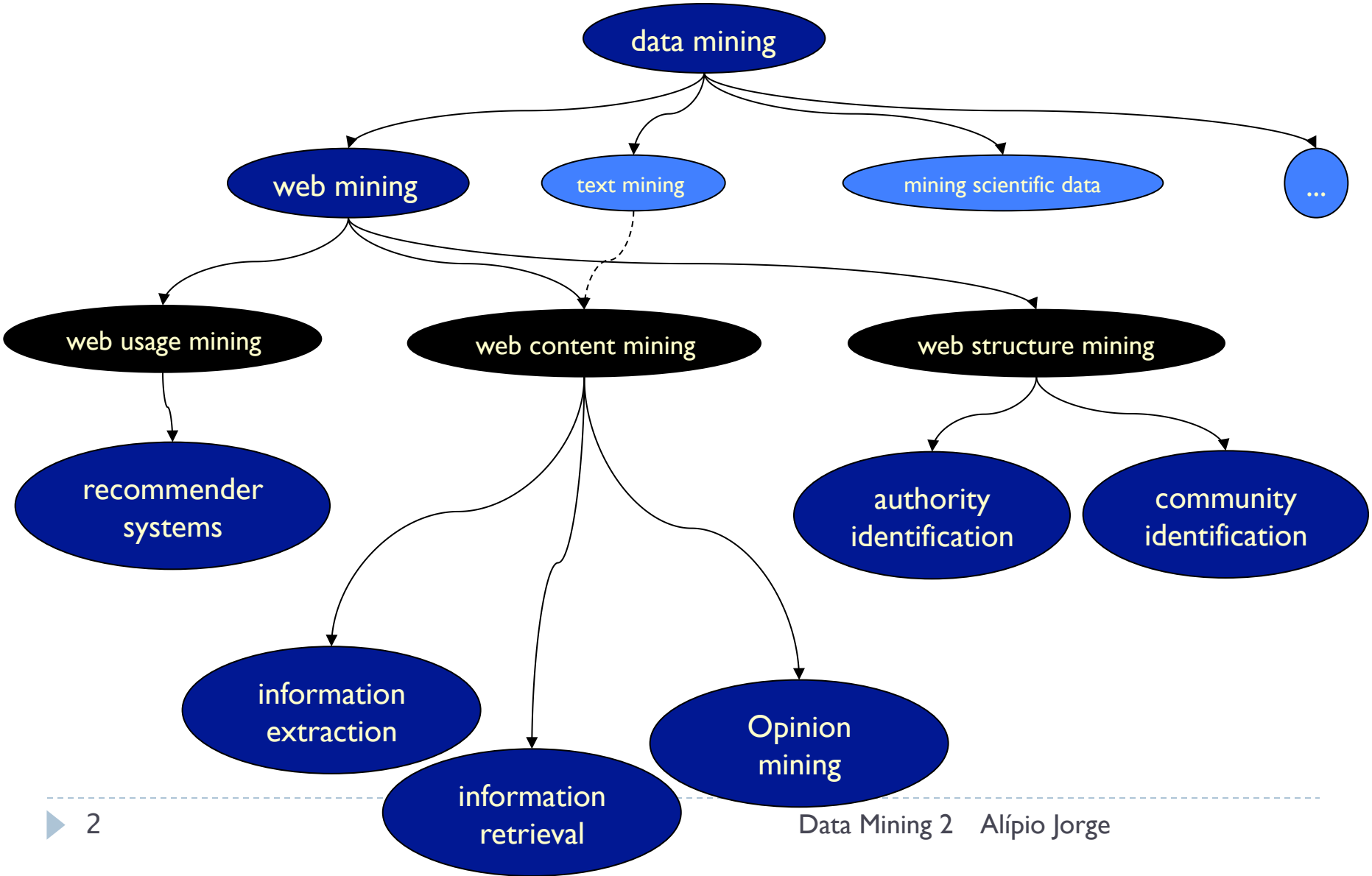
Introduction

Data Mining – a structured view

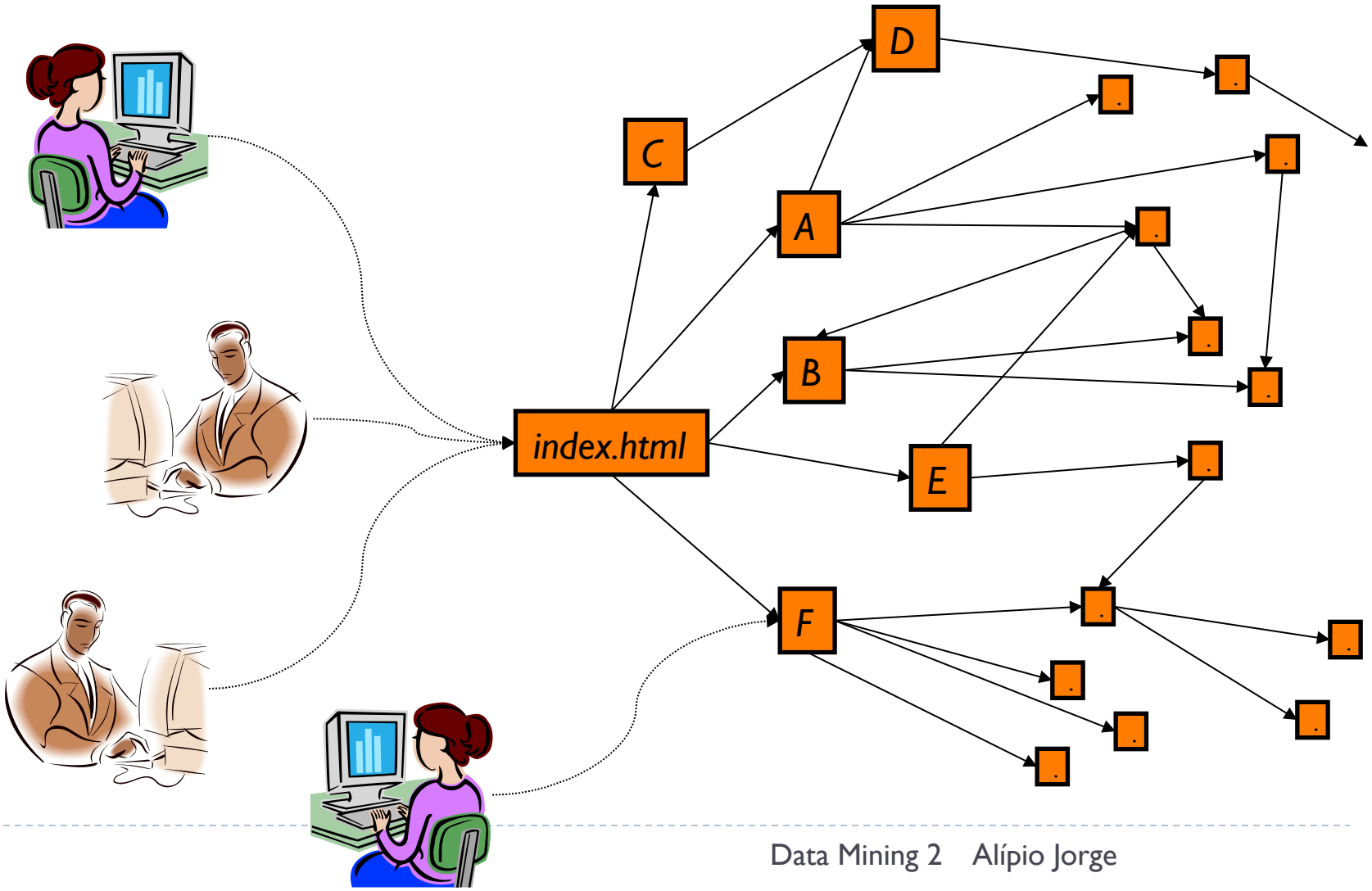


Web Mining

- Web Usage Mining
  - discovery from user access patterns from logs or alike
  - applications:
    - user segmentation, recommendation, personalization, adaptation, usability improvement
- Web Content Mining
  - extracts information from Web pages
  - applications
    - information extraction, summarization, topic extraction, opinion mining, sentiment analysis, information retrieval
- Web Structure Mining
  - discovery of useful knowledge from hyperlinks
  - applications:
    - discover important pages (information retrieval)
    - discover communities

Web usage mining

Web Site Usage Analysis



Web Usage Mining: Problems

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

Web Usage Mining: Techniques

- Clustering methods
  - 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 is doing next
- ...

Web Usage Mining: Problem

- User segmentation
  - we want to find user segments according to their activity
  - what is a good user segment?

- Examples
  - Web site targeting
  - Newsletter targeting
  - Study the evolution of usage styles

- Data
  - What is necessary?

