

## Data Mining II / Adv. Topics in Data Science

### Web Mining: Introduction

---

**Álvaro Figueira** (arfiguei@fc.up.pt), Room 1.32

Rita Ribeiro

2023/2024



DEPARTAMENTO DE CIÊNCIA DE COMPUTADORES  
UNIVERSIDADE DO PORTO

1

### Summary

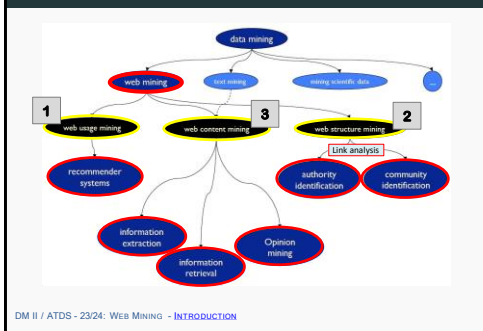
1. Web Mining - Introduction
  - Web Usage Mining
  - Web Structure Mining
  - Web Content Mining
2. Web Usage Mining → Recommender Systems
3. Web Structure Mining → Link Analysis
4. Web Content Mining → Information Retrieval

2

## Introduction

3

### Web Mining - a structured view



4

## Web Mining is

- Application of **data mining** techniques to discover patterns from the web.
- The main objective is to develop more intelligent tools to potentially help the user in **finding, extracting, filtering** and **evaluating** valuable information and resources.

DM II / ATDS - 23/24: WEB MINING - [INTRODUCTION](#)

5

## 1. Web **Usage** Mining

- Problem:
  - the web has available huge **raw log data** on access data, user profiles, registration data, user sessions user queries and so on;
  - this log data can be used to **discover user access patterns**.
- Objective:
  - prediction of the **user behavior and interaction** with the web.
- Applications:
  - user segmentation, recommendation, personalization, adaptation, usability improvement

DM II / ATDS - 23/24: WEB MINING - [INTRODUCTION](#)

6

## 2. Web **Structure** Mining

- Problem:
  - discover the **structural summary** about the web site and web pages;
  - Web Content Mining focuses on intra-document structure (within the document);
  - **Web Structure Mining** focuses on **inter-document structure** (within the web) that is, to discover the **link structure of hyperlinks**;
  - categorize the web pages and make it possible to compare or integrate different web pages.
- Objective:
  - discovery of useful **knowledge from hyperlinks**.
- Applications:
  - discover important pages (information retrieval)
  - discover communities.

DM II / ATDS - 23/24: WEB MINING - [INTRODUCTION](#)

7

## 3. Web **Content** Mining

- Problem:
  - there is a vast **data variety** in the internet;
  - text, image, audio, video, metadata and hyperlinks;
  - a query results in thousands of web pages;
  - Web Content Mining provides a **path to screen more specific data**.
- Objective:
  - extract information from **web pages' content**.
- Applications
  - information extraction, summarization, topic extraction and modeling, sentiment analysis, emotion recognition, information retrieval

DM II / ATDS - 23/24: WEB MINING - [INTRODUCTION](#)

6

8

## Web Mining Pros and Cons

- **Pros:**

- companies can understand the customers' actual need and they can react to the customer needs faster.
- personalize experiences
- higher trade volumes
- perform social media analysis
- ...

- **Cons:**

- personal information of an individual can be used or disseminated, without his knowledge or consent;
- companies collecting the data for a specific purpose might use the data for a totally different purpose;
- web data sets can be very large and may not be mined on a single server.
- ...

7

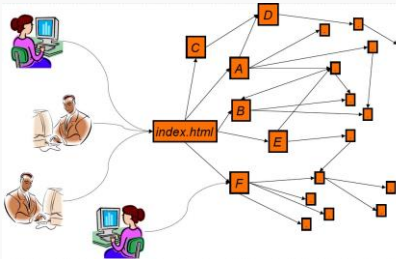
9

## Web Usage Mining

---

10

## Web Site Usage Analysis



DM II / ATDS - 23/24: WEB MINING - [WEB USAGE MINING](#)

11

## Web Usage Mining: Applications

- User Segmentation
- Content Bundling
- Item Recommendation
- Menu Customization
- User Action Prediction
- ...

