Web Information Processing and Applications: Part II—Web Mining

Instructor: Linli Xu
linlixu@ustc.edu.cn

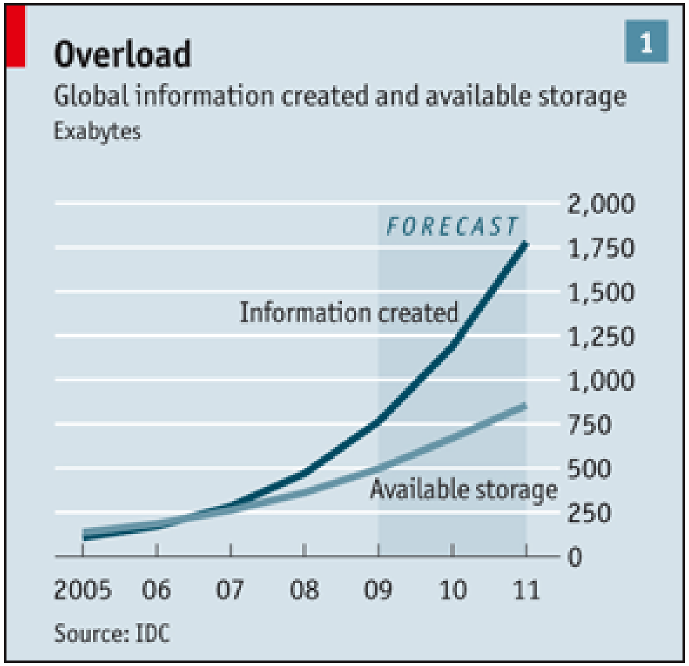
Introduction to Web Mining

Data Mining

Knowledge discovery from data

Data Mining

We are producing more data than we are able to store!



[The economist, 2010]

POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate
OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins?
NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics
PLUS
Juan Enriquez Reprograms Life
James Gleick Unsplits the Bit
AND
Lawrence Weschler Questions the Cloud
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE

Data Mining

Discovery of patterns（模式） and models（模型） that are:

- Valid: hold on new data with some certainty
- Useful: should be possible to act on the item
- Unexpected: non-obvious to the system
- Understandable: humans should be able to interpret the pattern

Data Mining: Cultures

Data mining overlaps with:

- Databases: Large-scale data, simple queries
- Machine learning: General data, Complex models
- Statistics: Predictive Models

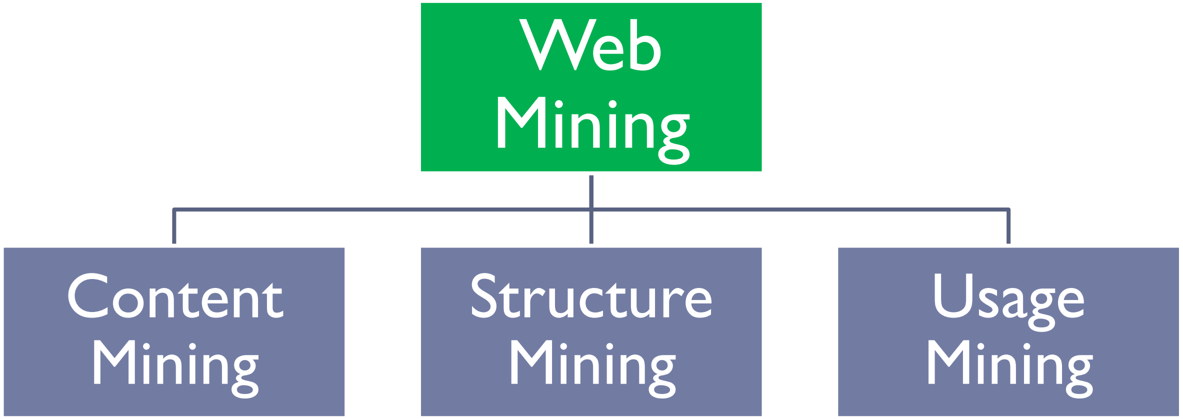
What is Web Mining?

Discovering useful information from the World-Wide Web and its usage patterns

Web Mining vs. Data Mining

- Structure (or lack of it)
  - Textual information and linkage structure
- Scale
  - Data generated per day is comparable to largest conventional data warehouses
- Speed
  - Often need to react to evolving usage patterns in real-time (e.g., merchandising)

Web Mining Topics



Web Mining Topics

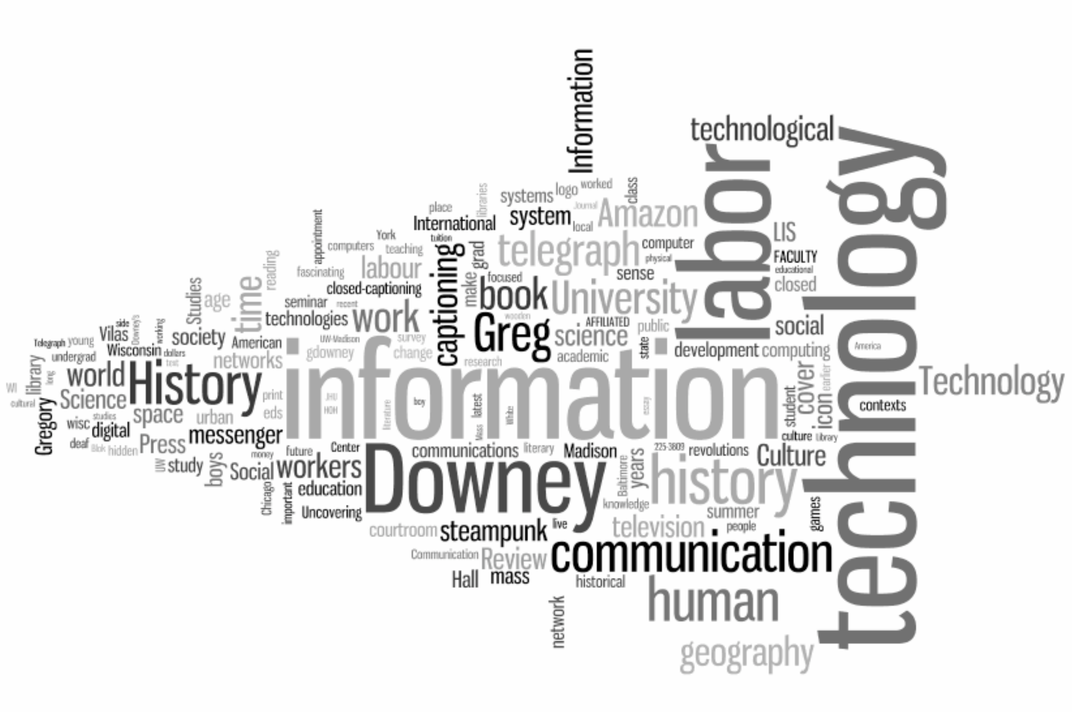
- Content mining

- Structure mining

- Usage mining

Web Content

- Text



Web Content Mining

- Text
- Image

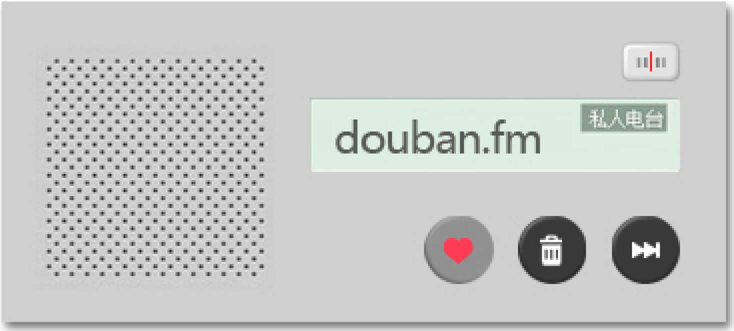
Web Content Mining

- Text
- Image
- Video

Web Content Mining

- Text
- Image
- Video
- Audio

etc.



