ASSOCIATION RULE MINING

Outline

- Introduction
- Big data techniques
- Association rule (Implementation)
- Big data technologies
- Future Work
- References

Introduction

What is Big

- No single standard definition…

- It represents massive data sets with large, more varied and complex structure with challenge of storing, analyzing and visualizing for extracting meaningful results [1].

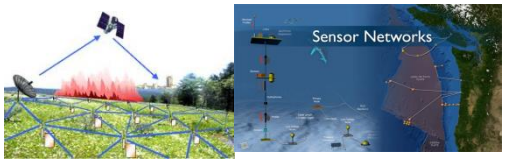
Factors generating bi a

- Medical records
- Scientific research
- Government
- Natural disaster and resource management
- Mobile phone
- Private sector
- Military surveillance
- Financial services
- Retail
- Social networks
- Web logs, text, document, photography, audio, video.
- Search indexing
- Call detail records
- Sensor networks and telecommunications

Scientific instrument

Mobile devices

Social media and networks



Sensor technology and networks

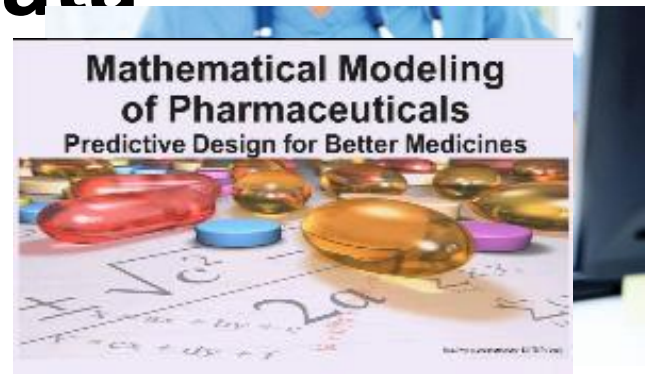
Motivation and benefits

- Exponential growth of digital world [2]

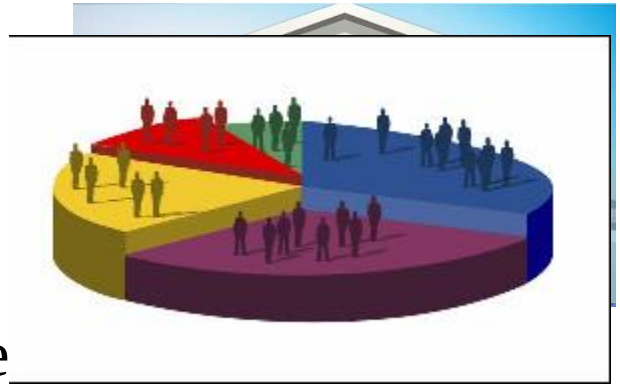
- Better aimed marketing
- Client based segmentation
- Automated decision making
- Greater return on investments
- Quantification of risks and market trending
- Better planning and forecasting
- Identification of consumer behaviour and production yield extension
- Predictive analytics on traffic flows
- Identification of threats from different video, audio and data feeds.

Potential of big data

- Health care [3]
  - Clinical data
  - Pharmaceutical R&D data
  - Activity (claims) and cost data



- Public sector [3]
  - Creating transparency
  - Population segmentation
  - Automatic decision making
  - Innovation of new product and service

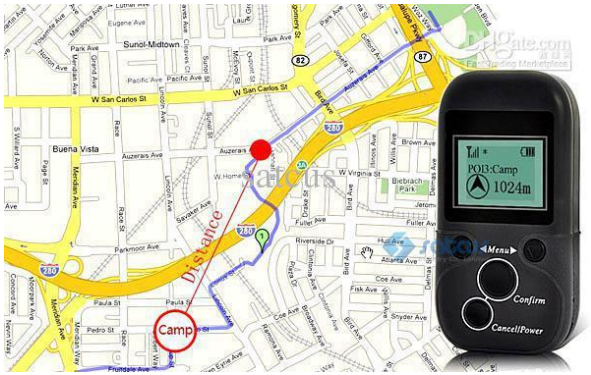


- Retail [3]
  - Marketing
  - Merchandising
  - Operations
  - Supply chain

Potential of big data

- Manufacturing [3]
  - Research and development and produ
  - Product lifecycle management.
  - Design to value.
  - Open innovation

- Personal location data [3]
  - Smart routing
  - Geo targeted advertising
  - Emergency response
  - Urban planning



- Social network analysis [3]
  - More targeted advertising
  - Marketing campaigns and capaci
  - Customer behavior and buying pa
  - Sentiment analytics

Association rule learning

Association rule in big data

- Association rules are the form A→B. A→B is different from B→A.
- This implies that if a customer purchase item A then he also purchase item B.

