Data Mining II / Adv. Topics in Data Science

Association Rules

Rita P. Ribeiro

2023/2024

Summary

Association Rules in Action

Association Rules: a New Data Mining Task

Data Mining Tasks:

- Prediction
  - Classification
  - Regression
  - ...
- Description
  - Clustering
  - Association Rules
    - find relationships / associations between groups of variables
  - ...

Motivation

Originally developed in the context of Market Basket Analysis

- Data consists of set of items bought by costumers, referred as transactions
- Find unexpected associations between sets of items using the frequency of sets of items
- Discovered sets of items are referred as frequent itemsets or frequent patterns
- Goals
  - Store layout - *Should products A and B be placed together?*
  - Promotions - *If the client is interested in {A,B,C,...}, can we guess other interests?*
  - ...

Actionable Knowledge: store layout

- Possible actions from rule $\{A1, A4\} \rightarrow \{A6\}$
  - Sell $A1$, $A4$, $A6$ together (pack)
  - Place article $A6$ next to articles $A1$, $A4$
  - Offer a discount coupon for $A6$ in articles $A1$, $A4$
  - Place a competitor of A6 next to $A1$, $A4$ (brand protection).
- Note
  - These actions must make sense from the business point of view.

Actionable Knowledge: cross selling

- Steps
  - Client puts article A in basket
  - Shop knows rule $A \rightarrow B$
  - Rule has enough confidence (> 20%)
  - Shop tells client he may be interested in B
  - Client decides whether to buy B or not
- Notes
  - Rules are discovered from business records
  - Discovery (mining) can be made off-line
  - Use of rules can be made on-line

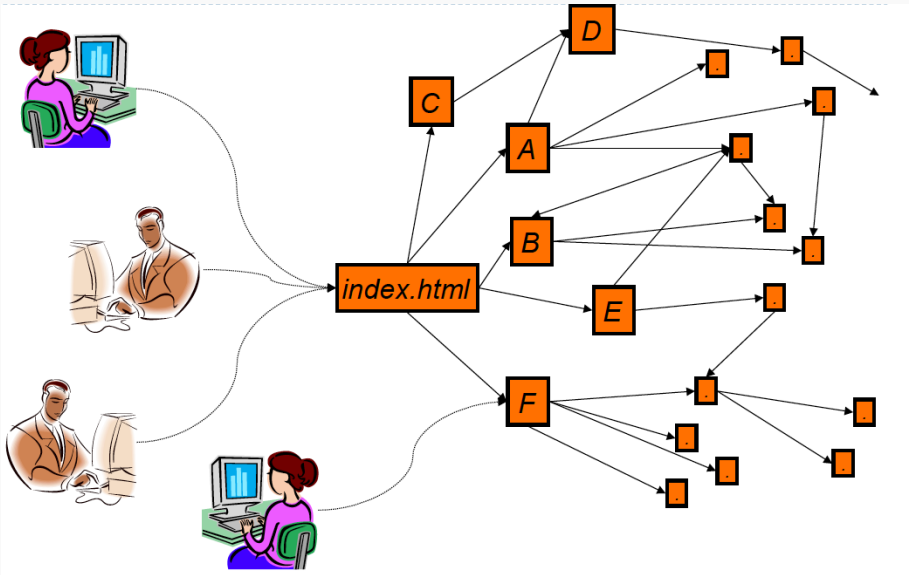
Actionable Knowledge: text mining

- Each document is treated as a “bag” of terms and keywords
  - doc1: Student, Teach, School (Education)
  - doc2: Student, School (Education)
  - doc3: Teach, School, City, Game (Education)
  - doc4: Baseball, Basketball (Sport)
  - doc5: Basketball, Player, Spectator (Sport)
  - doc6: Baseball, Coach, Game, Team (Sport)
  - doc7: Basketball, Team, City, Game (Sport)
- Goal: identify co-occurring terms and keywords
- Example:
  - Student, School $\rightarrow$ Education
  - Game $\rightarrow$ Sport

Actionable Knowledge: health

- Each patient visits a health unit one or more times
- We record the observations for each visit
  - Symptoms (headache, temperature)
  - Exam results (blood pressure, sugar level)
- A set of observations may fire a rule
  {Headache, blood pressure rise} $\rightarrow$ {stroke, immobilisation}
- Early prevention
- Rules obtained from the patient's records

Actionable Knowledge: web usage analysis



Actionable Knowledge: web usage analysis (cont.)

Usage patterns

- Most visited pages
- Frequent page sets
  - Site structure
- Pages associated to users
  - personalization
- Seasonal effects
  - operations, campaigns
- Cross-preferences
  - cross-selling

Actionable Knowledge: web usage analysis (cont.)

