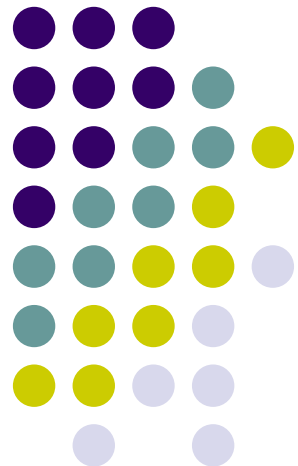
Web Mining : Accomplishments & Future Directions

Jaideep Srivastava
University of Minnesota
USA
srivasta@cs.umn.edu
http://www.cs.umn.edu/faculty/srivasta.html

Mr. Prasanna Desikan's help in preparing these slides is acknowledged



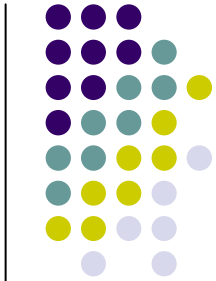
Overview

- Introduction to data mining
  - Data mining process
  - Data Mining techniques
    - Classification
    - Clustering
    - Topic Analysis
    - Concept Hierarchy
    - Content Relevance
- Web mining
  - Web mining definition
  - Web mining taxonomy
- Web Content Mining
  - Definition
  - Pre-processing of content
  - Common Mining techniques
    - Classification
    - Clustering
    - Topic Analysis
    - Concept Hierarchy
    - Content Relevance
  - Applications of Content Mining
- Web Structure Mining
  - Definition
  - Interesting Web Structures
  - Overview of Hyperlink Analysis Methodology
  - Key Concepts
    - PageRank
    - Hubs and Authorities
    - Web Communities
    - Information Scent
  - Conclusions
- Web Usage Mining
  - Definition
  - Preprocessing of usage data
    - Session Identification
    - CGI Data
    - Caching
    - Dynamic Pages
    - Robot Detection and Filtering
    - Transaction Identification
      - Identify Unique Users
      - Identify Unique User transaction

Overview

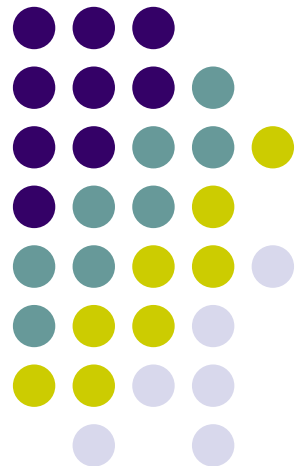
- Web Usage Mining (contd.)
  - Path and Usage Pattern Discovery
  - Pattern Analysis
  - Applications
  - Conclusions
- Web mining applications
  - Amazon.com
  - Google
  - Double Click
  - AOL
  - eBay
  - MyYahoo
  - CiteSeer
  - i-MODE
  - v-TAG Web Mining Server
- Related Concepts
  - Web Visualization
  - Topic Distillation
  - Web Page Categorization
  - Semantic Web Mining
  - Distributed Web Mining
- Web services & Web mining
  - Definitions
  - What they provide
  - Service Oriented Architecture
  - SOAP
  - WSDL
  - UDDI
  - How WM can help WS
  - Web Services Optimization

Overview

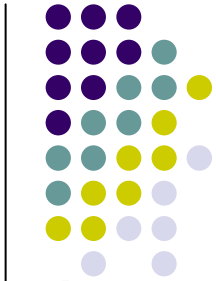


- Research Directions
  - Process Mining
  - Temporal Evolution of the Web
  - Web Services Optimization
  - Fraud at E-tailer
  - Fraud at online Auctioneer
  - Other threats
- Web Mining and Privacy
  - Public Attitude towards Privacy
  - Why this attitude
  - Does understanding implications help
  - What needs to be done
- Conclusions

