

Data Mining

Web Mining

A/Prof Heng Tao SHEN

©The University of Queensland, Brisbane Australia

<http://www.itee.uq.edu.au/~shenht>

1

Outline

- Web Content Mining
- Web Usage Mining
- Web Structure Mining

2

What is Web Mining?

- Web data mining - techniques to automatically discover and extract information from Web documents/services
- Web mining research
 - Database (DB)
 - Information retrieval (IR)
 - Machine learning (ML)
 - Natural language processing (NLP)

3

Mining the Web

- Web is an information source for:
 - Information services
 - news, advertisements, consumer information, financial management, education, government, e-commerce, etc.
 - Hyper-link information
 - User Behaviors (access and usage information)
 - Web Site contents and organization
 - Social media

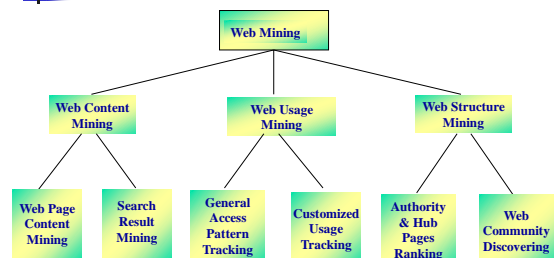
4

Web Mining: challenges

- Searches for
 - Regularity and dynamics of Web contents
 - Web user access patterns
 - Web structures
- Problems
 - The "abundance" problem (rich data but poor information)
 - Limited coverage of the Web: hidden web sources, majority of data in DBMS
 - Dynamic and semi-structured
 - Limited on keyword-oriented search
 - Limited customization to individual users

5

A taxonomy of Web Mining



6

Web content mining

- Discovery of useful information from Web contents / data / documents
 - Web data contents: text, image, audio, video, metadata and hyperlinks.
- Information retrieval view (Structured + Semi-Structured)
 - Assist / Improve information finding
 - Filtering Information to users on user profiles
 - Information extraction

7

Issues

- Developing intelligent tools for IR
 - Finding keywords and key phrases
 - Discovering grammatical rules and collocations
 - Hypertext classification/categorization
 - Extracting key phrases from text documents
 - Learning extraction models/rules
 - Hierarchical clustering
 - Predicting (words) relationship

8

Implementations

- Information Filtering/Categorization
 - Collaborative Filtering (CF)
- Personalized Web agents
 - Web Wrappers

9

Information Filtering/Categorization

- Using various information retrieval techniques and characteristics of open hypertext Web documents to automatically retrieve, filter, and categorize them.
 - **HyPursuit**: uses semantic information embedded in link structures and document content to create cluster hierarchies of hypertext documents, and structure an information space
 - **BO (Bookmark Organizer)**: combines hierarchical clustering techniques and user interaction to organize a collection of Web documents based on conceptual information

10

HyPursuit: similarity functions for web pages with hyperlinks

The hyperlink similarity between two Hypertext document:

$$S_{ij} = W_d \cdot S_{ij}^{dec} + W_a \cdot S_{ij}^{anc} + W_s \cdot S_{ij}^{spl}$$

Where Common Descendants: $S_{ij}^{dec} = \sum_{x \in \text{common}} \frac{1}{2^{(spl_x^i + spl_x^j)}}$

Common Ancestors:

Shortest path length between documents: $S_{ij}^{anc} = \sum_{x \in \text{common}} \frac{1}{2^{(spl_x^i + spl_x^j)}}$

W_d , W_a , and W_s are damping factors for normalization.

$$S_{ij}^{spl} = \frac{1}{2^{(spl_x^i)}} + \frac{1}{2^{(spl_x^j)}}$$

spl_{xy} ≡ length of a shortest path between d_x and d_y .

spl_{xy}^z ≡ length of a shortest path between d_x and d_y not travelling d_z .