Web Content Mining

Definition: Web content mining is the process of extracting useful information from the contents of Web documents.

- Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables.

- Research activities in this field also involve using techniques from other disciplines such as Information Retrieval (IR—信息检索) and Natural Language Processing (NLP—自然语言处理).

Pre-processing Content

- Content preparation
  - Extract text from HTML.
  - Perform Stemming（词根化）.
  - Remove Stop Words （停止词）.
  - Calculate Collection Wide Word Frequencies (DF).
  - Calculate per Document Term Frequencies (TF).

- Vector creation
  - Common Information Retrieval Technique.
  - Each document (HTML page) is represented by a sparse vector of term weights.
  - TFIDF weighting is most common.
  - Typically, additional weight is given to terms appearing as keywords or in titles.

- Semantic representation
  - Topic models
  - Word2vec, doc2vec

Common Web Content Mining Topics

The more basic and popular data mining techniques include:

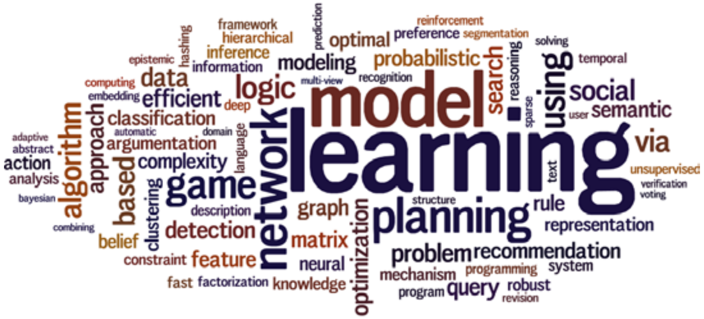
- Feature extraction / representation
- Classification（分类）
- Clustering（聚类）
- Associations

The other significant ideas:

- Topic identification, tracking and drift analysis
- Concept hierarchy creation
- Relevance of content

Web Content Representation

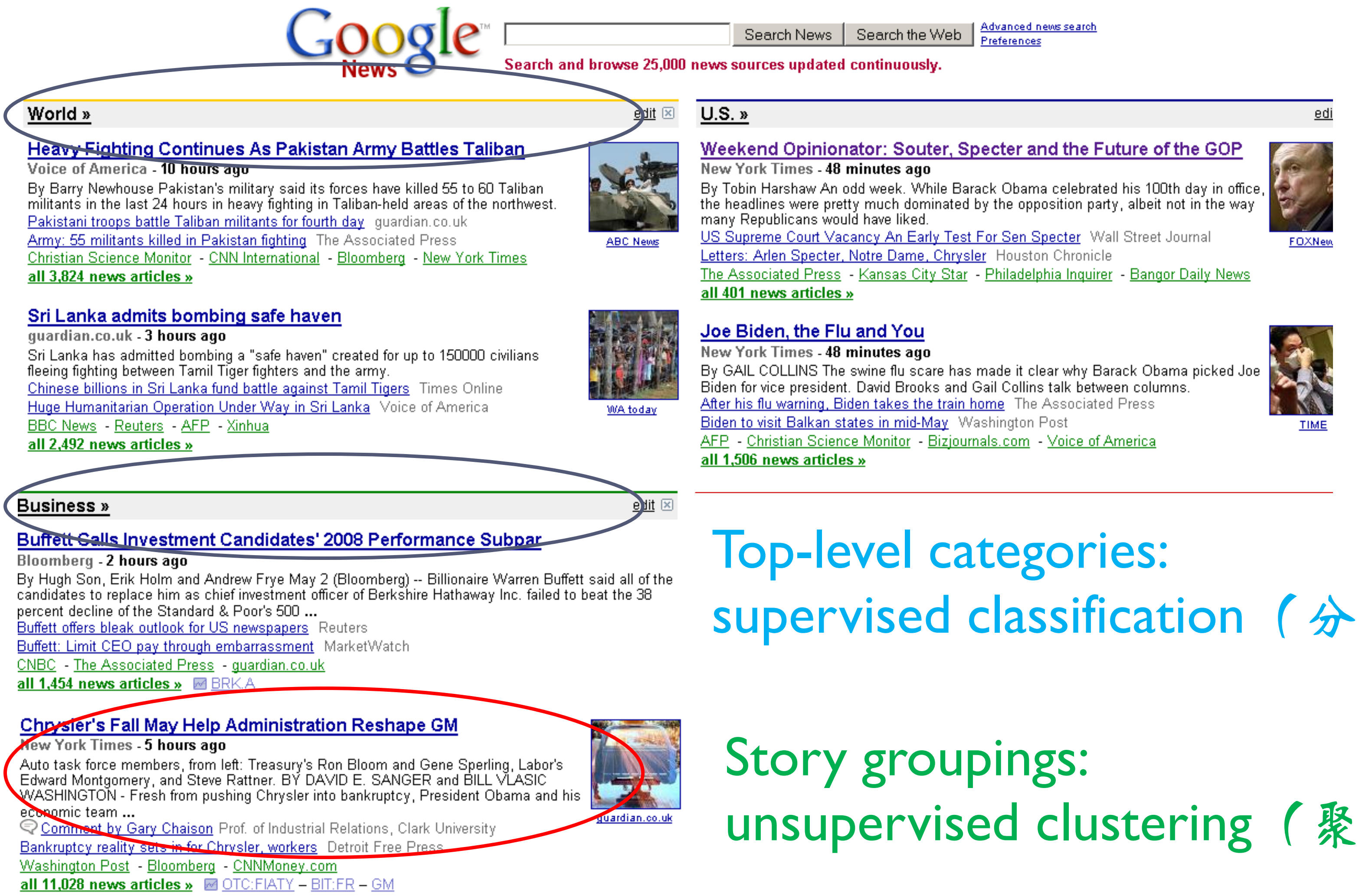
- A TF-IDF vector



- A representation of topics

- A vector with semantic relationships
  - Beijing – China + France ~= Paris

Web Content Mining



Search News | Search the Web | Advanced news search | Preferences

Search and browse 25,000 news sources updated continuously.

