

Brief Introduction to Data & Web Mining

Olfa Nasraoui

CECS 694:

Web mining for e-commerce and
information retrieval



Outline

- Knowledge Discovery in DB & Data Mining
 - Motivation & Definition of KDD
 - DM Tasks
- Web Mining
 - Motivation & Differences from DM
 - Types of Web Data to be Mined
 - Web Personalization & Profiling

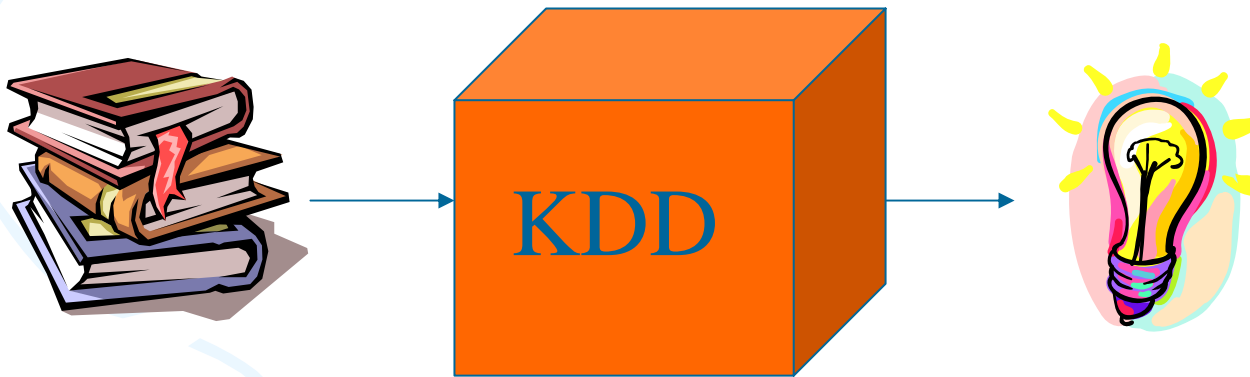
Knowledge Discovery in DB & Data Mining: Motivation

- *Explosion* in electronically stored data
- Huge DB's contain a *wealth* of info, *still* not *fully* exploited (valuable info (*gold!*) may be lurking within data).
- Accessing useful info. more and more difficult (Info. Retrieval in various data repositories: Image DB, WWW, ...etc).



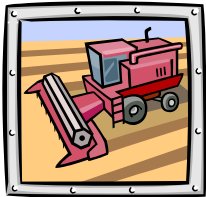
Knowledge Discovery in DB: Definition

KDD: discovering useful info. and knowledge from huge data repositories (patterns, associations, ...etc)



Knowledge Discovery in DB: Process

- 1. Data Preprocessing:** Cleaning, integration, transformation
- 2. Data Mining:** Intelligent methods for extracting knowledge/digging for gold
- 3. Pattern evaluation and presentation**

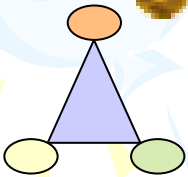


Data Mining Tasks



● **Class description:** summarization/characterization of a data collection

● **Mining associations:** Discovering association relationships/correlations among a set of items in the form of rules: $X \Rightarrow Y$ (DB tuples satisfying X are likely to satisfy Y)



Data Mining Tasks



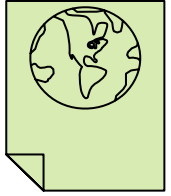
- **Classification:** Construct a model for each class of *labeled* training data based on its features and use it to classify *future* data
- **Prediction:** Predict the possible values of some missing data/attributes based on similar objects



Data Mining Tasks



- **Clustering:** Dividing *unlabeled* data into groups/clusters such that data in *same* cluster are as *similar* as possible while data from *distinct* clusters are *dissimilar*
- **Time-series analysis:** Discover regularities & interesting characteristics, search for similar sequences or subsequences, mining seq. patterns, trends/deviations

