COMP207
Database Development

Lecture 29

Data Warehousing, OLAP, and Data Mining: Association Rules & A-Priori Algorithm

Revision lecture

- Monday next week (the 9th)
- Review of topics & exam information
- Will also construct a second small mock exam

Frequent-Itemset Mining

- **Market-Basket Data**:
  - Set of items ***I***
  - Set of baskets ***B***

Purchase ID	Items bought
101	milk, bread, cookies, juice
792	milk, juice
1130	milk, bread, eggs
1735	bread, cookies, coffee

- Basic task:

Which sets ***J*** of items are **frequent**?

“Diapers and beer are frequently bought together.”

“Harry Potter 1 and Game of Thrones are frequently liked by the same viewers.”

- ***J*** is **frequent** if its **support** exceeds a pre-specified value (the support threshold).

More Complex Tables

- Sometimes data is more complex:

Purchase ID	Customer ID	Items bought
101	A	milk, bread, cookies, juice
792	B	milk, juice
1130	A	milk, bread, eggs
1735	C	bread, cookies, coffee

- Items are clear (milk, bread, cookies, juice, eggs, …), but *what are the baskets*?
- Solution: consider baskets with respect to a column
  - Exercise sheet for next week will make this more clear

Example (1/4): Customer Baskets

Purchase ID	Customer ID	Items bought
101	A	milk, bread, cookies, juice
792	B	milk, juice
1130	A	milk, bread, eggs
1735	C	bread, cookies, coffee

- **Baskets with respect to customer IDs**:
  - One basket per customer ID
  - Basket for customer ID ***cid***: all items that appear under "Items bought" in rows that have customer ID ***cid***
  - Example: basket for 'A': {milk, bread, cookies, juice, eggs}

Example (2/4): Customer Baskets

- Equivalently: restrict to customer IDs & items bought

~~Purchase ID~~	Customer ID	Items bought
~~101~~	A	milk, bread, cookies, juice
~~792~~	B	milk, juice
~~1130~~	A	milk, bread, eggs
~~1735~~	C	bread, cookies, coffee

- Then combine baskets for customer IDs:

Customer ID	Items bought
A	milk, bread, cookies, juice, eggs
B	milk, juice
C	bread, cookies, coffee

Example (3/4): Customer Baskets

Customer ID	Items bought
A	milk, bread, cookies, juice, eggs
B	milk, juice
C	bread, cookies, coffee

- Define all other notions using new baskets / table:
- **Support of a set *J*** with respect to the customers:
  support of ***J*** computed using the new set of baskets
- ***J* is frequent** with respect to the customers:
  if its support with respect to the customers is at least the support threshold

Example (4/4)

Purchase ID	Customer ID	Items bought
101	A	milk, bread, cookies, juice
792	B	milk, juice
1130	A	milk, bread, eggs
1735	C	bread, cookies, coffee

- Support of {milk, bread}:
  - With respect to the purchases = 2/4 = 0.5
  - With respect to the customers = 1/3 $\approx$ 0.33
- With support threshold s = 0.5
  - {milk, bread} is frequent with respect to the purchases
  - {milk, bread} is not frequent with respect to the customers

Exercise (3 mins)

Transaction ID	Customer ID	Items bought
101	A	X, Y, Z
792	B	W, X
1130	A	W, Z
1735	C	X

- What is the support for {W, X} with respect to **transactions**?
- What is the support for {W, X} with respect to **customers**?