World » edit

Heavy Fighting Continues As Pakistan Army Battles Taliban
Voice of America - **10 hours ago**
By Barry Newhouse Pakistan's military said its forces have killed 55 to 60 Taliban militants in the last 24 hours in heavy fighting in Taliban-held areas of the northwest.
Pakistani troops battle Taliban militants for fourth day guardian.co.uk
Army: 55 militants killed in Pakistan fighting The Associated Press
Christian Science Monitor - CNN International - Bloomberg - New York Times
all 3,824 news articles »
ABC News

Sri Lanka admits bombing safe haven
guardian.co.uk - **3 hours ago**
Sri Lanka has admitted bombing a "safe haven" created for up to 150000 civilians fleeing fighting between Tamil Tiger fighters and the army.
Chinese billions in Sri Lanka fund battle against Tamil Tigers Times Online
Huge Humanitarian Operation Under Way in Sri Lanka Voice of America
BBC News - Reuters - AFP - Xinhua
all 2,492 news articles »
WA today

Business » edit

Buffett Calls Investment Candidates' 2008 Performance Subpar
Bloomberg - **2 hours ago**
By Hugh Son, Erik Holm and Andrew Frye May 2 (Bloomberg) -- Billionaire Warren Buffett said all of the candidates to replace him as chief investment officer of Berkshire Hathaway Inc. failed to beat the 38 percent decline of the Standard & Poor's 500 ...
Buffett offers bleak outlook for US newspapers Reuters
Buffett: Limit CEO pay through embarrassment MarketWatch
CNBC - The Associated Press - guardian.co.uk
all 1,454 news articles » BRK.A

Chrysler's Fall May Help Administration Reshape GM
New York Times - **5 hours ago**
Auto task force members, from left: Treasury's Ron Bloom and Gene Sperling, Labor's Edward Montgomery, and Steve Rattner. BY DAVID E. SANGER and BILL VLASIC WASHINGTON - Fresh from pushing Chrysler into bankruptcy, President Obama and his economic team ...
Comment by Gary Chaison Prof. of Industrial Relations, Clark University
Bankruptcy reality sets in for Chrysler, workers Detroit Free Press
Washington Post - Bloomberg - CNNMoney.com
all 11,028 news articles » OTC:FIATY – BIT:FR – GM
guardian.co.uk

U.S. » edi

Weekend Opinionator: Souter, Specter and the Future of the GOP
New York Times - **48 minutes ago**
By Tobin Harshaw An odd week. While Barack Obama celebrated his 100th day in office, the headlines were pretty much dominated by the opposition party, albeit not in the way many Republicans would have liked.
US Supreme Court Vacancy An Early Test For Sen Specter Wall Street Journal
Letters: Arlen Specter, Notre Dame, Chrysler Houston Chronicle
The Associated Press - Kansas City Star - Philadelphia Inquirer - Bangor Daily News
all 401 news articles »
FOXNew

Joe Biden, the Flu and You
New York Times - **48 minutes ago**
By GAIL COLLINS The swine flu scare has made it clear why Barack Obama picked Joe Biden for vice president. David Brooks and Gail Collins talk between columns.
After his flu warning, Biden takes the train home The Associated Press
Biden to visit Balkan states in mid-May Washington Post
AFP - Christian Science Monitor - Bizjournals.com - Voice of America
all 1,506 news articles »
TIME

Top-level categories: supervised classification（分类）

Story groupings: unsupervised clustering（聚类）

Document Classification

- “Supervised（有监督）” technique
- Categories are defined and documents are assigned to one or more existing categories
- Training is performed through the use of documents that have already been classified (often by hand) as belonging to a category

Document Clustering

- "Unsupervised（无监督）" technique
- Documents are divided into groups based on a similarity metric
- No pre-defined notion of what the groups should be
- Most common similarity metric is the dot product between two document vectors

Topic Identification and Tracking

- Combination of Clustering and Classification
- As new documents are added to a collection
  - An attempt is made to assign each document to an existing topic (category)
  - The collection is also checked for the emergence of new topics
  - The drift in the topic(s) are also identified

Concept Hierarchy Creation

- Creation of concept hierarchies is important to understand the category and sub categories a document belongs to

- Key factors
  - Organization of categories; e.g. Flat, Tree, or Network
  - Maximum number of categories per document.
  - Category Dimensions; e.g. Subject, Location, Time, Alphabetical, Numerical

Relevance of Content

Relevance can be measured with respect to any of the following criteria

- Document
- Query based
- User Based
- Role/Task Based

Document Relevance

- Measure of how useful a given document is in a given situation
- Commonly seen in the context of queries - results are ordered by some measure of relevance
- In general, a query is not necessary to assign a relevance score to a document

Query Based Relevance

- Most common
- Well established in Information Retrieval
- Similarity between query keywords and document is calculated
- Can be enhanced through additional information such as popularity (Google) or term positions (AltaVista)

User Based Relevance

- Often associated with personalization
- Profile for a particular user is created
- Similarity between a profile and document is calculated
- No query is necessary

Role/Task Based Relevance

- Similar to User Based Relevance
- Profile is based on a particular role or task, instead of an individual
- Input to profile can come from multiple users

Web Content Mining Applications

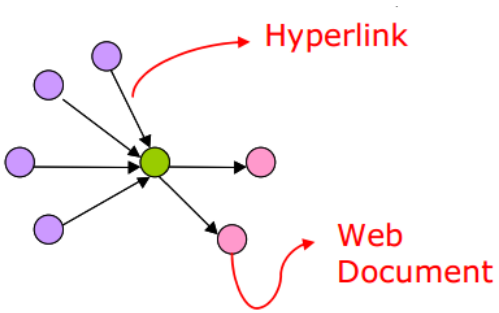
- Identify the topics represented by Web Documents
- Categorize Web Documents
- Find Web Pages across different servers that are similar
- Applications related to relevance
  - Queries – Enhance standard Query Relevance with User, Role, and/or Task Based Relevance
  - Recommendations – List of top "n" relevant documents in a collection or portion of a collection.
  - Filters – Show/Hide documents based on relevance score

Web Mining Topics

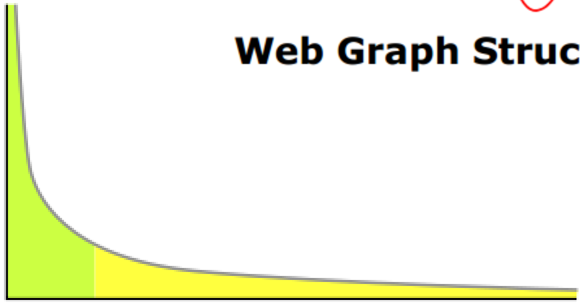
- Content mining
- Structure mining
- Usage mining