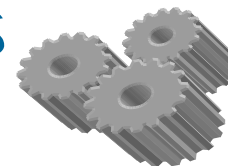
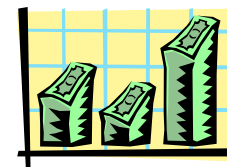


Web Mining

- ☀ WWW: Vital, popular source of information
- ☀ Searching for info.:
 - One of the most common tasks (71% of users)
 - Can be frustrating
- ☀ Navigation (self-guided, sometimes aimless search)
- ☀ Design of good Web sites important

Applications of Web Mining

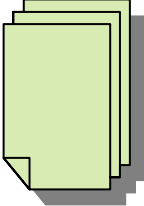

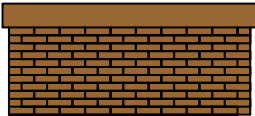
- Automatic personalization: Adaptive sites can facilitate navigation, search
- E-commerce Web sites can be made more user friendly
- Optimized marketing efforts for trading products, services, information
- Improved search engines



Differences from Regular DM

- **Huge**, semi/unstructured, highly dynamic data
 - ✦ Content: > 8 Billion pages
 - ✦ Usage: > daily visitors to popular sites: in millions
- WWW data **corrupted with noise**
(unintentional access, incorrect logging, imperfect crawling)
- Data is **dynamic** (expired links, changing user interests/activities, changing Web content & structure, ..., etc)

Types of Web Data \Rightarrow Types of Web Mining

- **Content:** Web pages HTML content, snippets, multimedia data (*Web content mining*) 
- **Usage:** Web access log files/ clickstream data (*Web usage mining*) 
- **Structure:** *Link* topology of the Web (*Web structure mining*) 

WINDOWS TO THE UNIVERSE

Teacher Resources

Become a Sponsor

About the Site

Space Weather

Space Missions

Myths

Art, Books, Film

History & People

Geology

Life

Physics

Images & Multimedia

SEARCH:

GO

Advanced Search





Our Planet



Our Solar System



Astronomy & The Universe

Stuff to do...

 Create a Journal	 Play Games	 Take a Guided Tour
 Read the News	 Buy Science Stuff	 Become a Member

Highlights:

Learn about how Earth's climate is changing through our [Climate and Global Change](#) section!

OUR OTHER PROJECTS:

CSEM HIGH TIDE SPARC Space Weather Today

- History & People
- Geology
- Life
- Physics
- Images & Multimedia



Stuff to do...

- Create a Journal
- Play Games
- Take a Guided Tour
- Read the News
- Buy Science Stuff
- Become a Member

Highlights:

Interested in collaborating with Windows to the Universe? Research Education Partnerships offer an opportunity to become part of the Windows to the Universe educational community.

SEARCH:

GO

Advanced Search

OUR OTHER PROJECTS:

CSEM HIGH TIDE SPARC Space Weather Today

- Sun
- Mercury
- Venus
- Earth
- Mars
- Asteroid
- Jupiter
- Saturn
- Uranus
- Neptune
- Pluto
- Comet

- Enter with CD
- Tools
- Credits
- Site Map
- Contact us
- My Windows Settings



- History & People
- Geology
- Life
- Physics
- Images & Multimedia



Stuff to do...

- Create a Journal
- Play Games
- Take a Guided Tour
- Read the News
- Buy Science Stuff
- Become a Member

Highlights:

Interested in collaborating with Windows to the Universe? Research Education Partnerships offer an opportunity to become part of the Windows to the Universe educational community.

OUR OTHER PROJECTS:
 CSEM HIGH TIDE SPARC Space Weather Today

SEARCH:

 GO

Advanced Search

- Sun
- Mercury
- Venus
- Earth
- Mars
- Asteroid
- Jupiter
- Saturn
- Uranus
- Neptune
- Pluto
- Comet