- For support level that generate less than **100,000 rules**, Apriori finishes on all datasets in less than **1 minute**.

- For support level that generate less than **1,000,000 rules**, which are sufficient for prediction purposes Apriori finishes processing in less than **10 minutes**. [14]

Real life application

Field of work	Problem	Method applied	Outcome
Government sector. Researchers of King's College London [19]	Fraud at Consignia, UK's Post office group	Use of "if…then" association rule. E.g. Normal behavior rule "IF time < 1200 AND item = stamps THEN $2 < cost < $4."	Detectors that successfully spot abnormal transactions. They also copy themselves, so CIFD adapts itself to create detectors that correspond to the most prevalent patterns of fraud.
[20]	Issues concerning accessibility of an urban area	Spatial association rule mining to geo-referenced U.K. census data of 1991	Helped in transportation planning in area near a local Stepping Hill Hospital.
Health care sector. [21]	Anomaly detection and classification in Breast Cancer.	In training Apriori algorithm was applied and association rules were extracted. The support was set to 10% and the confidence to 0%.	Success rate of classifier was 69.11%. Time required for training was much less then neural network.

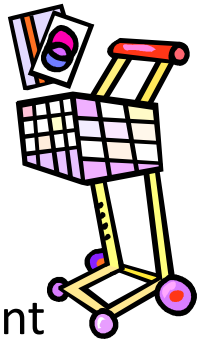
Real life application

Field of work	Problem	Method applied	Outcome
Retail Sector. [22]	Purchasing behavior of customer	On a dataset of 353,421 records from 1903 households about 1,022,812 association rules were generated for promotion sensitivity analysis i.e., analysis of customer responses to various types of promotions, including advertisements, coupons, and various types of discounts.	In a time duration of 1.5 hours about 2.6% of discovered rules were accepted and rest were rejected. Thus total rules reduced to about 14 rules per household from 537 rules per household.
Telecom Sector. [23]	Which country pairs or triples or quadruples customers are currently calling	Use of association rule by treating the top-k country item set as a market basket for each of account. Exploiting temporal nature of data by using traffic from last month as a baseline for current month.	Successful in detecting high rate of fraud calls trends associated with adult entertainment services, that move from country to country through time.

Real life application

Field of work	Problem Statement	Method applied	Outcome
Manufacturing sector. VAM Drilling industries France [24]	Setting up a system which provides result identical to human observation related to performance and dysfunctions during forging.	Use of Rule-Growth that mines sequential rules by FP-growth with varying the parameters minimum support and minimum confidence	Found the main dysfunction responsible for delay. Finding that generator is cause for exceeding maximum time in starting phase The third major problem was the lack of effectiveness of metal strippers

Market Basket Analysis



- In Retail each customer purchases different set of products, different quantities, different times
- Retailers uses this information to:
  - Gain insight about its merchandise (products):
    - Fast and slow movers
    - Products which are purchased together
    - Products which might benefit from promotion
  - Take action:
    - Store layouts
    - Which products to put on specials, promote, coupons…
- Combining all of this with a customer loyalty card it becomes even more valuable

DATASET [18]

S.No.	Item 1	Item 2	Item 3
1.	Bread	Butter	Milk
2.	Ice-cream	Bread	Butter
3.	Bread	Butter	Noodles
4.	Bread	Noodles	Ice-cream
5.	Butter	Milk	Bread
6.	Bread	Noodles	Ice-cream
7.	Milk	Butter	Bread
8.	Ice-cream	Milk	Bread
9.	Butter	Milk	Noodles
10.	Noodles	Butter	Ice-cream

S.No.	Item 1	Item 2	Item 3
1.	Bread	Butter	Milk
2.	Ice-cream	Bread	Butter
3.	Bread	Butter	Noodles
4.	Bread	Noodles	Ice-cream
5.	Butter	Milk	Bread
6.	Bread	Noodles	Ice-cream
7.	Milk	Butter	Bread
8.	Ice-cream	Milk	Bread
9.	Butter	Milk	Noodles
10.	Noodles	Butter	Ice-cream

- The support for ten transactions where bread and noodles occur together is three. Support for {Bread, Noodles} = 3/10= 0.30.

- This means the association of data set or item set, the bread and noodles brought together with 30 percent support.