Introduction to Data Mining

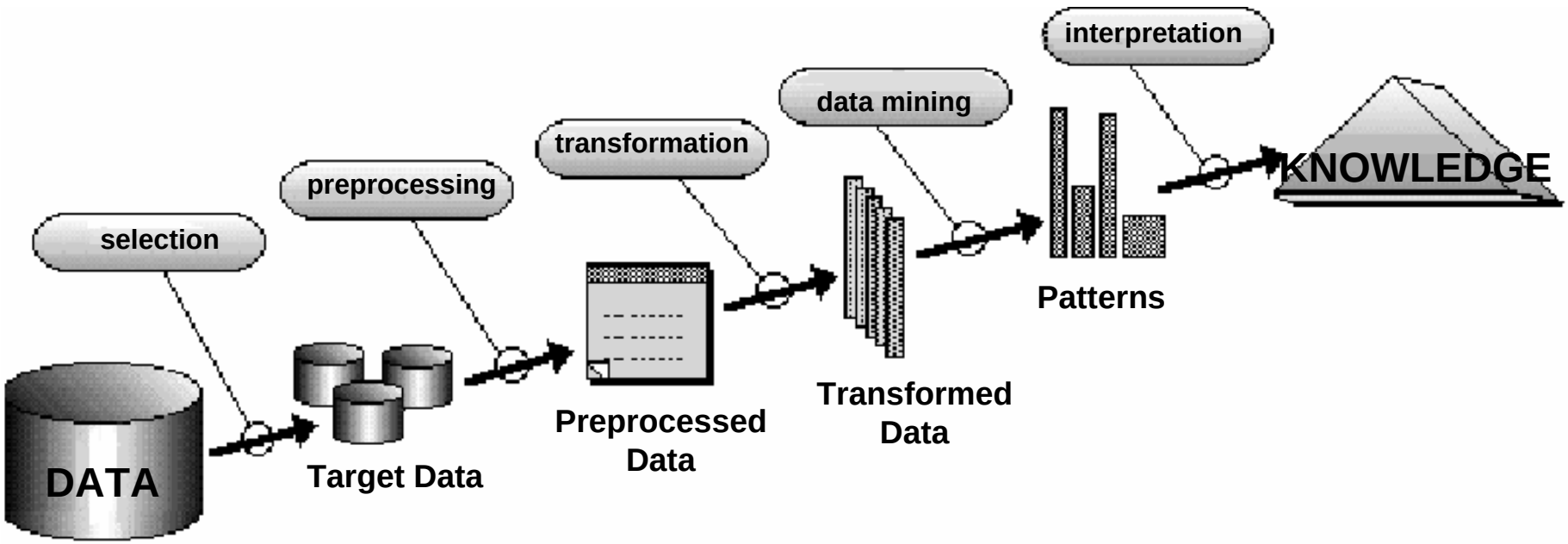


Why Mine Data?



- Computerization and automated data gathering has resulted in extemely large data repositories.
  - E.g., Walmart: 2000 stores, 20 M transactions/day
- Raw Data ⟶ Patterns ⟶ Knowledge
- Scalability issues and desire for more automation makes more traditional techniques less effective:
  - Statistical Methods
  - Relational Query Systems
  - OLAP

The Data Mining (KDD) Process

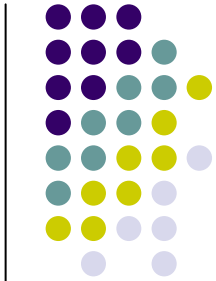


Data Mining Techniques

- Classification
- Clustering
- Association Rules
- Sequential Patterns
- Regression
- Deviation Detection