Web Usage Mining: Tec: Clustering

- Example task
  - we want to have different entry pages for different user groups

- Strategy
  - users who tend to visit the same pages are regarded as a group
  - In DM, this is the goal of Clustering Algorithms

Web Site Usage Analysis

Web client

HTTP request to URI

Information (HTML+) sent back

Internet

Metrics / Statistics

Mining

OLAP

Logs

Other data

Web Server

Gathering access data

- From logs
  - Web server log files are used
    - However log data is far from perfect, so an alternative is

- Tagging
  - A piece of programming code is added to each page or template
  - access data is stored in a database (or wherever chosen)
  - more events can be captured

  - e.g. Google Analytics

Log Data

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

Data (access log)

```
216.35.116.27 - - [12/May/2002:05:30:23 +0100] "GET HTTP/1.0" 404 349
...
66.28.250.173 - - [12/May/2002:09:05:32 +0100] "GET /~amjorge/Aulas/madsad/
  ecd1/ HTTP/1.0" 403 348
213.121.90.73 - - [13/May/2002:13:23:28 +0100] "GET /scripts/..%25%35%63../
  winnt/system32/cmd.exe?/c+dir HTTP/1.0" 404 355
213.121.90.73 - - [13/May/2002:13:23:28 +0100] "GET /scripts/..%252f../winnt/
  system32/cmd.exe?/c+dir HTTP/1.0" 404 355
195.221.214.20 - - [13/May/2002:13:39:20 +0100] "GET /niaad/ECO/Rel_Pub00.html
  HTTP/1.0" 200 21493
193.126.80.59 - - [13/May/2002:13:42:22 +0100] "GET /cgi-bin/Count.cgi?
  trgb=ffffff&df=events.dat&ft=0&dd=E HTTP/1.0" 500 652
193.137.36.165 - - [13/May/2002:13:54:44 +0100] "GET /cgi-bin/Count.cgi?
  trgb=ffffff&df=events.dat&ft=0&dd=E HTTP/1.0" 500 652
192.168.1.7 - - [13/May/2002:14:14:21 +0100] "GET /proxy.pac HTTP/1.0" 200 1961
213.150.167.8 - - [13/May/2002:14:34:05 +0100] "GET /niaad/statlog/datasets/
  heart/heart.doc.html HTTP/1.1" 200 611
213.150.167.8 - - [13/May/2002:14:34:53 +0100] "GET /niaad/statlog/datasets/
  heart/heart.doc.html HTTP/1.1" 200 611
```

Data

- **Common log format (NCSA)**

`216.35.116.27`	**Visitors IP**	***nslookup: j3407.inktomi.com***
`-`	**Identification**	***Never used***
`-`	**Authenticated ID**	***If there is login***
`[12/May/ 2002:05:30:23 +0100]`	**Date / hour of transaction**	***With the difference to GMT (+0100)***
`"GET /niaad/ Software/c50/ purchase.html HTTP/1.0"`	**Method (GET / POST) and accessed file**	***GET – normal access*** ***POST – submit*** ***HEAD – used by crawlers*** ***HTTP /1.0 (protocol)***
`404`	**Error code**	***200 – sucess, 300 - redirect*** ***400 – failure(404 – not found)*** ***500 – server errors***
`349`	**Size of transaction (bytes)**	

Web Usage Mining: Clustering

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

- Activity
  - 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:
    - #pages seen by both / # pages seen by any of them
  - We can also use Euclidean distance
    - each user is described as a point in the space

      <A,B,C,D,E,F,G,I,J>
    - then we calculate the distance between the two points

$$\sqrt{(A1-A2)^2+(B1-B2)^2+\ldots+(J1-J2)^2}$$

Web Usage Mining: Clustering

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

- Agglomerative (bottom up) Hierarchical Clustering
  - we can obtain a predefined number of clusters
- Kmeans clustering
  - another popular clustering method

Web Usage Mining: Clustering

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
- Other variables to consider
  - Aggregate
    - number of page views
    - size of average session
    - number of sessions
    - pageview duration

Activity(R)

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

```
## read data (copy and…)
d <-read.table(file("clipboard"),header=TRUE)
## Mac OS X: read.table(pipe("pbpaste"),…)
## or paste into a file and use read.csv
d <-read.csv("toy-session-data.csv")

## transform data into a matrix
dat <-table(d$USER,d$PAGE)

## obtain distance matrix (Euclidean)
dm <-dist(dat)

## cluster and view dendrogram
plot(hclust(dm))

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

Case summary

- We want to differentiate users
- We use access data
- Then clustering
- A known user can be assigned to an appropriate group
  - And be shown a specific version of the site

- We could also
  - Cluster pages

Web content mining

a small example

Web content mining: tag bundles

- Problem
  - Social sites have tagged items
  - Users provide tags (social tagging)
  - We can bundle tags for helping users to provide tags
  - How?

- Evaluation:
  - How can it be done?

Web content mining: tag bundles

- Process [Kammergruber et al. 2010]
  - Get tags provided by users
  - A transaction is the set of tags given by a single user
  - Use Association Rule discovery to find tags that are associated to each other

Resources

- Book
  - Web Data Mining, Bing Liu
- Papers
  - Walter Kammergruber, M. Viermetz, K. Ehms "Using association rules for discovering tag bundles in social tagging data", CISIM 2010.