DM II / ATDS - 23/24: WEB MINING - [WEB USAGE MINING](#)

12

## Web Usage Mining: **Techniques**

- **Clustering** methods
  - user segmentation
  - content bundling
- **Association rule discovery** methods
  - recommendation and personalization
- **Collaborative filtering**
- **Markov chains**
- **Classification**
  - predicting if a user is leaving the site or what will be doing next
- ...

DM II / ATDS - 23/24: WEB MINING - [WEB USAGE MINING](#)

13

## User Segmentation

- We want to find **user segments** according to their activity
- Examples
  - Web site targeting
  - Newsletter targeting
  - Study the evolution of usage styles
- What is a good user segment?
- Data
  - What is necessary to do so?

DM II / ATDS - 23/24: WEB MINING - [WEB USAGE MINING](#)

14

## User Segmentation (cont.)

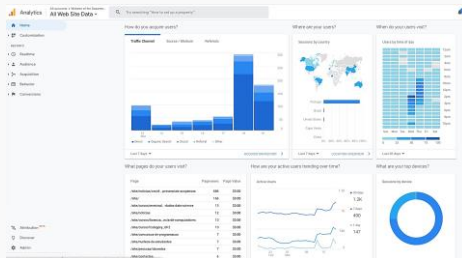
- Example Task:
  - we want to have **different entry pages** for different user segments
- Strategy:
  - users who tend to visit the same pages are regarded as a group
  - in Data Mining, this is the goal of **clustering algorithms**

DM II / ATDS - 23/24: WEB MINING - [WEB USAGE MINING](#)

15

## User Segmentation (cont.)

- Web Site Usage Analysis - user segmentation



DM II / ATDS - 23/24: WEB MINING - [WEB USAGE MINING](#)

16



## User Segmentation (cont.)

- Gathering Access Data
  - From logs
    - web server log files are used
    - however, log data is far from perfect
  - Tagging
    - a piece of programming code is added to each page or template (cookies)
    - access data is stored in a database (or wherever chosen)
    - more events can be captured...
- e.g. Google Analytics

DM II / ATDS - 23/24: WEB MINING - [WEB USAGE MINING](#)

17

## User Segmentation (cont.)

- Server Log Files Data
  - logs in ASCII of the httpd, or CSV files
  - registers each hit: who, what, when, how, from where ...
  - Transfer/Access log: what was seen by the visitor
  - Error log: connection errors
  - Referrer log (English mistake part of the jargon): how each visitor has found the page
  - Agent log: which browser was used

DM II / ATDS - 23/24: WEB MINING - [WEB USAGE MINING](#)

18

## User Segmentation (cont.)

- Data from Access Log - httpd

```
216.239.46.60 - - [04/Jan/2003:14:56:50 +0200] "GET
/-lpis/curriculum/C+Unix/Ergastiria/Week-7/filetype.c.txt HTTP/1.0"
304 -
216.239.46.100 - - [04/Jan/2003:14:57:33 +0200] "GET
/~oswinds/top.html HTTP/1.0" 200 869
64.68.82.70 - - [04/Jan/2003:14:58:25 +0200] "GET /-lpis/systems/r-
device/r_device_examples.html HTTP/1.0" 200 16792
216.239.46.133 - - [04/Jan/2003:14:58:27 +0200] "GET
/-lpis/publications/crc-chapter1.html HTTP/1.0" 304 -
209.237.238.161 - - [04/Jan/2003:14:59:11 +0200] "GET /robots.txt
HTTP/1.0" 404 276
209.237.238.161 - - [04/Jan/2003:14:59:12 +0200] "GET
/teachers/pitas1.html HTTP/1.0" 404 286
216.239.46.43 - - [04/Jan/2003:14:59:45 +0200] "GET
/~oswinds/publications.html HTTP/1.0" 200 48966
```

DM II / ATDS - 23/24: WEB MINING - [WEB USAGE MINING](#)