Primary techniques

Classification: Definition

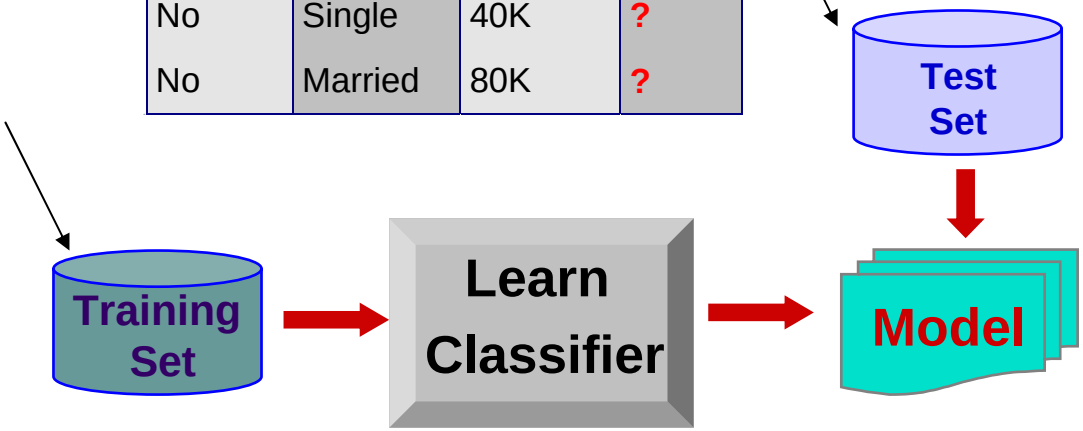


- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification Example

Tid	Refund (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Cheat (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

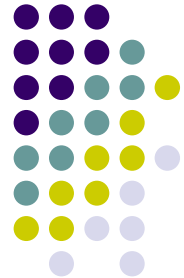
Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification Techniques

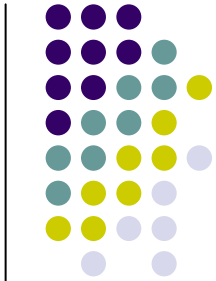
- Decision Tree based Methods
- Rule-based Methods
- Memory based reasoning
- Neural Networks
- Genetic Algorithms
- Naïve Bayes and Bayesian Belief Networks
- Support Vector Machines

What is Cluster Analysis?



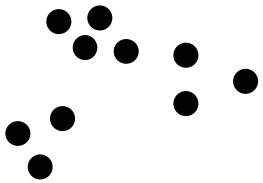
- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.
  - Based on information found in the data that describes the objects and their relationships.
  - Also known as unsupervised classification.
- Many applications
  - **Understanding:** group related documents for browsing or to find genes and proteins that have similar functionality.
  - **Summarization:** Reduce the size of large data sets.
- Web Documents are divided into groups based on a similarity metric.
  - Most common similarity metric is the dot product between two document vectors.

What is not Cluster Analysis?

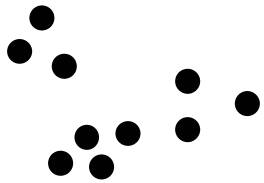


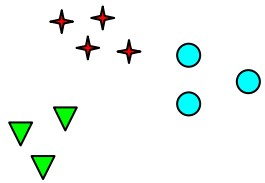
- Supervised classification.
  - Have class label information.
- Simple segmentation.
  - Dividing students into different registration groups alphabetically, by last name.
- Results of a query.
  - Groupings are a result of an external specification.
- Graph partitioning
  - Some mutual relevance and synergy, but areas are not identical.

Notion of a Cluster is Ambiguous

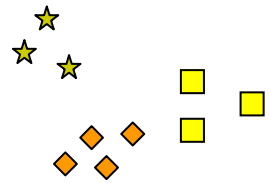


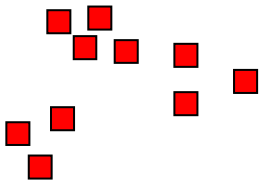
Initial points.