- Enter with CD
- Tools
- Credits
- Site Map
- Contact us
- My Windows Settings



Web Content for file <http://www.windows.ucar.edu/>

```
<html>
<head>
<title>Windows to the Universe</title>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<!-- Fireworks MX Dreamweaver MX target. Created Wed Jan 28 11:46:21 GMT-0800
(Pacific Standard Time) 2004-->
<style type="text/css">
BODY,TD { background-color: black;
  color: #ffd782;
  font-family: Arial, Helvetica, sans-serif;
  font-size: 10pt
}
A { color:#66ccff }
}
</style>
<script type="text/javascript" language="JavaScript"
  src="JavaScript/news.js"></script>
<script type="text/javascript" language="JavaScript"
  src="JavaScript/win_home.js"></script>
</head>
```


Web Content for file <http://www.windows.ucar.edu/>

```
<body bgcolor="#000000" background="/images_home/orion_star_field_1.gif"
onLoad="showNews();
  MM_preloadImages('/images_home/jupiterB.jpg','/images_home/memberB.jpg','/i
images_home/storeB.jpg','/images_home/newsB.jpg','/images_home/toursB.jpg','/i
images_home/gamesB.jpg','/images_home/journalB.jpg','/images_home/planets.gif
','/images_home/earth_slideshow.gif','/images_home/solar_system_slideshow.gif','
/images_home/universe_slideshow.gif')">
<div id="planets" style="position:absolute; z-index:1; visibility:hidden; width:512;
height:51" >
  
</div>
<map name="m_planets">
  <area shape="rect" coords="461,0,507,51" onMouseOver="menuOver();"
onMouseOut="menuOut();" href="/tour/link=/comets/comets.html"
title="Comets" alt="Comets" >
  <area shape="rect" coords="424,0,461,51" onMouseOver="menuOver();"
onMouseOut="menuOut();" href="/tour/link=/pluto/pluto.html" title="Pluto"
alt="Pluto" >
```

Web Content for file <http://www.windows.ucar.edu/>

```
<area shape="rect" coords="377,0,424,51" onMouseOver="menuOver();"
onMouseOut="menuOut();" href="/tour/link=/neptune/neptune.html"
title="Neptune" alt="Neptune" >
  <area shape="rect" coords="331,0,377,51"
onMouseOver="menuOver();" onMouseOut="menuOut();"
href="/tour/link=/uranus/uranus.html" title="Uranus" alt="Uranus" >
  <area shape="rect" coords="287,0,331,51"
onMouseOver="menuOver();" onMouseOut="menuOut();"
href="/tour/link=/saturn/saturn.html" title="Saturn" alt="Saturn" >
  <area shape="rect" coords="248,0,287,51"
onMouseOver="menuOver();" onMouseOut="menuOut();"
href="/tour/link=/jupiter/jupiter.html" title="Jupiter" alt="Jupiter" >
  <area shape="rect" coords="198,0,248,51"
onMouseOver="menuOver();" onMouseOut="menuOut();"
href="/tour/link=/asteroids/asteroids.html" title="Asteroids"
alt="Asteroids" >
```

Usage Data

- 66.151.181.4 - - [01/Jan/2005:00:00:01 -0700] "GET /tour/link=/mythology/nut_sky.sp.html HTTP/1.0" 200 10028 "-" "FAST-WebCrawler/3.8/Scirus (scirus-crawler@fast.no; http://www.scirus.com/srsapp/contactus/)"
- 83.130.199.108 - - [01/Jan/2005:00:00:01 -0700] "GET /tour/link=/uranus/uranus.html HTTP/1.1" 200 10910 "http://www.windows.ucar.edu/tour/link=/mars/mars.html" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"
- 68.41.241.123 - - [01/Jan/2005:00:00:01 -0700] "GET /favicon.ico HTTP/1.1" 404 294 "-" "Mozilla/5.0 (Windows; U; Windows NT 5.1; rv:1.7.3) Gecko/20040913 Firefox/0.10"
- 83.130.199.108 - - [01/Jan/2005:00:00:02 -0700] "GET /our_solar_system/moon_counts.js HTTP/1.1" 200 1109 "http://www.windows.ucar.edu/tour/link=/uranus/uranus.html" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"
- 83.130.199.108 - - [01/Jan/2005:00:00:02 -0700] "GET /images/uranus.jpg HTTP/1.1" 200 23897 "http://www.windows.ucar.edu/tour/link=/uranus/uranus.html" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"