19

## User Segmentation (cont.)

## NCSA Common Log Format

216.35.116.27	Visitors IP	<i>nslookup: j3407.lnktml.com</i>
-	Identification	<i>Never used</i>
-	Authenticated ID	<i>If there is login</i>
[12/May/ 2002:05:30:23 +0100]	Date / hour of transaction	<i>With the difference to GMT (+0100)</i>
"GET /niaad/ Software/c59/ purchase.html HTTP/1.0"	Method (GET / POST) and accessed file	<i>GET - normal access POST - submit HEAD - used by crawlers HTTP/1.0 (protocol)</i>
404	Error code	<i>200 - success, 300 - redirect 400 - failure(404 - not found) 500 - server errors</i>
349	Size of transaction (bytes)	

DM II / ATDS - 23/24: WEB MINING - [WEB USAGE MINING](#)

20

## • Data from Moodle access log

Date/hour	User	Affected user	Content	Component	Event name	Description	Source	IP
23/03/22 14:08	User4	User1	Unidade: DM2-TACD (FCUP: CC4024, CC4061-2022/2023-META-25)	Sistema	Disciplina visualizada	The user with id '1479' viewed the course with id '4871'.	web	149.90.64.48
23/03/22 14:06	User55	User5	Ficheiro: Slides "Sequence Mining"	Sistema	Atualização da conclusão de atividades da UC	The user with id '17457' updated the completion state for the course module with id '145792' for the user with id '17467'.	web	193.136.24.134
23/03/22 14:06	User55	User5	Ficheiro: Slides "Sequence Mining"	Sistema	Atualização da conclusão de atividades da UC	The user with id '17457' updated the completion state for the course module with id '145792' for the user with id '17467'.	web	193.136.24.134
23/03/22 14:06	User55	User1	Ficheiro: Slides "Sequence Mining"	Ficheiro	Módulo de disciplina visualizado	The user with id '17457' viewed the 'resource' activity with course module id '145792'.	web	193.136.24.134
23/03/22 14:06	User55	User1	Unidade: DM2-TACD (FCUP: CC4024, CC4061-2022/2023-META-25)	Sistema	Disciplina visualizada	The user with id '17457' viewed the course with id '4871'.	web	193.136.24.134
23/03/22 14:04	User33	User1	Unidade: DM2-TACD (FCUP: CC4024, CC4061-2022/2023-META-25)	Sistema	Disciplina visualizada	The user with id '28224' viewed the course with id '4871'.	web	213.22.146.169
23/03/22 14:03	User57	User57	Ficheiro: Slides "Sequence Mining"	Sistema	Atualização da conclusão de atividades da UC	The user with id '17593' updated the completion state for the course module with id '145792' for the user with id '17593'.	web	193.136.24.138
23/03/22 14:03	User57	User57	Ficheiro: Slides "Sequence Mining"	Sistema	Atualização da conclusão de atividades da UC	The user with id '17593' updated the completion state for the course module with id '145792' for the user with id '17593'.	web	193.136.24.138
23/03/22 14:03	User57	User1	Ficheiro: Slides "Sequence Mining"	Ficheiro	Módulo de disciplina visualizado	The user with id '17593' viewed the 'resource' activity with course module id '145792'.	web	193.136.24.138
23/03/22 14:03	User57	User1	Unidade: DM2-TACD (FCUP: CC4024, CC4061-2022/2023-META-25)	Sistema	Disciplina visualizada	The user with id '17593' viewed the course with id '4871'.	web	193.136.24.138
23/03/22 13:39	User60	User60	Ficheiro: Slides "Sequence Mining"	Sistema	Atualização da conclusão de atividades da UC	The user with id '34739' updated the completion state for the course module with id '145792' for the user with id '34739'.	web	193.136.24.134
23/03/22 13:39	User60	User60	Ficheiro: Slides "Sequence Mining"	Sistema	Atualização da conclusão de atividades da UC	The user with id '34739' updated the completion state for the course module with id '145792' for the user with id '34739'.	web	193.136.24.134
23/03/22 13:39	User60	User1	Ficheiro: Slides "Sequence Mining"	Ficheiro	Módulo de disciplina visualizado	The user with id '34739' viewed the 'resource' activity with course module id '145792'.	web	193.136.24.134