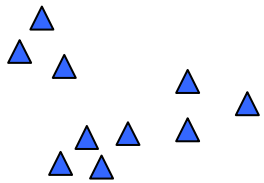


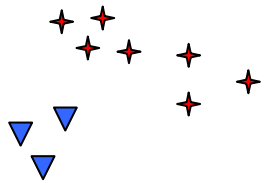
Six Clusters



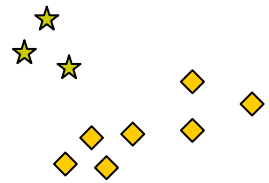


Two Clusters

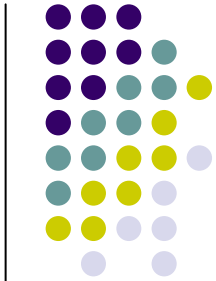




Four Clusters

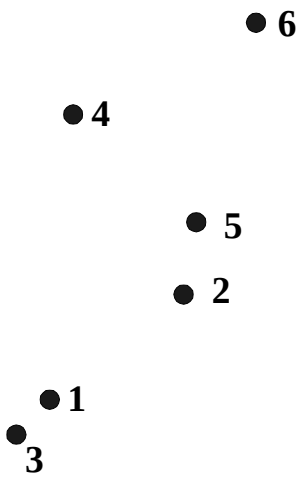


Types of Clusterings

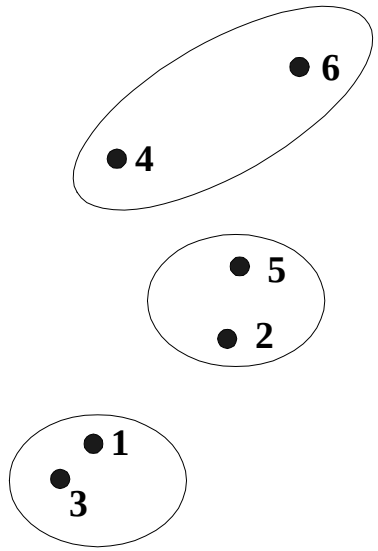


- A clustering is a set of clusters.
- One important distinction is between hierarchical and partitional sets of clusters.
- Partitional Clustering
  - A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset.
- Hierarchical clustering
  - A set of nested clusters organized as a hierarchical tree.

