ Plagiarism detection, Web page deduplication
- Similarity of shopping baskets
  - ✓ Customer profiling/clustering, market segmentation
- Similarity of friend-lists (social network)
  - ✓ Friend recommendation, who to follow, etc.

What you should know

- How to measure the similarity between itemsets
- How to calculate the Jaccard similarity
- Use itemset similarity for complex data mining tasks and real applications

Thank You

Questions?