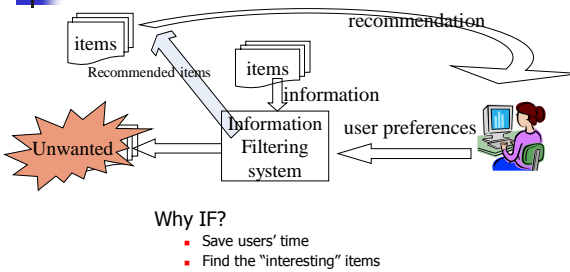
11

How do we find similar Web pages?

- Content based approach
- Structure based approach
- Combing both content and structure approach

12

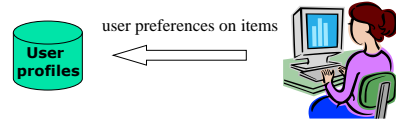
Information Filtering



13

Assumption

- Data about users' preferences can be collected.



14

Definition of CF problem

Given a dataset D as a tuple $\langle U, I, O_{ij} \rangle$

Where,

U_i identifies the i -th user of the system,

I_j identifies the j -th items of the system,

O_{ij} represents the i -th user's opinion on the j -th item

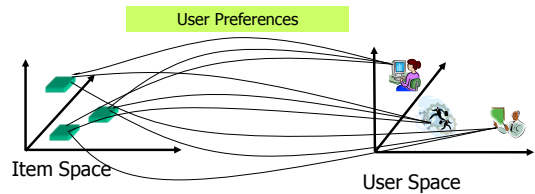
Find a list of k recommended items for each user.

15

A mapping of two high-dimensional spaces

Q1: For a given kind of items, what kind of customers would like it?

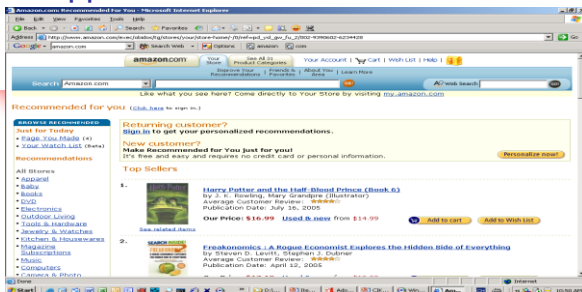
Q2: For certain type of customers, what kind of items do they like?



16

Applications of CF

E-commerce



Digital library [Seikyung Jung, CIKM04]
Recommend TV show [Kamal Ali, KDD04]

17

Challenges to CF

- Prediction Precision**
- Scalability:** the number of users and items increase dramatically, how is the performance of the algorithm?
- Robustness:** given some degree of noise in the data, how is the algorithm to provide accurate prediction?
- Sparsity:** the user-item rating matrix is very sparse
- Cold start:** how to make recommendations for new users or new items

18

Web usage mining

- Web Log Mining
 - Pre-processing
 - Pattern mining
 - Pattern analysis

19

Applications

- Target potential customers for E-commerce
- Enhance the quality and delivery of Internet information services
- Improve web server performance (Load Balancing)
- Identify potential prime advertisement locations
- Facilitates personalization/adaptive sites
- Improve site design
- Fraud/intrusion detection
- Predict user's actions (allows pre-fetching)

20

Outcomes

- Association rules
 - Find pages that are often viewed together
- Clustering
 - Cluster users based on browsing patterns
 - Cluster pages based on content
- Classification
 - Relate user attributes to patterns

21

Phases

- Three distinctive phases:
 - Pre-processing
 - Pattern discovery
 - Pattern analysis

22

Phase 1: pre-processing

- Converts the raw data into the data abstraction necessary for the further applying the data mining algorithm
 - Mapping the log data into **relational tables** before an adapted data mining technique is performed.
 - **Using the log data directly** by utilizing special pre-processing techniques.

23

Raw data – Web log

- **Click stream**: a sequential series of page view request
- **User session**: a delimited set of user clicks (click stream) across one or more Web servers.
- **Server session (visit)**: a collection of user clicks to a single Web server during a user session.
- **Episode**: a subset of related user clicks that occur within a user session.