From Web Access Logs to Association Rules

IP	date	time	url
194.65.227.7	30-12-1997	0:00:02	/verdemo/tema16/tema16.HTM
194.65.227.7	30-12-1997	0:00:02	/verdemo/gifs/infoline.gif
194.65.227.7	30-12-1997	0:00:12	/verdemo/tema16/sb1601/info1601.htm
194.65.227.7	30-12-1997	0:00:13	/verdemo/tema16/tema16.htm
194.65.227.7	30-12-1997	0:00:13	/verdemo/tema16/sb1601/sub1601.htm
194.65.255.18	30-12-1997	0:00:13	/si/apresent/apresent.html
194.65.227.7	30-12-1997	0:00:15	/verdemo/gifs/back3.gif
194.65.255.18	30-12-1997	0:00:15	/si/gifs/bg6.GIF
194.65.255.18	30-12-1997	0:00:17	/si/gifs/botapr.GIF
194.65.255.18	30-12-1997	0:00:17	/si/gifs/bola.GIF
194.65.255.18	30-12-1997	0:00:18	/si/gifs/barr1-2.GIF

→ ? →

Regras	Suporte	Confiança
diversos & economia-e-finanças & população-condições-sociais -> estatísticas-gerais	0,06	0,97
diversos & economia-e-finanças -> estatísticas-gerais	0,09	0,85
comércio-turismo-serviços & diversos & população-condições-sociais -> estatísticas-gerais	0,05	0,84
comércio-turismo-serviços & estatísticas-gerais & população-condições-sociais -> diversos	0,05	0,84
território-e-ambiente & diversos -> estatísticas-gerais	0,06	0,83
comércio-turismo-serviços & diversos & estatísticas-gerais -> população-condições-sociais	0,05	0,82
indústria-construção-energia & estatísticas-gerais -> diversos	0,06	0,77
indústria-construção-energia & economia-e-finanças -> estatísticas-gerais	0,06	0,77
economia-e-finanças & estatísticas-gerais & população-condições-sociais -> diversos	0,06	0,77
indústria-construção-energia & população-condições-sociais -> diversos	0,05	0,76

web access logs → user sessions → page “baskets” → association rules

Actionable Knowledge: web usage analysis (cont.)

Web Access Logs

IP	date	time	url
194.65.227.7	30-12-1997	0:00:02	/verdemo/tema16/tema16.HTM
194.65.227.7	30-12-1997	0:00:02	/verdemo/gifs/infoline.gif
194.65.227.7	30-12-1997	0:00:12	/verdemo/tema16/sb1601/info1601.htm
194.65.227.7	30-12-1997	0:00:13	/verdemo/tema16/tema16.htm
194.65.227.7	30-12-1997	0:00:13	/verdemo/tema16/sb1601/sub1601.htm
194.65.255.18	30-12-1997	0:00:13	/si/apresent/apresent.html
194.65.227.7	30-12-1997	0:00:15	/verdemo/gifs/back3.gif
194.65.255.18	30-12-1997	0:00:15	/si/gifs/bg6.GIF
194.65.255.18	30-12-1997	0:00:17	/si/gifs/botapr.GIF
194.65.255.18	30-12-1997	0:00:17	/si/gifs/bola.GIF
194.65.255.18	30-12-1997	0:00:18	/si/gifs/barr1-2.GIF

Taxonomy of pages

Theme
↓
Sub-theme
↓
Topic
↓
URL

Actionable Knowledge: web usage analysis (cont.)

Sessions / users

IP	date	time	url
194.65.227.7	30-12-1997	0:00:02	/verdemo/tema16/tema16.HTM
194.65.227.7	30-12-1997	0:00:02	/verdemo/gifs/infoline.gif
194.65.227.7	30-12-1997	0:00:12	/verdemo/tema16/sb1601/info1601.htm
194.65.227.7	30-12-1997	0:00:13	/verdemo/tema16/tema16.htm
194.65.227.7	30-12-1997	0:00:13	/verdemo/tema16/sb1601/sub1601.htm
194.65.255.18	30-12-1997	0:00:13	/si/apresent/apresent.html
194.65.227.7	30-12-1997	0:00:15	/verdemo/gifs/back3.gif
194.65.255.18	30-12-1997	0:00:15	/si/gifs/bg6.GIF
194.65.255.18	30-12-1997	0:00:17	/si/gifs/botapr.GIF
194.65.255.18	30-12-1997	0:00:17	/si/gifs/bola.GIF
194.65.255.18	30-12-1997	0:00:18	/si/gifs/barr1-2.GIF

Actionable Knowledge: web usage analysis (cont.)