Usage Data

- 66.151.181.4 - - [01/Jan/2005:00:00:01 -0700] "GET /tour/link=/mythology/nut_sky.sp.html HTTP/1.0" 200 10028 "-" "FAST-WebCrawler/3.8/Scirus (scirus-crawler@fast.no; http://www.scirus.com/srsapp/contactus/)"
- 83.130.199.108 - - [01/Jan/2005:00:00:01 -0700] "GET /tour/link=/uranus/uranus.html HTTP/1.1" 200 10910 "http://www.windows.ucar.edu/tour/link=/mars/mars.html" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"
- 68.41.241.123 - - [01/Jan/2005:00:00:01 -0700] "GET /favicon.ico HTTP/1.1" 404 294 "-" "Mozilla/5.0 (Windows; U; Windows NT 5.1; rv:1.7.3) Gecko/20040913 Firefox/0.10"
- 83.130.199.108 - - [01/Jan/2005:00:00:02 -0700] "GET /our_solar_system/moon_counts.js HTTP/1.1" 200 1109 "http://www.windows.ucar.edu/tour/link=/uranus/uranus.html" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"
- 83.130.199.108 - - [01/Jan/2005:00:00:02 -0700] "GET /images/uranus.jpg HTTP/1.1" 200 23897 "http://www.windows.ucar.edu/tour/link=/uranus/uranus.html" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"

One Entry of the Log File

- 83.130.199.108 - -
- [01/Jan/2005:00:00:01 -0700]
- "GET
- /tour/link=/uranus/uranus.html
- HTTP/1.1"
- 200
- 10910
- http://www.windows.ucar.edu/tour/link=/mars/mars.html
- "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"

Usage Data

- 66.151.181.4 - - [01/Jan/2005:00:00:01 -0700] "GET /tour/link=/mythology/nut_sky.sp.html HTTP/1.0" 200 10028 "-" "FAST-WebCrawler/3.8/Scirus (scirus-crawler@fast.no; http://www.scirus.com/srsapp/contactus/)"
- 83.130.199.108 - - [01/Jan/2005:00:00:01 -0700] "GET /tour/link=/uranus/uranus.html HTTP/1.1" 200 10910 "http://www.windows.ucar.edu/tour/link=/mars/mars.html" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"
- 68.41.241.123 - - [01/Jan/2005:00:00:01 -0700] "GET /favicon.ico HTTP/1.1" 404 294 "-" "Mozilla/5.0 (Windows; U; Windows NT 5.1; rv:1.7.3) Gecko/20040913 Firefox/0.10"
- 83.130.199.108 - - [01/Jan/2005:00:00:02 -0700] "GET /our_solar_system/moon_counts.js HTTP/1.1" 200 1109 "http://www.windows.ucar.edu/tour/link=/uranus/uranus.html" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"
- 83.130.199.108 - - [01/Jan/2005:00:00:02 -0700] "GET /images/uranus.jpg HTTP/1.1" 200 23897 "http://www.windows.ucar.edu/tour/link=/uranus/uranus.html" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"

Another Entry of the Log File

- **66.151.181.4**
- **[01/Jan/2005:00:00:01 -0700]**
- **"GET**
- **/tour/link=/mythology/nut_sky.sp.html**
- **HTTP/1.0"**
- **200**
- **10028 "-"**
- **"FAST-WebCrawler/3.8/Scirus (scirus-crawler@fast.no;http://www.scirus.com/srsapp/contactus/)"**

Structure Data

```
<area shape="rect" coords="461,0,507,51"  
onMouseOver="menuOver();" onMouseOut="menuOut();"  
href="/tour/link=/comets/comets.html" title="Comets"  
alt="Comets" >
```

```
<area shape="rect" coords="424,0,461,51"  
onMouseOver="menuOver();" onMouseOut="menuOut();"  
href="/tour/link=/pluto/pluto.html" title="Pluto" alt="Pluto" >
```


- History & People
- Geology
- Life
- Physics
- Images & Multimedia

Our Planet our solar system

Stuff to do...