21

## User Segmentation - Example (1)

USER	PAGE
1	A
1	B
1	C
2	A
2	C
3	B
3	G
3	F
3	I
4	B
4	C
5	G
5	F
5	I
5	J
6	A
6	C

Look for groups of users (e.g. two groups)

- Two users are similar if they tend to view the same pages

- similarity of two users X and Y can be given by

$$\frac{\# \text{ pages seen by both}}{\# \text{ pages seen by any of them}}$$

- ... or by the **Euclidean Distance**

- each user is described as a point in the space  $\langle A, B, C, D, E, F, G, I, J \rangle \rightarrow 9$ -dimensional space

- then, we calculate the distance between the two users:

$$ED = \sqrt{(A_1 - A_2)^2 + (B_1 - B_2)^2 + \dots + (J_1 - J_2)^2}$$

22

## User Segmentation - Example (2)

USER	PAGE
1	A
1	B
1	C
2	A
2	C
3	B
3	G
3	F
3	I
4	B
4	C
5	G
5	F
5	I
5	J
6	A
6	C

Looking for groups of users:

- **Agglomerative (bottom up) Hierarchical Clustering**
  - we can obtain a hierarchy of clusters
- **K-means clustering**
  - a popular clustering, totally unsupervised

**Obs:** In the latter case the number of clusters must be specified

DM II / ATDS - 23/24: WEB MINING - [WEB USAGE MINING](#)

23

## User Segmentation - Example (3)

USER	PAGE
1	A
1	B
1	C
2	A
2	C
3	B
3	G
3	F
3	I
4	B
4	C
5	G
5	F
5	I
5	J
6	A
6	C

- Applications
  - Entry page customization
    - with known users
    - with unknown users
  - Newsletter customization
  - Segmented usability study
  - Dynamics of the site/app
- Other variables to consider
  - Aggregation of
    - number of page views
    - size of average session
    - number of sessions
    - pageview duration

DM II / ATDS - 23/24: WEB MINING - [WEB USAGE MINING](#)

24

## User Segmentation - Working example

USER	ITEM
1	A
1	B
1	G
2	A
2	C
3	B
3	G
3	F
3	I
4	B
4	C
5	G
5	F
5	I
5	J
6	A
6	C

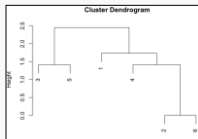
```
# read data
d <- read.csv("toy-session-data.csv")
# transform data into a matrix
dat <- table(d$USER, d$ITEM)

# obtain distance matrix (Euclidean)
dm <- dist(dat)
# cluster and view dendrogram
plot(hclust(dm))

# check parameters of dist and hclust for
# alternatives
```

The hclust function starts by treating each observation as a separate cluster and iteratively merges the closest pairs of clusters until all the observations are in a single cluster. The distance between clusters is defined as the maximum or minimum distance between any pair of observations in the two clusters, depending on the method selected by the user.

	A	B	C	F	G	I	J
1	1	1	0	0	1	0	0
2	1	0	1	0	0	0	0
3	0	1	0	1	1	1	0
4	0	1	1	0	0	0	0
5	0	0	0	1	1	1	1
6	1	0	1	0	0	0	0



DM II / ATDS - 23/24: WEB MINING - [WEB USAGE MINING](#)

25

## User Segmentation - Summary

- We want to differentiate users
  - we use user access data, and then
  - [cluster the users](#)
- A known user can be assigned to an appropriate group
  - and be shown a specific version of the site
  - We can then study group behavior



- We could also [cluster the pages](#)

DM II / ATDS - 23/24: WEB MINING - [WEB USAGE MINING](#)






26

## References

---

27

## References

-  Adomavicius, G., Sankaranarayanan, R., Sen, S., and Tuzhilin, A. (2005).  
**Incorporating contextual information in recommender systems using a multidimensional approach.**  
*ACM Trans. Inf. Syst.*, 23(1):103-145.
-  Adomavicius, G. and Tuzhilin, A. (2005).  
**Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions.**  
*IEEE Trans. on Knowl. and Data Eng.*, 17(6):734-749.
-  Aggarwal, C. C. (2015).  
**Data Mining, The Textbook.**  
Springer.
-  Breese, J. S., Heckerman, D., and Kadie, C. (1998).  
**Empirical analysis of predictive algorithms for collaborative filtering.**  
In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, UAI'98*, pages 43-52. San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
-  Craven, P.  
**Google's pagerank explained and how to make the most of it.**  
<http://www.webworkshop.net/pagerank.html>.