Processed data (user_id and theme)

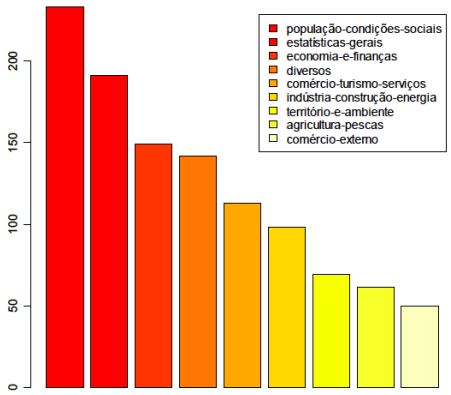
cesto

USER ID	TEMA
acporto	comércio-externo
acporto	comércio-turismo-serviços
agine181	estatísticas-gerais
alggp0157	estatísticas-gerais
alggp0218	economia-e-finanças
aline003	estatísticas-gerais
aline003	território-e-ambiente
aline003	população-condições-sociais
aline003	comércio-turismo-serviços
aline024	comércio-turismo-serviços
aline025	economia-e-finanças
aline025	diversos
aline029	estatísticas-gerais
aline029	economia-e-finanças
aline029	comércio-turismo-serviços
aline032	população-condições-sociais
aline043	economia-e-finanças
aline043	comércio-turismo-serviços
aline065	população-condições-sociais
aline086	agricultura-pescas
...	

Actionable Knowledge: web usage analysis (cont.)

Frequency of visited pages (by theme)

	Frequência
população-condições-sociais	233
estatísticas-gerais	191
economia-e-finanças	149
diversos	142
comércio-turismo-serviços	113
indústria-construção-energia	98
território-e-ambiente	69
agricultura-pescas	61
comércio-externo	50



Actionable Knowledge: web usage analysis (cont.)

Derived association rules

Regras	Suporte	Confiança
diversos & economia-e-finanças & população-condições-sociais -> estatísticas-gerais	0,06	0,97
diversos & economia-e-finanças -> estatísticas-gerais	0,09	0,85
comércio-turismo-serviços & diversos & população-condições-sociais -> estatísticas-gerais	0,05	0,84
comércio-turismo-serviços & estatísticas-gerais & população-condições-sociais -> diversos	0,05	0,84
território-e-ambiente & diversos -> estatísticas-gerais	0,06	0,83
comércio-turismo-serviços & diversos & estatísticas-gerais -> população-condições-sociais	0,05	0,82
indústria-construção-energia & estatísticas-gerais -> diversos	0,06	0,77
indústria-construção-energia & economia-e-finanças -> estatísticas-gerais	0,06	0,77
economia-e-finanças & estatísticas-gerais & população-condições-sociais -> diversos	0,06	0,77
indústria-construção-energia & população-condições-sociais -> diversos	0,05	0,76

- diverse & economy-and-finance $\rightarrow$ general-statistics (sup=9%,conf=85%)
- This means that:
  - "9% of the users visit pages of these 3 themes"
  - "85% of the users interested in diverse and economy-and-finance are also interested in general statistics"

Classification *versus* Association

	Classification	Association
Consequent of rule	1 atom	n atoms
Rule redundancy	little or none	high
Nr. of rules	low	high
Data mining task	supervised	unsupervised
	one target attribute	all attributes are "equal"

Association Rules
Basic Concepts

Market Basket Analysis

Market Baskets data set

TID	Products
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

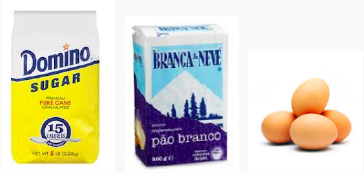
Products are converted in binary flags

→

TID	A	B	C	D	E
1	1	1	0	0	1
2	0	1	0	1	0
3	0	1	1	0	0
4	1	1	0	1	0
5	1	0	1	0	0
6	0	1	1	0	0
7	1	0	1	0	0
8	1	1	1	0	1
9	1	1	1	0	0

Market Basket Analysis: how frequent is an itemset?

- Sugar, Flower and Eggs are sold together



- How important is this set?
- **Support** measures the importance of a set
  - Percentage of transactions t containing the set S
  - Absolute support: number of transactions t containing the set S

Market Basket Analysis: how predictive is an itemset?

