COMP207
Database Development

Lecture 28

Data Warehousing, OLAP, and Data Mining: Data Mining

From OLAP to Data Mining

- Previous lecture: **OLAP (Online Analytic Processing)**
  - What is OLAP?
  - Opposed to **OLTP (Online Transaction Processing)**
- This lecture: **Data Mining**
  - Can be seen as extended form of OLAP
  - *“Find factors that have had the most influences over sales of product X?”* rather than trying out queries like

```
SELECT model, SUM(price)
FROM Sales NATURAL JOIN Products
WHERE type=‘X’
GROUP BY model;
```

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

Data Mining – Real life examples

1. US phone company
   - Found 10,000 residential customers that paid above 1000$ a month
   - They were businesses that tried to use cheaper residential rates
2. Supermarkets (Target, US)
   - Managed to send promotional offers for diapers before some of the families had announced they were pregnant
   - (Outcome: partially hides this nowadays by adding “wrong” offers)
3. Crime prevention
   - Where to deploy police, who to search at borders, what intelligence to take seriously and credit card fraud detection
4. Personalities
   - Facebook likes + data mining is better at judging peoples personality than humans
   - See “**Computer-based personality judgments are more accurate than those made by humans**“ by Wu Youyou et. al
5. Email spam detection
   - e.g. in GMail
6. Recommendations
   - Related items in e.g. Netflix, Amazon or others
   - People you may know in Facebook

Ethics and how to use

2. Supermarkets (Target, US)
   - Managed to send promotional offers for diapers before some of the families had announced they were pregnant
   - (Outcome: partially hides this nowadays by adding “wrong” offers)

- Many other dubious examples
- This module does not cover:
  - Ethics
  - How to use the information

Scenario 1: Supermarket Checkout

- Imagine a supermarket…

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

- Use for decision making: “Which items are frequently bought together?”
  - Can influence item placement, decisions on prices, etc.
  - Can uncover interesting relationships, e.g.: diapers-beer

Scenario 2: You Might Also Like…

- Imagine a film streaming service…

Films	Liked by…
Bourne Identity	Anna, Chloe, Dave, Emma, Fred, Henry
Bourne Supremacy	Anna, Dave, Emma, Gwen, Henry
Bourne Ultimatum	Anna, Emma, Fred
Harry Potter 1	Ben, Chloe, Dave, Emma, Gwen, Henry
Harry Potter 2	Ben, Dave, Gwen, Henry
Harry Potter 3	Ben, Henry

- Possible question: “Which viewers frequently like the same films?”
  - Could be used for recommending films

Market-Basket Data

- Data that can be described by:
  - A set of **items *I***
  - A set of **baskets *B***: each basket ***b*** ∈ ***B*** is a subset of ***I***
- Example:

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

b_1 = {milk, bread, cookies, juice}

b_2 = {milk, juice}

b_3 = {milk, bread, eggs}

b_4 = {bread, cookies, coffee}

I = {bread, coffee, cookies, eggs, juice, milk}

$B = \{b_1, b_2, b_3, b_4\}$

Another Example

Films	Liked by…
Bourne Identity	Anna, Chloe, Dave, Emma, Fred, Henry
Bourne Supremacy	Anna, Dave, Emma, Gwen, Henry
Bourne Ultimatum	Anna, Emma, Fred
Harry Potter 1	Ben, Chloe, Dave, Emma, Gwen, Henry
Harry Potter 2	Ben, Dave, Gwen, Henry
Harry Potter 3	Ben, Henry

- I = {Anna, Ben, Chloe, Dave, Emma, Fred, Gwen, Henry}
- $B = \{b_1, b_2, b_3, b_4, b_5, b_6\}$
  - b_1 = {Anna, Chloe, Dave, Emma, Fred, Henry}
  - b_2 = {Anna, Dave, Emma, Gwen, Henry}
  - …

Frequent-Itemset Mining

- Given:
  - 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 problem:

Which items occur frequently together in a basket?

“Diapers and beer are frequently bought together.”

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

- How frequent is “frequently”?

Support of an Itemset

- Given: set of items ***I*** and set of baskets ***B***

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

Support of {milk, juice}:
$\frac{2}{4}$ = 0.5

Support of {bread, juice}:
$\frac{1}{4}$ = 0.25

- **Support** of a subset ***J*** of ***I*** = frequency with which the items in ***J*** occur together in a basket in ***B***

$$\frac{\text{number of baskets in } \boldsymbol{B} \text{ containing all items in } \boldsymbol{J}}{\text{number of baskets in } \boldsymbol{B}}$$

Frequent Itemsets

- Given: set of items ***I*** and set of baskets ***B***

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

Support of {milk, juice}: $\frac{2}{4} = 0.5$

Support of {bread, juice}: $\frac{1}{4} = 0.25$

- A subset ***J*** of ***I*** is **frequent** if its support is at least ***s***.
  - ***s***: **support threshold** (specified by user)
- Example (s = 0.5)
  - {milk, juice} is frequent
  - {bread, juice} is not frequent

Exercise (5 min)

Viewer	Liked videos
Anna	BI, BS, BU
Ben	HP1, HP2, HP3
Chloe	BI, HP1
Dave	BI, BS, HP1, H2
Emma	BI, BS, BU, HP1
Fred	BI, BU
Gwen	BS, HP1, HP2
Henry	BI, BS, HP1, HP2, HP3

- What is the support of {BI, BS}?
- Is it frequent if the support threshold is s = 3/8?
- Can you find other frequent sets of two items?

Solution

Viewer	Liked videos
Anna	BI, BS, BU
Ben	HP1, HP2, HP3
Chloe	BI, HP1
Dave	BI, BS, HP1, H2
Emma	BI, BS, BU, HP1
Fred	BI, BU
Gwen	BS, HP1, HP2
Henry	BI, BS, HP1, HP2, HP3

- Support of {BI, BS} = $\frac{4}{8}$ = 0.5
- All frequent pairs of two items with support threshold s = 3/8: {BI, BS}, {BI, HP1}, {BS, HP1}, {HP1, HP2}

End of Module Questionnaires / Survey