Web/Networks Structure Mining

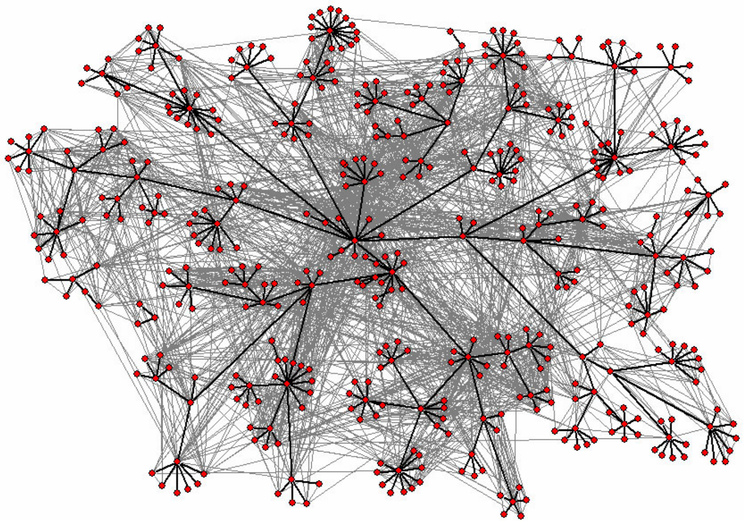
- Web as a "graph"
  - Pages = nodes, hyperlinks = edges
    - Directed graph
    - High linkage:
      - □ 10-20 links per page on average
      - □ Power-law degree distribution



Web Graph Structure



- Network as a "graph"
  - Users/Items=nodes, relations=edges
    - Social networks (facebook, weibo...)
    - Research publication networks (dblp)



Web Structure Terminology (1)

- Web-graph: A directed graph that represents the Web.
- Node: Each Web page is a node of the Web-graph.
- Link: Each hyperlink on the Web is a directed edge of the Web-graph.
- In-degree: The in-degree of a node, p, is the number of distinct links that point to p.
- Out-degree: The out-degree of a node, p, is the number of distinct links originating at p that point to other nodes.

Web Structure Terminology (2)

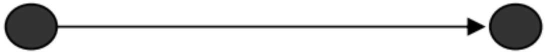
- Directed Path: A sequence of links, starting from p that can be followed to reach q.

- Shortest Path: Of all the paths between nodes p and q, which has the shortest length, i.e. number of links on it.

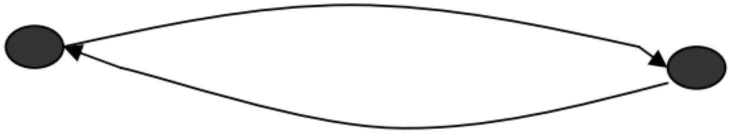
- Diameter: The maximum of all the shortest paths between a pair of nodes p and q, for all pairs of nodes p and q in the Web-graph.

Interesting Web Structures

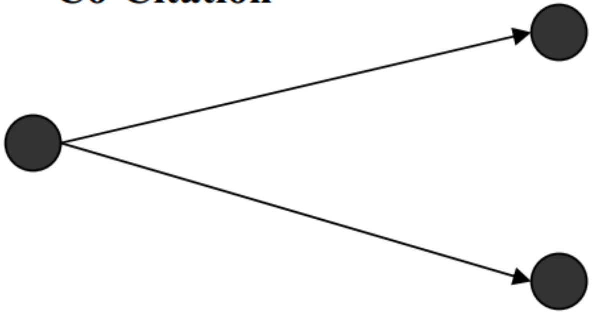
Endorsement



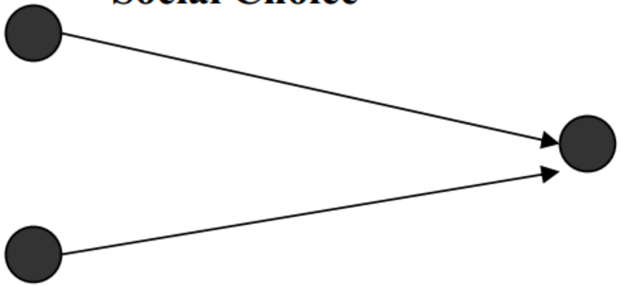
Mutual Reinforcement



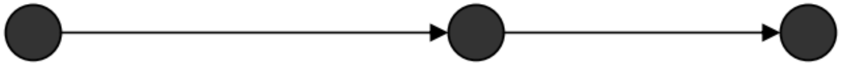
Co-Citation



Social Choice



Transitive Endorsement



Web Structure Mining

- Generate *structural summary* about the Web site and Web page
  - Hierarchy of hyperlinks in the website and its structure.
- Finding information about web pages
  - Retrieving information about the relevance and the quality of the web page.
  - Finding the authoritative（权威性，可信度） on the topic and content.
- Inference on hyperlinks
  - The web page contains not only information but also hyperlinks, which contains huge amount of annotation.
  - Hyperlink identifies author's endorsement of the other web page

What can the graph tell us?

- Distinguish “important” pages from unimportant ones
  - Page rank
- Discover communities of related pages
  - Hubs and Authorities
- Detect web spam
  - Trust rank

Web Communities

Definition:

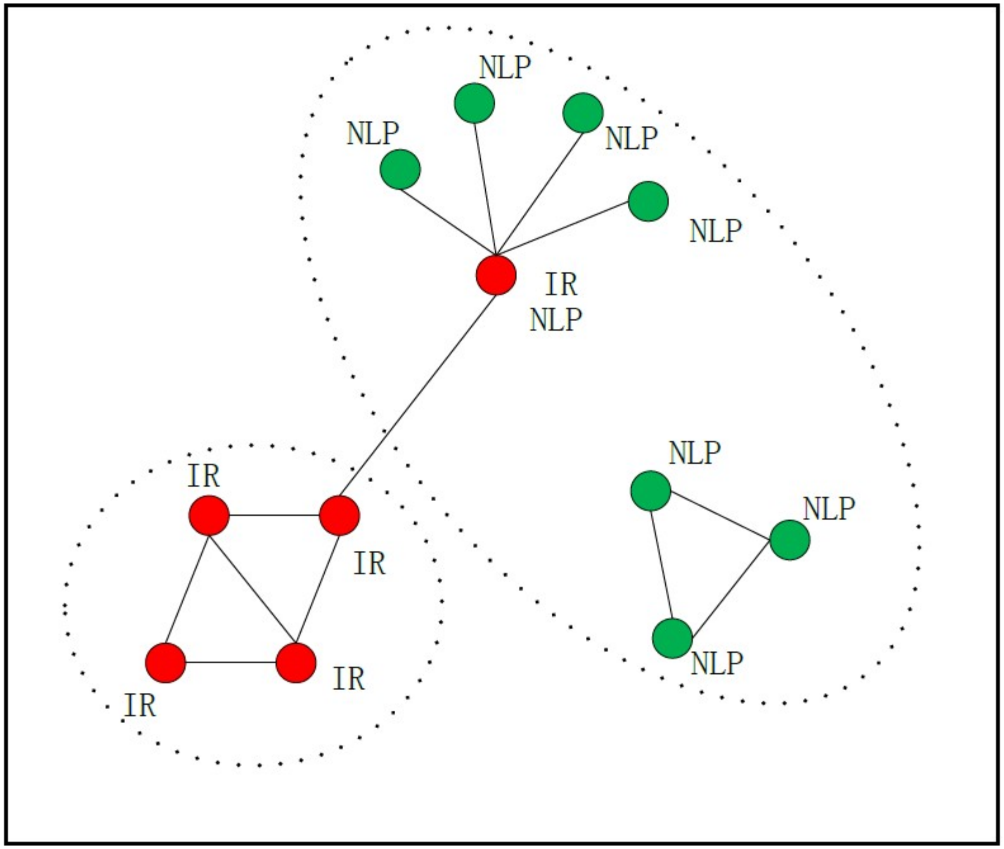
Web communities can be described as a collection of web pages such that each member node has more hyperlinks (in either direction) within the community than outside the community.