Exercise (3 mins)

Transaction ID	Customer ID	Items bought
101	A	X, Y, Z
792	B	W, X
1130	A	W, Z
1735	C	X

- What is the support for {W, X} with respect to **transactions**? 1/4
- What is the support for {W, X} with respect to **customers**? 2/3

Association Rules

- Variant of frequent-item mining query
  - “Customers who buy diapers frequently also buy beer.”
  - “People watching Harry Potter and Game of Thrones frequently also watch Twin Peaks.”
- General form:

$$\boxed{\{i_1, i_2, \ldots, i_n\} \Rightarrow j}$$

items

- Examples:
  - {diapers} ⇒ beer
  - {Harry Potter, Game of Thrones} ⇒ Twin Peaks

Association Rules: Properties

$$\{i_1, i_2, \ldots, i_n\} \Rightarrow j$$

- **Support**: support of $\{i_1, i_2, \ldots, i_n, j\}$
  - Want high support (not much benefit in exploiting the rule otherwise)
- **Confidence**: percentage of baskets for $\{i_1, i_2, \ldots, i_n\}$ containing j

$$\frac{\text{support of } \{i_1, \ldots, i_n, j\}}{\text{support of } \{i_1, \ldots, i_n\}}$$

  - Should be high
  - ”67% of all customers who bought milk also bought juice.”
  - Should differ significantly from fraction of baskets containing *j*

Example

{milk} ⇒ juice

Purchase ID	Items bought
101	milk, bread, cookies, juice
792	milk, juice
1130	milk, bread, eggs
1735	bread, cookies, coffee

- **Support** = support of {milk, juice} = $\frac{2}{4}$ = 0.5
  - So, the rule applies in 50% of all the cases

- **Confidence** = $\dfrac{\text{support of \{milk, juice\}}}{\text{support of \{milk\}}} = \dfrac{2/4}{3/4} = \dfrac{2}{3} \approx 0.67$
  - "67% of all customers who bought milk also bought juice."
  - Differs significantly from % of baskets containing juice (0.5)

Exercise 2 (5 min)

Viewer	Liked videos
Anna	BI, BS, BU
Ben	HP1, HP2, HP3
Chloe	BI, HP1
Dave	BI, BS, HP1, HP2
Emma	BI, BS, BU, HP1
Fred	BI, BU
Gwen	BS, HP1, HP2
Henry	BI, BS, HP1, HP2, HP3

- What is the confidence for {BS} ⇒ HP1?
- Are there movies X, Y such that {X} ⇒ Y with confidence $> \frac{2}{3}$
- Are there any movies X, Y, Z so that X and Y has support at least $\frac{2}{8} = \frac{1}{4}$ and such that {X,Y} ⇒ Z have confidence $>\frac{3}{4}$?

Solution

Viewer	Liked videos
Anna	BI, BS, BU
Ben	HP1, HP2, HP3
Chloe	BI, HP1
Dave	BI, BS, HP1, HP2
Emma	BI, BS, BU, HP1
Fred	BI, BU
Gwen	BS, HP1, HP2
Henry	BI, BS, HP1, HP2, HP3

- Confidence for {BS} ⇒ HP1: $\frac{4}{5}$
- {BS} ⇒ BI ($\frac{4}{5}$), {BS} ⇒ HP1 ($\frac{4}{5}$), {BU} ⇒ BI (1), {HP2} ⇒ BS ($\frac{3}{4}$), {HP2} ⇒ HP1 (1), {HP3} ⇒ HP1 (1) {HP3} ⇒ HP2 (1)

Solution 2

Viewer	Liked videos
Anna	BI, BS, BU
Ben	HP1, HP2, HP3
Chloe	BI, HP1
Dave	BI, BS, HP1, HP2
Emma	BI, BS, BU, HP1
Fred	BI, BU
Gwen	BS, HP1, HP2
Henry	BI, BS, HP1, HP2, HP3

- {BI,HP2} ($\frac{1}{4}$) ⇒ BS (1), {BI,HP2} ($\frac{1}{4}$) ⇒ HP1 (1), {BS,BU} ($\frac{1}{4}$) ⇒ BI (1), {BS,HP2} ($\frac{3}{8}$) ⇒ HP1 (1), {HP1,HP3} ($\frac{1}{4}$) ⇒ HP2 (1), {HP2,HP3} ($\frac{1}{4}$) ⇒ HP1 (1)