24

Phase 2: pattern discovery

- Pattern discovery uses techniques such as statistical analysis, association rules, clustering, classification, sequential pattern, dependency modeling.

25

Phase 3: pattern analysis

- A process to gain Knowledge about how visitors use Website in order to
 - Prevent disorientation and help designers to place important information/functions exactly where the visitors look for and in the way users need it.
 - Build up adaptive Website server

26

Web structure mining

- To discover the link structure of the hyperlinks at the inter-document level to generate structural summary about Websites and Web pages.
 - Direction 1: based on the hyperlinks, categorizing the Web pages and generated information.
 - Direction 2: discovering the structure of Web document itself.
 - Direction 3: discovering the nature of the hierarchy or network of hyperlinks in the Website of a particular domain.

27

Applications

- Web pages categorization/ranking
- Communities discovery
- Schema discovery in semi-structured environment

28

Well-known methods

- HITS (Topic distillation)
- PageRank (Ranking web pages used by Google)
- Algorithms in cyber-community

29

HITS: Hyperlink Induced Topic Search

- View Web as a directed graph
- Assumption: if document A has hyperlink to document B, then the author of document A thinks that document B contains valuable information
- Concerned with the identification of the most authoritative, or definitive, Web pages on a broad-topic
- A purely link structure-based computation, ignoring the textual content

30

HITS: Hubs and Authority

- It determines two values for a page
- **Hub**: the value of its links to other pages
- **Authority**: the value of the content of the page
- **Mutual reinforcing relationship**
 - An authority value is computed as the sum of the scaled hub values that point to that page
 - A hub value is the sum of the scaled authority values of the pages it points to
 - A good authority is a page that is pointed to by many good hubs, while a good hub is a page that points to many good authorities

31

HITS: algorithm

- A **sampling** component, which constructs a focused collection of several thousand Web pages from query results
- A **weight-propagation** component, which determines estimates of hub and authority by an iterative procedure
 - Authority Update: Update each node's Authority score to be equal to the sum of the Hub Scores of each node that points to it
 - Hub Update: Update each node's Hub Score to be equal to the sum of the Authority Scores of each node that it points to
- As the result, pages with highest weights are returned as hubs and authorities for the research topic

32

HITS: drawbacks

- **Limit on narrow topics**
 - Not enough authoritative pages
 - Frequently returns resources for a more general topic
 - Adding a few edges can potentially change scores considerably
- **Topic drifting**
 - Appear when hubs discuss multiple topics

33

PageRank

- Introduced by Brin and Page (1998)
 - Mine hyperlink structure of Web to produce 'global' importance ranking of every web page
- **Assumption**: Highly linked pages are more 'important' than pages with a few links
- A page has a high rank if the **sum of the ranks of its back-links** is high
- Google utilizes several factors to rank the results
 - proximity, anchor text, PageRank, query, etc

34

PageRank: main idea

- PageRank results from a "ballot" among all the other pages on Web about how important a page is
- A hyperlink to a page counts as a vote of support
- PageRank of a page is defined recursively and depends on the number and PageRank metric of all pages that link to it ("incoming links")
- A page that is linked to by many pages with high PageRank receives a high rank itself
- If there are no links to a web page there is no support for that page

35

PageRank: algorithm

- PageRank is a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page
- It is assumed that the distribution is evenly divided among all documents in the collection at the beginning of the computational process
- PageRank computations require several passes, called "iterations", through the collection to adjust approximate PageRank values to more closely reflect the theoretical true value

36

HITS vs. PageRank

- Both are iterative algorithms based on the linkage of the documents on the Web
- HITS is executed at query time (the hub and authority scores assigned to a page are query-specific), while PageRank is run at indexing time
- HITS is not commonly used by search engines.
- HITS computes two scores per document, hub and authority, while PageRank computes a single score.
- HITS is processed on a small subset of 'relevant' documents, while PageRank ranks all Web pages.