S.No.	Item 1	Item 2	Item 3
1.	Bread	Butter	Milk
2.	Ice-cream	Bread	Butter
3.	Bread	Butter	Noodles
4.	Bread	Noodles	Ice-cream
5.	Butter	Milk	Bread
6.	Bread	Noodles	Ice-cream
7.	Milk	Butter	Bread
8.	Ice-cream	Milk	Bread
9.	Butter	Milk	Noodles
10.	Noodles	Butter	Ice-cream

$$\text{Confidence}\,(A \rightarrow B) = \frac{\text{Number of tuples containing both A and B}}{\text{Number of tuples containing A}}$$

- Confidence for Bread → Noodles = 3/8 = 0.375

- This means that a customer who buy bread then there is a confidence of 37.5 percent that it also buy noodles.

APRIORI ALGORITHM

- Apriori is an algorithm for finding frequent item-sets using candidate generation. [18]

- Given minimum required support ***'S'*** as interestingness criterion: -

  (1) Search for all individual elements (1-element item-set) that have a minimum support of 'S'.

  (2) From the results of the previous search for 'i' element item-set, search for all 'i+ 1' element item-sets that have a minimum support of 'S'. This becomes the set of all frequent '(i+ 1)' item-sets that are interesting.

  (3) Repeat step 2 until item-set size reaches maximum.

EXPLANATION

- In the given dataset every item occurs three or more than three times and total number of transaction is ten so,

  Minimum Support = 0.3

- Interestingness of 1- element item-sets: - {Bread}, {Butter}, {milk}, {ice-cream}, {noodles}

Item-set	Support
Bread	0.8
Butter	0.7
Noodles	0.5
Ice-cream	0.5
Milk	0.5

EXPLANATION

- Interestingness 2-element item-sets
- {Bread, Butter}, {Bread, Milk}, {Bread, Noodles}, {Bread, Ice-cream}, {Butter, Milk}, {Butter, Noodles}, {Noodles, ice-cream }, etc.

Item-sets	Support
{Bread, Butter}	0.5
{Bread, Milk}	0.4
{Bread, Noodles}	0.3
{Bread, Ice-cream}	0.4
{Butter, Milk}	0.4
{Butter, Noodles}	0.3
{Butter, Ice-cream}	0.2
{Noodles, Milk}	0.1
{Noodles, Ice-cream }	0.3
{Milk, Ice-cream}	0.1

S.No.	Item 1	Item 2	Item 3
1.	Bread	Butter	Milk
2.	Ice-cream	Bread	Butter
3.	Bread	Butter	Noodles
4.	Bread	Noodles	Ice-cream
5.	Butter	Milk	Bread
6.	Bread	Noodles	Ice-cream
7.	Milk	Butter	Bread
8.	Ice-cream	Milk	Bread
9.	Butter	Milk	Noodles
10.	Noodles	Butter	Ice-cream

EXPLANATION

- Interestingness 3-element item-sets.

Item-set	Support
{Bread, Butter, Milk}	0.3
{Bread, Ice-cream, Noodles}	0.2
{Bread, Butter, Noodles}	0.1

- The main advantage of the Apriori algorithm is that it only takes data from previous iteration not from the whole data.

MINING ASSOCIATION RULES

RULES: - [18]

(1) Use Apriori to generate item-sets of different sizes.

(2) At each iteration divide each frequent item-set X into two parts antecedent (LHS) and consequent (RHS) this represents a rule of the form LHS→RHS.

(3) Discard all rules whose confidence is less than minimum confidence

S.No.	Item 1	Item 2	Item 3
1.	Bread	Butter	Milk
2.	Ice-cream	Bread	Butter
3.	Bread	Butter	Noodles
4.	Bread	Noodles	Ice-cream
5.	Butter	Milk	Bread
6.	Bread	Noodles	Ice-cream
7.	Milk	Butter	Bread
8.	Ice-cream	Milk	Bread
9.	Butter	Milk	Noodles
10.	Noodles	Butter	Ice-cream

RULE	CONFIDENCE(Percentage)
{Bread} → {Butter, Milk}	37
{Butter} → {Bread, Milk}	42
{Milk} → {Bread, Butter}	60
{Bread, Butter} → {Milk}	60
{Bread, Milk} → {Butter}	75
{Butter, Milk} → {Bread}	75

Final outcome

- If the minimum confidence threshold is 70 percentage, and the minimum support is 30 percentage, then discovered rules are

  {Bread, Milk} → {Butter}

  {Butter, Milk}→ {Bread}

EXPERIMENT

- Bakery Dataset
- On a database with number of items = 50
- Total number of receipt = 75000
- Minimum Support = 0.04