Finding Association Rules

- Focus on association rules with high support (e.g., at least support threshold ***s***)
- Compute the set ***F*** of all itemsets ***J*** with support $\geq$ ***s***
- From ***F*** it is easy to obtain the association rules:
  - For each set ***J*** in ***F*** and each *j* in ***J***, consider rule ***J*** \ {*j*} ⇒ *j*
  - Compute confidence for ***J*** \ {*j*} ⇒ *j* and compare with % of baskets containing *j*
- Remains: Finding frequent itemsets

A-Priori Algorithm

- Goal: compute all itemsets $\boldsymbol{J}$ with support $\geq \boldsymbol{s}$
- Crux:

If $\boldsymbol{J}$ has support $\geq \boldsymbol{s}$, then all subsets of $\boldsymbol{J}$ have support $\geq \boldsymbol{s}$.

If some subset of $\boldsymbol{J}$ has support < $\boldsymbol{s}$, then $\boldsymbol{J}$ has support < $\boldsymbol{s}$.

- Compute frequent item sets from smaller ones:
  - F_1 := item sets $\{i\}$ with support $\geq$ **s**
  - F_2 := item sets $\{i,j\}$ with $\{i\}$, $\{j\} \in F_1$ and support $\geq$ **s**
  - F_3 := item sets $\{i,j,k\}$ with $\{i,j\},\{i,k\},\{j,k\} \in F_2$ and support $\geq$ **s**
  - …

A-Priori Algorithm

- Input: set of items ***I***, set of baskets ***B***, max. size ***q***, support threshold ***s***
- Output: subsets ***J*** of ***I*** with $|\boldsymbol{J}| \leq \boldsymbol{q}$ and support $\geq$ ***s***

- Algorithm:

```
– $C_1 := \{ \{i\} : i \in \boldsymbol{I} \}$
– for $k = 1$ to $\boldsymbol{q}$ do
    • $F_k := \{ \boldsymbol{J} \in C_k : \boldsymbol{J}$ has support $\geq \boldsymbol{s} \}$
    • if $k = \boldsymbol{q}$ then stop
    • $C_{k+1} := \{ \boldsymbol{J} \subseteq \boldsymbol{I} : |\boldsymbol{J}| = k+1$ and all subsets of $\boldsymbol{J}$ of size $k$ occur in $C_k\}$
```

- Efficient implementation in SQL for fixed ***q*** (e.g., ***q***=2)

Wrapping Up…

Data Mining

- Refers to discovery of patterns/knowledge in data
  - "50% of people who buy hot dogs also buy mustard."
  - "These three individual's pattern of bank transactions indicate that they are running a terrorist cell."
- Combines many different areas of computer science and mathematics
  - Machine Learning: bit in COMP219, might see methods in other modules
  - Statistics
  - Data management
- Here: illustration of some basic methods

Many Applications

- **Deviation Detection**
  - Identify anomalies (e.g., intruders trying to break into a system)
- **Link Analysis**
  - Try to discover links between attributes (e.g., association rules)
- **Predictive Modelling (“Prediction”)**
  - Try to predict future behaviour of certain attributes in the data based on past behaviour
- **Database Segmentation**
  - Group data by similar “behaviour” (e.g., group customers based on spending habits, reaction to a marketing campaign, etc.)

Types of Discovered Knowledge

- **Association rules**
- **Classification hierarchies**
  - Example: Classification of mutual funds based on performance data characteristics such as growth, income, stability, …
- **Sequential patterns**
  - Example: “If a patient underwent cardiac bypass surgery for blocked arteries and an aneurysm and later developed high blood urea within a year of surgery, he/she is likely to suffer from kidney failure within the next 18 months.”
- **Clustering**
  - Example: Group treatment data on a disease based on similarity of side effects.
- ...

The Final Slide

- **Databases**: exciting field of computer science
  - Especially with emergence of applications that require efficient access to large data sets
  - Intersects with many other fields: algorithms, machine learning, distributed systems, engineering, …
- Here: foundations
- Where to go from here?
  - Read more on data management (textbooks, web, …)
  - Experiment with systems (relational DBMS, NoSQL stores, …), use them in a software project, …
  - Contribute to open source projects (e.g., in the NoSQL area) or do your own project

Thank you!