- Create a Journal
- Play Games
- Take a Guided Tour
- Read the News
- Buy Science Stuff
- Become a Member

Highlights:

Interested in collaborating with Windows to the Universe? Research Education Partnerships offer an opportunity to become part of the Windows to the Universe educational community.

SEARCH:

GO

Advanced Search

OUR OTHER PROJECTS:

CSEM HIGH TIDE SPARC Space Weather Today

- Sun
- Mercury
- Venus
- Earth
- Mars
- Asteroid
- Jupiter
- Saturn
- Uranus
- Neptune
- Pluto
- Comet



- Enter with CD
- Tools
- Credits
- Site Map
- Contact us
- My Windows Settings

NASA

Uranus

NCAR National Center for Atmospheric Research

CSEM

CISM



Windows to the Universe

Beginner Intermediate Advanced



Comets



Image courtesy of NASA Click on image for full size version

Not long ago, many people thought that comets were a sign that something bad was about to happen to them. People didn't understand how objects in the sky moved, so the sight of a comet must have been very disturbing. There are many historical records and works of art which record the appearance of comets and link them with terrible events such as wars or plagues.

Now we know that comets are lumps of ice and dust that periodically come into the center of the solar system from somewhere in its outer reaches, and that some comets make repeated trips. When comets get close enough to the Sun, heat makes them start to evaporate. Jets of gas and dust form long tails that we can see from Earth. These tails can sometimes be millions of miles long.

In 1985-1986, a spacecraft called Giotto visited the most famous comet, Halley, on Halley's most recent visit to the inner solar system. In 1994, comet Shoemaker-Levy became trapped by the gravity of Jupiter and plunged into Jupiter's atmosphere!

In 1996 and 1997 we saw comet Hyakutake, and comet Hale-Bopp. Hale-Bopp was one of the brightest comets ever seen from Earth. Comet Linear was discovered in 1999 and made its closest approach to the Sun in July 2000. The Stardust spacecraft flew by Comet Wild 2 in January 2004, collecting samples of the comet to return to Earth. The newest comet mission is Rosetta -- it will land on a comet named Churyumov-Gerasimenko!

Example of Application: Web Personalization



- **WWW Personalization:** Tailor user's interaction w/ Web info space **based on info** about user
- Need to **gather info.** about user
- **Manually** entered profiles are subjective, static, not always available, and continue raising *privacy* concerns
- Alternative: Extract profiles based on **all** users' *access patterns*: **Mass profiling** \Rightarrow **anonymous profiles**
- **Typical profile = {URLs user is interested in, with corresponding URL significance weights}**

Example of profiles description discovered using web usage mining, with corresponding interestingness measures

i	$ \mathcal{X}_i $	$ \mathcal{X}_i^* $	N_i^*	description	σ_i^{*2}
1	219	132	140.5	main page, class list, course enquiries and people	0.16
2	119	73	77.0	main page, class list, course and undergraduate degree enquiries	0.27
3	140	85	91.6	main page and class list	0.13
4	129	71	80.7	main page, people, individual faculty, research and graduate degree pages	0.39

Example Profiles for MU-CECS1 at $L = 2$

- **General outside visitor: Profiles 1 and 3**
- **Prospective students: Profiles 2 and 4**
- **Insiders (students): Profiles 6, 7, ...etc**



Conclusion

- Web mining is a special discipline of data mining that is concerned with mining web data
- Web data: usage, structure, content.
- Increasing dependence on the Web to do most information enquiries and daily business + Special characteristics of web data (huge volumes, dynamic, noisy, missing, heavy pre-processing, domain knowledge) make web mining a crucial and challenging area of research.
- Many interesting and challenging applications of Web mining: profiling, personalization, intelligent search and retrieval, automatic website organization, ...etc.
- An interesting area of research since late 90's, still many open areas of research for the future!