DM II / ATDS - 23/24: WEB MINING

28






## References (cont.)

-  Good, N., Schafer, J. B., Konstan, J. A., Borchers, A., Sarwar, B., Herlocker, J., and Riedl, J. (1999).  
**Combining collaborative filtering with personal agents for better recommendations.**  
*In Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence, AAAI '99/IAAI '99*, pages 439-446, Menlo Park, CA, USA. American Association for Artificial Intelligence.
-  Han, J., Kamber, M., and Pei, J. (2011).  
**Data Mining: Concepts and Techniques.**  
Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.
-  Huang, Z., Chen, H., and Zeng, D. (2004).  
**Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering.**  
*ACM Trans. Inf. Syst.*, 22(1):116-142.
-  Jayanthi, S. (2013).  
**Web mining issues.**  
<http://webminingissues.blogspot.pt/>.
-  Jorge, A. (2016).  
**Web mining.**  
Slides.

DM II / ATDS - 23/24: WEB MINING

29

## References (cont.)

-  Kammergruber, W. C., Viermetz, M., Ehms, K., and Langen, M. (2010).  
**Using association rules for discovering tag bundles in social tagging data.**  
*In 2010 International Conference on Computer Information Systems and Industrial Management Applications, C/SIM, Krakow, Poland, October 8-10, 2010*, pages 414-419.
-  Kohavi, R., Henne, R. M., and Sommerfield, D. (2007).  
**Practical guide to controlled experiments on the web: Listen to your customers not to the hippo.**  
*In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '07*, pages 959-967, New York, NY, USA. ACM.
-  Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. M. (2009).  
**Controlled experiments on the web: Survey and practical guide.**  
*Data Min. Knowl. Discov.*, 18(1):140-181.
-  Liu, B. (2011).  
**Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data.**  
Springer, 2nd edition.
-  Mobasher, B., Dai, H., Luo, T., and Nakagawa, M. (2001).  
**Effective personalization based on association rule discovery from web usage data.**  
*In Proceedings of the 3rd International Workshop on Web Information and Data Management, WIDM '01*, pages 9-15, New York, NY, USA. ACM.

DM II / ATDS - 23/24: WEB MINING

30

## References (cont.)



Oliveira, M. D. B. and Gama, J. (2012).

**A framework to monitor clusters evolution applied to economy and finance problems.**

*Intell. Data Anal.*, 16(1):93-111.



Palmisano, C., Gorgoglione, M., and Tuzhilin, A. (2008).

**Using context to improve predictive modeling of customers in personalization applications.**

*IEEE Transactions on Knowledge & Data Engineering*, 20:1535-1549.



Resnick, P. and Varian, H. R. (1997).

**Recommender systems.**

*Commun. ACM*, 40(3):56-58.



Samatova, N. F., Hendrix, W., Jenkins, J., Padmanabhan, K., and Chakraborty, A. (2013).

**Practical Graph Mining with R.**

Chapman & Hall/CRC.



Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001).

**Item-based collaborative filtering recommendation algorithms.**

In *Proceedings of the 10th International Conference on World Wide Web, WWW '01*, pages 285-295, New York, NY, USA. ACM.

DM II / ATDS - 23/24: WEB MINING

31

## References (cont.)



Shani, G. and Gunawardana, A. (2011).

**Evaluating Recommendation Systems, pages 257–297.**

Springer US, Boston, MA.



Tan, P.-N., Steinbach, M., and Kumar, V. (2005).

**Introduction to Data Mining.**

Addison Wesley.

DM II / ATDS - 23/24: WEB MINING

32