Approach:

- Maximal-flow model
- Graph substructure identification

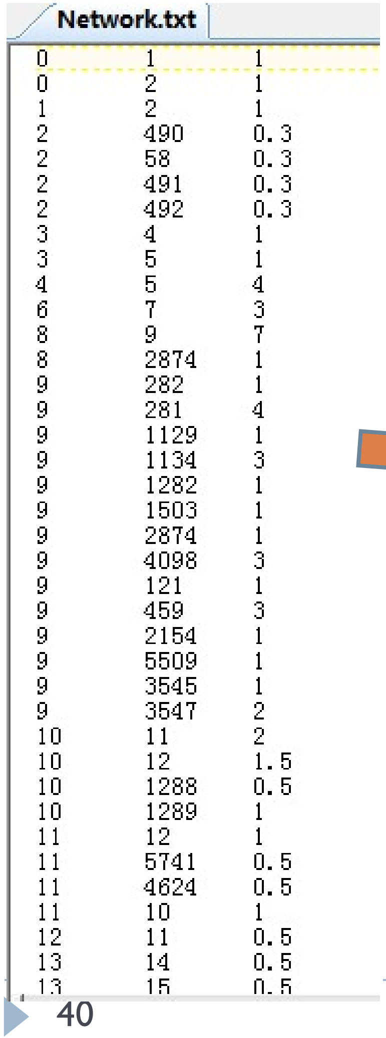
Social Network Mining

Community mining



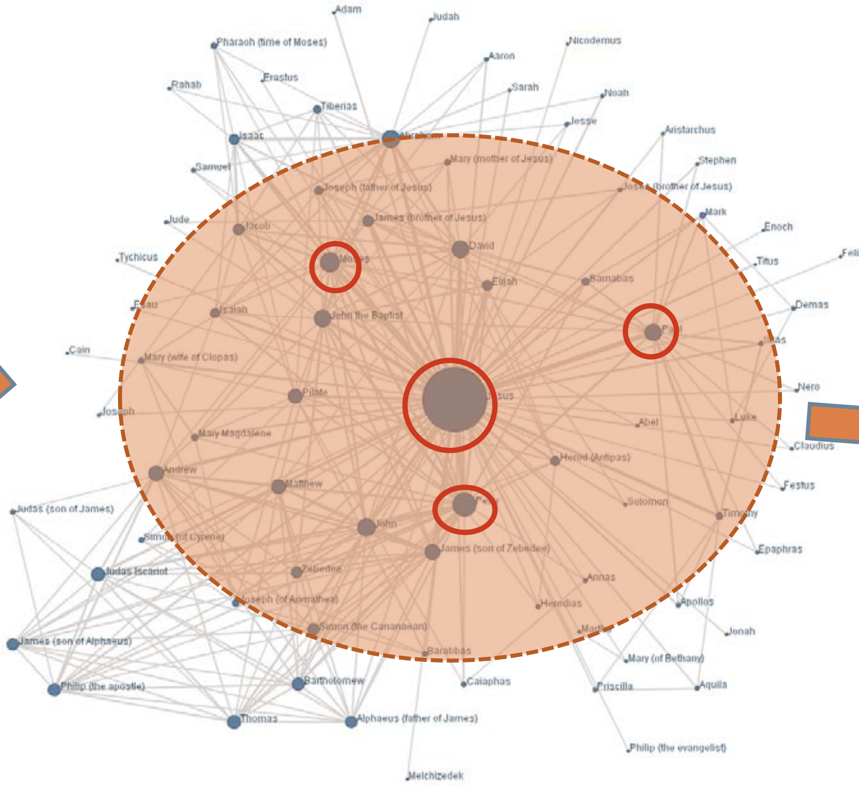
Social Network Mining

Influence analysis

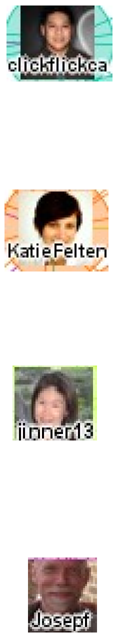


Network.txt

```
0      1      1
0      2      1
1      2      1
2      490    0.3
2      58     0.3
2      491    0.3
2      492    0.3
3      4      1
3      5      1
4      5      4
6      7      3
8      9      7
8      2874   1
9      282    1
9      281    4
9      1129   1
9      1134   3
9      1282   1
9      1503   1
9      2874   1
9      4098   3
9      121    1
9      459    3
9      2154   1
9      5509   1
9      3545   1
9      3547   2
10     11     2
10     12     1.5
10     1288   0.5
10     1289   1
11     12     1
11     5741   0.5
11     4624   0.5
11     10     1
12     11     0.5
13     14     0.5
13     15     0.5
```



最有影响力用户



Web Mining Topics

- Content mining
- Structure mining
- Web usage mining

Web Usage Mining

Navigation Patterns

- Examples:

  70% of users who accessed /company/product2 did so by starting at /company and proceeding through /company/new, /company/products and company/product1

  80% of users who accessed the site started from /company/products

  65% of users left the site after four or less page references

Web Usage Mining

Sequential Patterns

Customer	Transaction Time	Purchased Items
John	6/21/05 5:30 pm	Beer
John	6/22/05 10:20 pm	Brandy
Frank	6/20/05 10:15 am	Juice, Coke
Frank	6/20/05 11:50 am	Beer
Frank	6/20/05 12:50 am	Wine, Cider
Mary	6/20/05 2:30 pm	Beer
Mary	6/21/05 6:17 pm	Wine, Cider
Mary	6/22/05 5:05 pm	Brandy

Sequential Patterns with Support >= 40%	Supporting Customers
(Beer) (Brandy)	John, Frank
(Beer) (Wine, Cider)	Frank, Mary