- Frequent itemsets are used to generate association rules.
- If you buy sugar and flower, you also buy eggs.
- How strong is this rule?
- **Confidence** measures the strength of the rule
  - Percentage of transactions t that having sugar and flower also have eggs

Association Rules: Basic Concepts

- Consider a set of items I
- A transaction t is a subset of items, i.e. $t \subseteq I$
- Given a data set of transactions $D = \{t_i\}_{i=1}^{N}$
- An association rule is defined as an implication $X \rightarrow Y$, where
  - X and Y are itemsets, i.e. $X, Y \subseteq I$
  - $X \neq \emptyset$, $Y \neq \emptyset$ and $X \cap Y = \emptyset$
- $sup(X)$ is the proportion of transactions in D that include the itemset X
- support: $sup(X \rightarrow Y) = sup(X \cup Y)$
- confidence: $conf(X \rightarrow Y) = sup(X \cup Y)/sup(X)$

Association Rules: an example

Given the data

Transactions ID	Items Bought
100	A, B, C
200	A, C
150	A, D
500	B, E, F

$\rightarrow$

TID	A	B	C	D	E	F
100	1	1	1	0	0	0
200	1	0	1	0	0	0
150	1	0	0	1	0	0
500	0	1	0	0	1	1

- The itemsets with a minimum support of 50%

Frequent Itemsets	Support
{A}	75%
{B}	50%
{C}	50%
{A,C}	50%

- Rules with minimum support of 50% and minimum confidence of 50%
  - $A \rightarrow C$
    - $sup(A \rightarrow C) = sup(\{A, C\}) = 50\%$
    - $conf(A \rightarrow C) = sup(\{A, C\})/sup(\{A\}) = 66.6\%$
  - $C \rightarrow A$
    - $sup(C \rightarrow A) = sup(\{A, C\}) = 50\%$
    - $conf(C \rightarrow A) = sup(\{A, C\})/sup(\{C\}) = 100\%$

Association Rules: exercise

Given

Cliente	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13
1	1	1	0	0	1	0	0	0	0	0	0	0	0
2	0	0	1	0	0	1	0	0	0	0	0	0	0
3	1	0	1	1	1	0	0	0	0	0	0	0	0
4	1	1	1	0	1	0	0	0	0	0	0	0	0
5	0	0	1	0	0	1	0	1	1	1	0	0	0
6	0	1	0	0	0	0	0	1	0	1	0	0	0
7	1	0	0	0	0	0	1	1	0	1	0	1	1
8	0	1	0	0	0	0	0	1	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1	0	1	0

Calculate

- Support of
  - $\{A3\}$
  - $\{A3, A5\}$
  - $\{A3, A5, A1\}$
- Confidence of
  - $\{A3\} \rightarrow \{A4\}$
  - $\{A3\} \rightarrow \{A5\}$
  - $\{A3, A5\} \rightarrow \{A1\}$
  - $\{A3, A5\} \rightarrow \{A1, A4\}$

References

References

Aggarwal, C. C. (2015).
Data Mining, The Texbook.
Springer.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A. I. (1996).
Fast discovery of association rules.
In *Advances in Knowledge Discovery and Data Mining*, pages 307–328. American Association for Artificial Intelligence.

Agrawal, R. and Srikant, R. (1994).
Fast algorithms for mining association rules in large databases.
In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499. Morgan Kaufmann Publishers Inc.

Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. (1997).
Dynamic itemset counting and implication rules for market basket data.
In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, volume 26, pages 255–264. ACM.

Domingo, C., Gavalda, R., and Watanabe, O. (1998).
On-line sampling methods for discovering association rules.

References (cont.)

Gama, J. (2016).
Association rules.
Slides.

Gama, J., Oliveira, M., Lorena, A. C., Faceli, K., and de Leon Carvalho, A. P. (2015).
Extração de Conhecimento de Dados - Data Mining.
Edições Sílabo, 2nd edition.

Han, J., Kamber, M., and Pei, J. (2011).
Data Mining: Concepts and Techniques.
Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

Han, J., Pei, J., Yin, Y., and Mao, R. (2004).
Mining frequent patterns without candidate generation: A frequent-pattern tree approach.
Data Mining and Knowledge Discovery, 8(1):53–87.

Jorge, A. (2016).
Association rules.
Slides.

Liu, B. (2011).
Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data.
Springer, 2nd edition.

References (cont.)

Savasere, A., Omiecinski, E., and Navathe, S. B. (1995).
An efficient algorithm for mining association rules in large databases.
In *Proceedings of the 21th International Conference on Very Large Data Bases*, VLDB '95, pages 432–444. Morgan Kaufmann Publishers Inc.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005).
Introduction to Data Mining.
Addison Wesley.

Toivonen, H. (1996).
Sampling large databases for association rules.
In *Proceedings of the 22th International Conference on Very Large Data Bases*, VLDB '96, pages 134–145. Morgan Kaufmann Publishers Inc.

Torgo, L. (2017).
Data Mining with R: Learning with Case Studies.
Chapman and Hall/CRC, 2nd edition.