37

What is Web community?

- A cyber **community** on the Web is a group of web pages sharing a common interest
- Main properties:
 - Pages in the same community should be similar to each other in content
 - The pages in one community should differ from the pages in another community
 - Similar to cluster

38

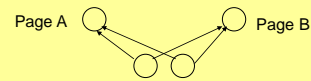
Community discovery

- Discovering Web communities is similar to clustering. So, we must define the **similarity of two pages**.

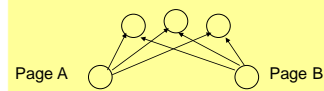
39

Similarity of Web pages

- Co-citation:** the similarity of A and B is measured by the number of pages cite both A and B.



- Bibliographic coupling:** the similarity of A and B is measured by the number of pages cited by both A and B.



40

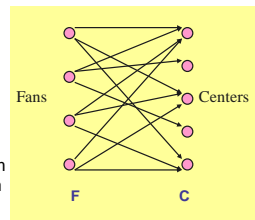
The CT-algorithm

- The method from IBM Almaden Research Center, **Clever** search engine
- They call their method **Communities Trawling (CT)**
- They implemented it on the graph of 200 millions pages, it worked very well

41

CT: main idea

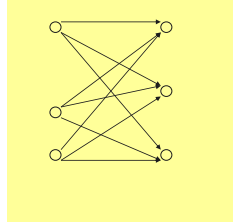
- Definition of Web community
 - Dense directed bipartite sub graphs**
- Bipartite graph: Nodes are partitioned into two sets, F and C
 - Every directed edge in the graph is directed from a node in F to a node in C
 - Dense if many of the possible edges between F and C are present



42

Bipartite cores

- Bipartite cores
 - A complete bipartite subgraph with at least i nodes from F and at least j nodes from C
 - i and j are tunable parameters
- Every community have such a core with a certain i and j .



A ($i=3, j=3$) bipartite core

CT: algorithm

- A bipartite core is the identity of a community
- To extract all the communities is to enumerate all the bipartite cores on the web
- Authors invent an efficient algorithm to enumerate the bipartite cores by **iterative pruning**
 - **elimination-generation pruning**

CT: drawbacks

- The bipartite graph cannot suit all kinds of communities
- The density of the community is hard to adjust

Summary

- Web mining
 - Content mining
 - Usage mining
 - Structure mining

References

(HyPursuit) Weiss, R., Velez, B., Sheldon, M., Nemprempe, C., Szilagy, P., Duda, A. and Gifford, D. "HyPursuit: A hierarchical network search engine that exploits contentlink hypertext clustering" in Proc. of the ACM Hypertext'96 (Washington, DC, March, 1996). <http://www.psrp.lcs.mit.edu/ftpdir/papers/>

(CF Approach) Sarwar B. Karypis G., Konstan J., and Riedl J., "Item-Based Collaborative Filtering Recommendation Algorithms", Proceedings of ACM 10th WWW Conference, Hong Kong, May 2001, 285-295.

References

Kosala, R. and Blockeel, H. *Web Mining Research: A Survey*. SIGKDD Explorations, 2(1):1-15, 2000

J. Srivastava, R. Cooley, M. Deshpande, Pang-Ning Tan, *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*, ACM SIGKDD Explorations, Vol. 1, Issue 2, 2000.

(Community Trawling) S.R. Kumar et al., "Trawling Emerging-Cyber-Communities Automatically," Proc. 8th World Wide Web Conf., Elsevier Science, Amsterdam, 1999, pp. 403-415.

(HITS Algorithm) Kleinberg J.M. "Authoritative Sources in a Hyperlinked Environment", Journal of ACM, 46(5), September 1999, pp604-632.

(Page Rank) S. Brin, L. Page. *The Anatomy of a Large-Scale Hypertextual Web Search Engine*. Proceedings of the Seventh World Wide Web Conference, Brisbane, Australia, April 1998.