Web Usage Mining

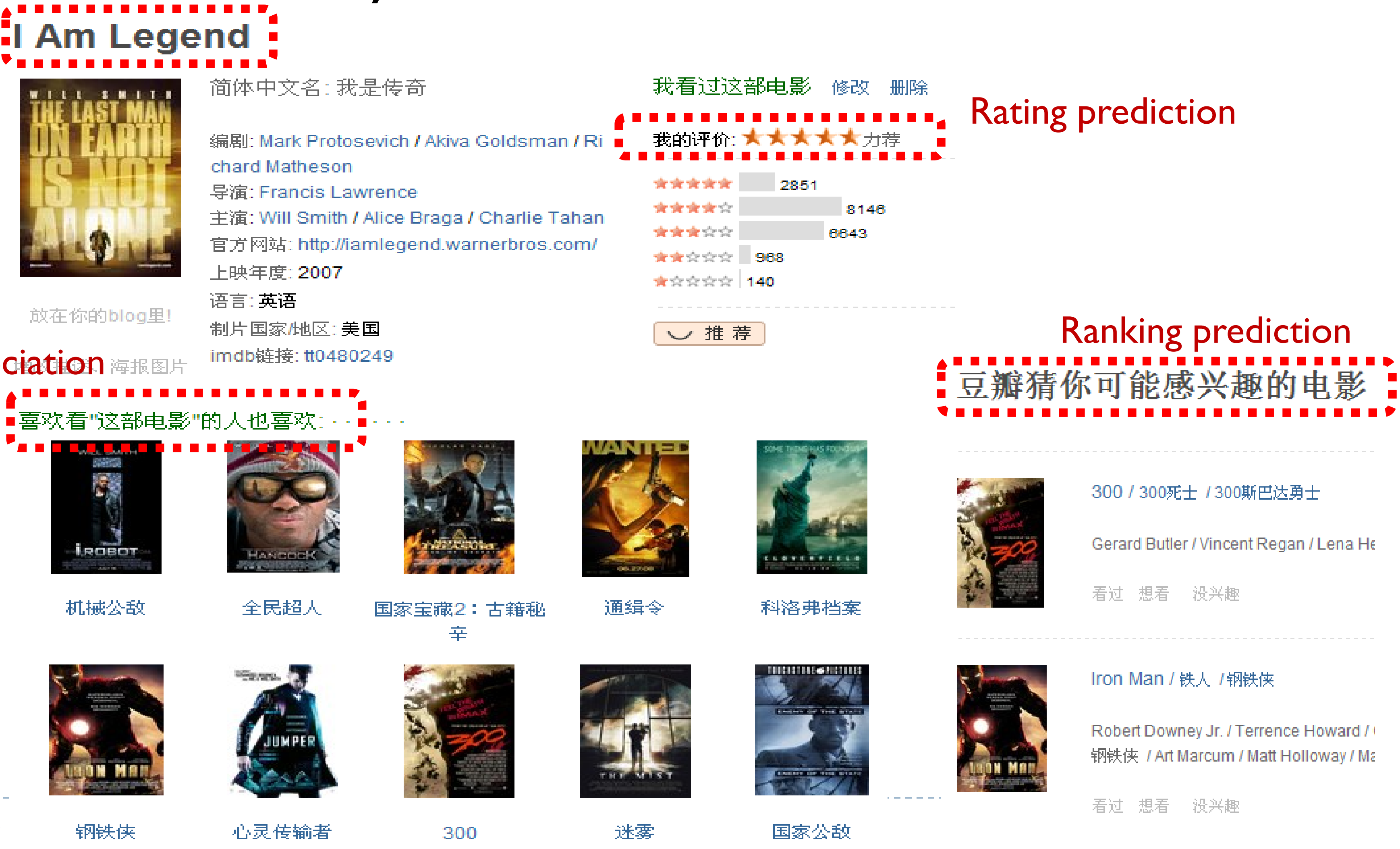
Association Rules

Example:

- 60% of users who placed an online order in /company/product1 also placed an order in /company/product4 within 15 days

Web Usage Mining

Recommender Systems



I Am Legend

简体中文名: 我是传奇

编剧: Mark Protosevich / Akiva Goldsman / Richard Matheson

导演: Francis Lawrence

主演: Will Smith / Alice Braga / Charlie Tahan

官方网站: http://iamlegend.warnerbros.com/

上映年度: 2007

语言: 英语

制片国家/地区: 美国

imdb链接: tt0480249

放在你的blog里!

海报图片

我看过这部电影 修改 删除

我的评价: ★★★★★ 力荐

★★★★★	2851
★★★★☆	8146
★★★☆☆	6643
★★☆☆☆	968
★☆☆☆☆	140

推 荐

Rating prediction

Association Rule

喜欢看"这部电影"的人也喜欢: · · · · · ·

机械公敌 全民超人 国家宝藏2：古籍秘辛 通缉令 科洛弗档案

钢铁侠 心灵传输者 300 迷雾 国家公敌

Ranking prediction

豆瓣猜你可能感兴趣的电影

300 / 300死士 / 300斯巴达勇士

Gerard Butler / Vincent Regan / Lena He

看过 想看 没兴趣

Iron Man / 铁人 / 钢铁侠

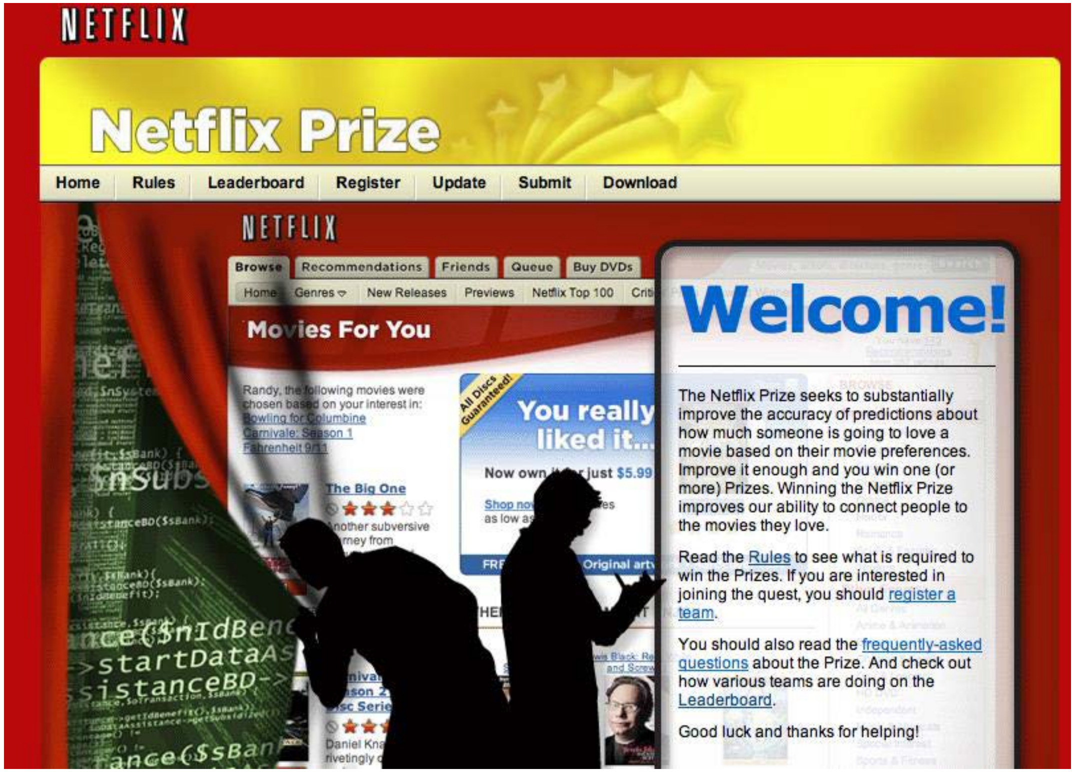
Robert Downey Jr. / Terrence Howard /

钢铁侠 / Art Marcum / Matt Holloway / Ma

看过 想看 没兴趣

Web Usage Mining

Recommender Systems



The Netflix Prize seeks to substantially improve the accuracy of predictions about how much someone is going to love a movie based on their movie preferences. Improve it enough and you win one (or more) Prizes. Winning the Netflix Prize improves our ability to connect people to the movies they love.

Read the Rules to see what is required to win the Prizes. If you are interested in joining the quest, you should register a team.

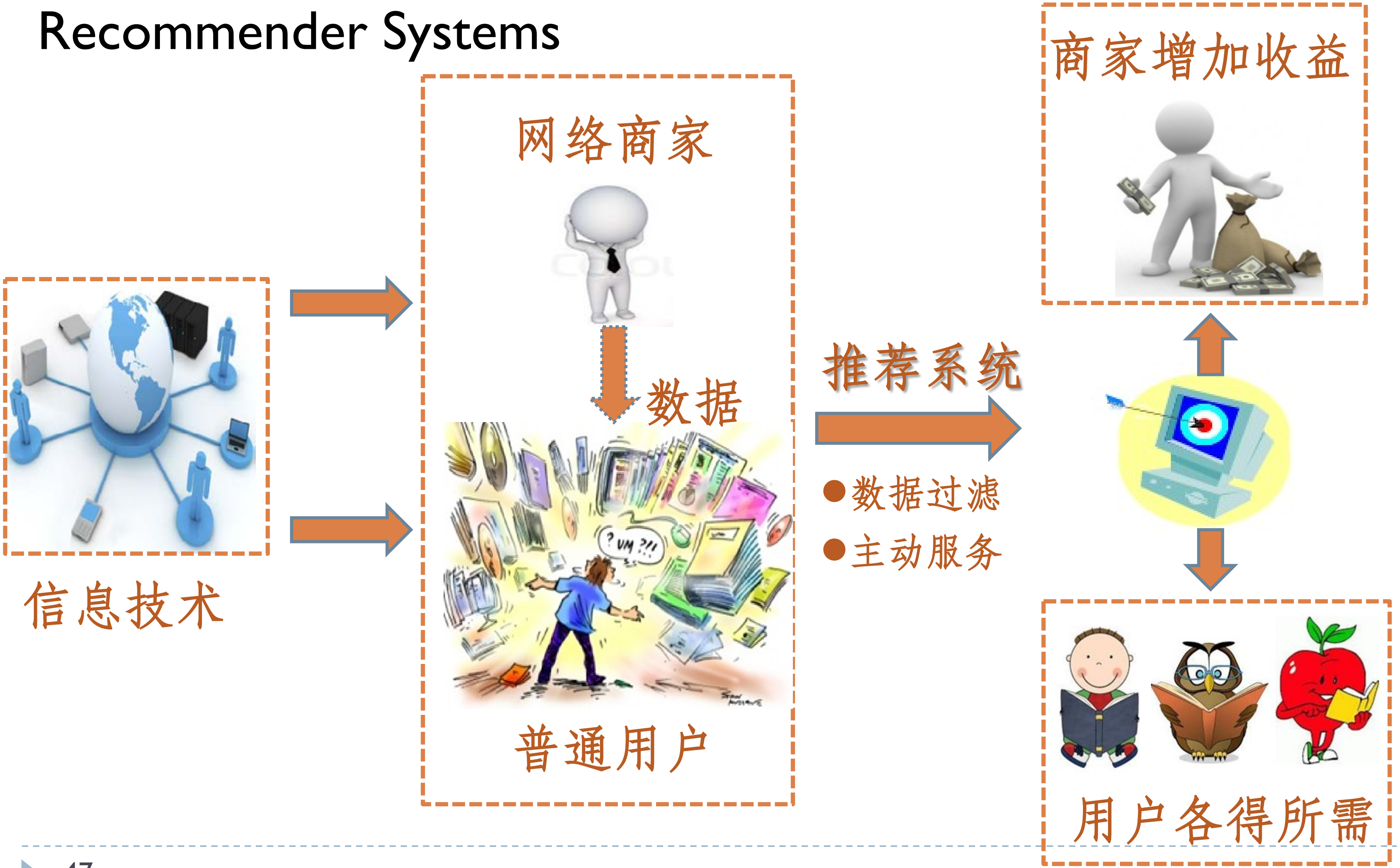
You should also read the frequently-asked questions about the Prize. And check out how various teams are doing on the Leaderboard.

Good luck and thanks for helping!

Challenge: to improve the accuracy of movie preference predictions Netflix $1m Prize.

Web Usage Mining

Recommender Systems



信息技术

网络商家

数据

普通用户

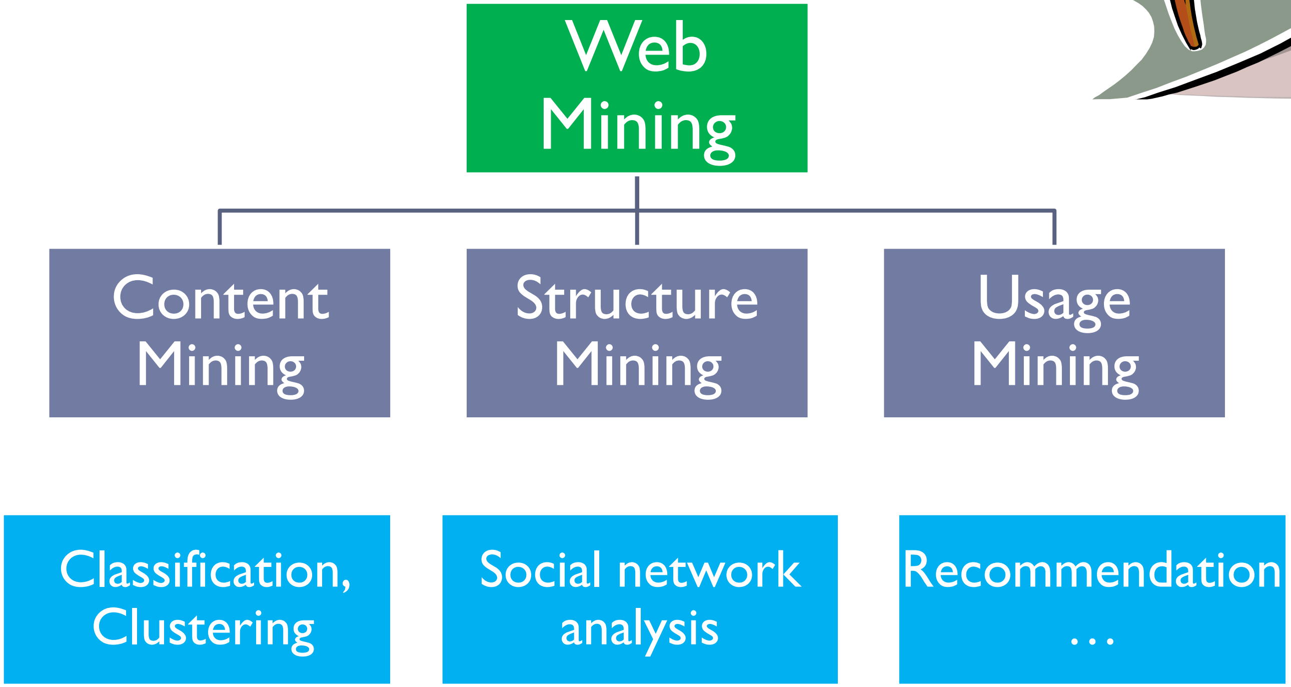
推荐系统

- ●数据过滤
- ●主动服务

商家增加收益

用户各得所需

Roadmap



Web Mining

Content Mining

Structure Mining

Usage Mining

Classification, Clustering

Social network analysis

Recommendation …

Note: Helpful to combine usage with content and structure

What will we learn?

- We will learn to mine different types of web data:
  - Data is high dimensional
  - Data is a graph
  - Data is labeled / unlabeled

What will we learn?

- We will learn to solve real-world problems:
  - Social network analysis
  - Recommender systems

  …

- We will learn various "tools":
  - Linear algebra (MATRIX analysis)
  - Optimization

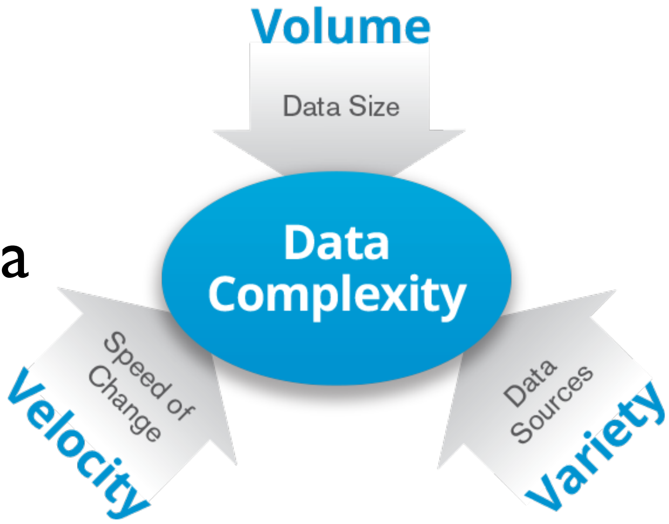
  …

Philosophy

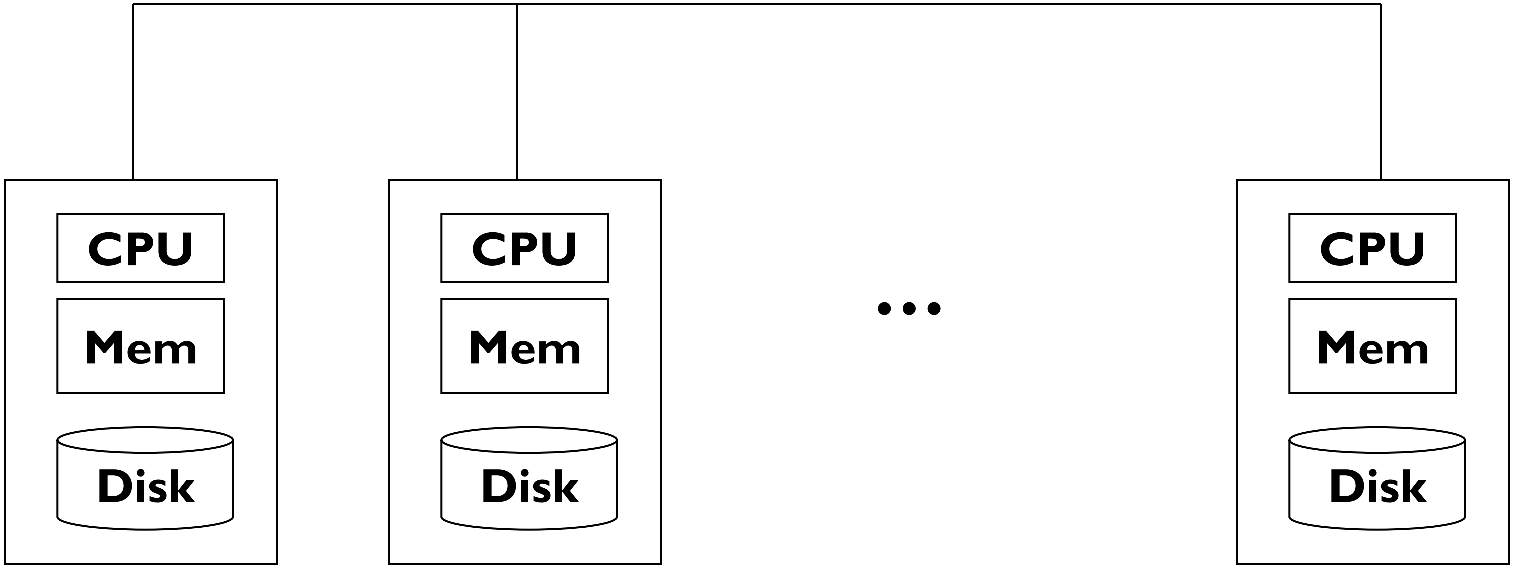
- In many cases, adding more data leads to better results that improving algorithms
  - Netflix
  - Google search
  - Google ads

Challenges of Web Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data
- Data from Multi-Sources



Very Large-Scale Data Mining



Cluster of commodity nodes

Systems Issues

- Web data sets can be very large
  - Tens to hundreds of terabytes
- Cannot mine on a single server!
  - Need large farms of servers
- How to organize hardware/software to mine multi-terabye data sets
  - Without breaking the bank!

Web Mining and Privacy

Public attitude to privacy

- We willingly agree to be tracked, use cookies, fill in forms, answer fairly personal questions on the web
- Different cases regarding medical data

- People don't even know that so much data is being collected about them – e.g. approx. 30GB/day of click-stream data per day at Amazon.com a few years ago

Web Mining and Privacy

What needs to be done?

- Raising public awareness through debate and education
  - Most of the industry doesn’t want this
- Regulations that can prevent/reduce threats
- Good laws on cyber crimes and their enforcement
- Better technology and tools for
  - Security
  - Data analysis
  - Auditing
  - Privacy preserving web mining
  - …

Summary

- Web has been adopted as a critical communication and information medium by a majority of the population
- Web data is growing at a significant rate
- A number of new Computer Science concepts and techniques have been developed
- Many successful applications exist
- Fertile area of research
- Privacy – real debate needed