Partitional Clustering

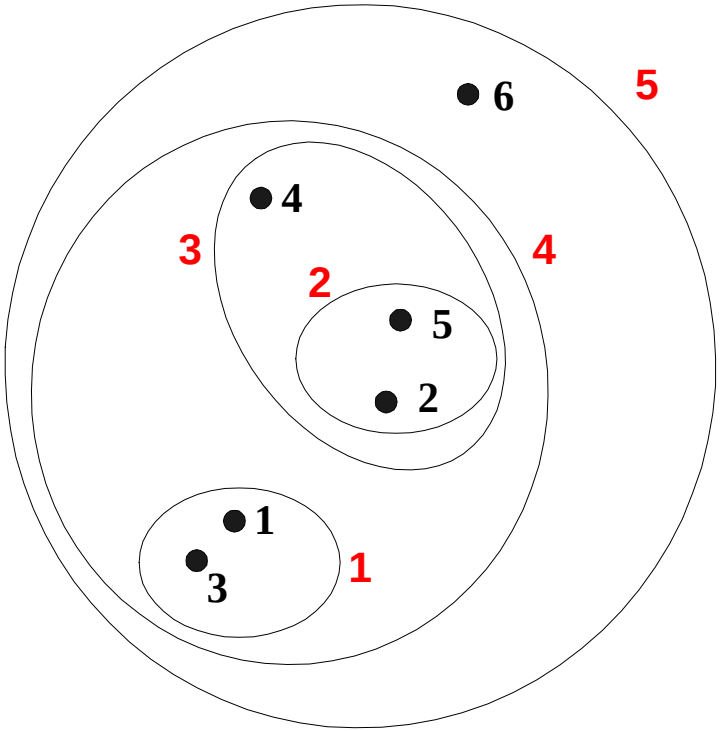


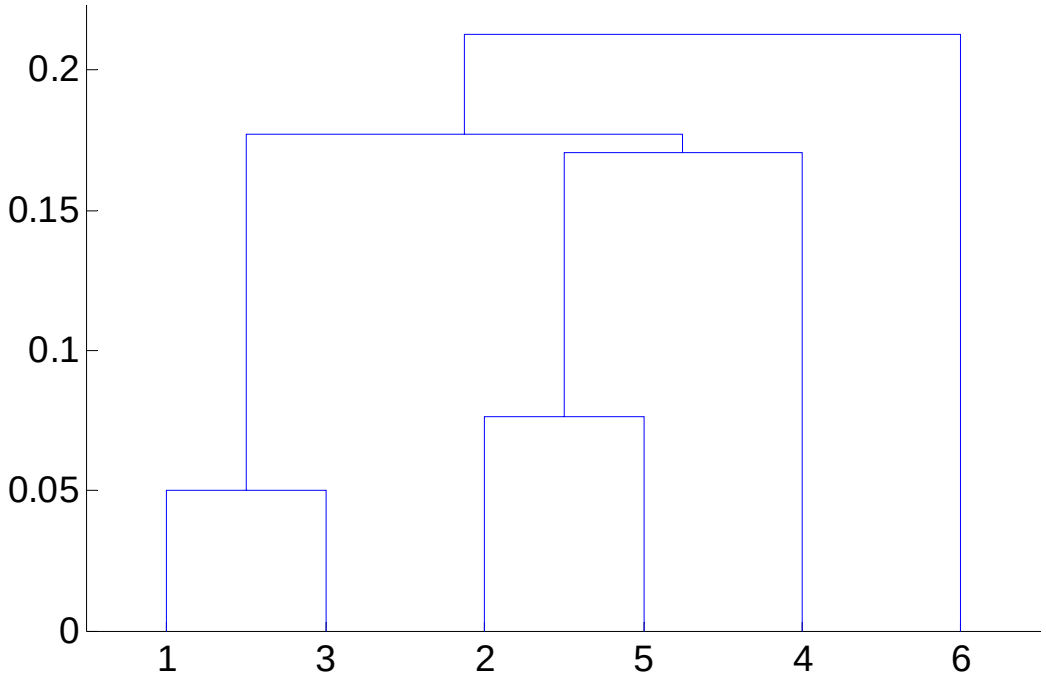
Original Points



A Partitional Clustering

Hierarchical Clustering (agglomerative clustering)





Traditional Hierarchical Clustering

Traditional Dendrogram

Other Distinctions Between Sets of Clusters

- Exclusive versus non-exclusive
  - In non-exclusive clusterings, points may belong to multiple clusters.
  - Can represent multiple classes or ‘border’ points
- Fuzzy versus non-fuzzy
  - In fuzzy clusterings, a point belongs to every cluster with some weight between 0 and 1.
  - Weights must sum to 1.
  - Probabilistic clustering has similar characteristics.
- Partial versus complete.
  - In some cases, we only want to cluster some of the data.

Mining Associations

- **Given a set of records, find rules that will predict the occurrence of an item based on the occurrences of other items in the record**

Market-Basket transactions

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example:

TID	Bread	Milk	Diaper	Beer	Eggs	Coke
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

Definition of Association Rule

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Association Rule: $X \overset{s,c}{\Rightarrow} y$

Support: $s = \frac{\sigma(X \cup y)}{|T|} (s = P(X, y))$

Confidence: $c = \frac{\sigma(X \cup y)}{\sigma(X)} (c = P(y \mid X))$

Goal:

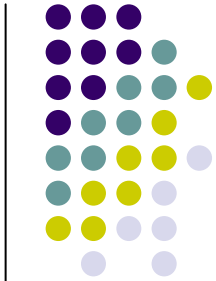
Discover all rules having support $\geq$ *minsup* **and confidence** $\geq$ *minconf* **thresholds.**

Example: $\{Milk, Diaper\} \Rightarrow Beer$

$$s = \frac{\sigma(Milk, Diaper, Beer)}{|T|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(Milk, Diaper, Beer)}{\sigma(Milk, Diaper)} = \frac{2}{3} = 0.67$$

Association Rule Mining



TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Rules:

{Milk,Diaper} → {Beer} (s=0.4, c=0.67)
{Milk,Beer} → {Diaper} (s=0.4, c=1.0)
{Diaper,Beer} → {Milk} (s=0.4, c=0.67)
{Beer} → {Milk,Diaper} (s=0.4, c=0.67)
{Diaper} → {Milk,Beer} (s=0.4, c=0.5)
{Milk} → {Diaper,Beer} (s=0.4, c=0.5)

Observations:

- All the rules above correspond to the same itemset: {Milk, Diaper, Beer}
- Rules obtained from the same itemset have identical support but can have different confidence

Association Rule Mining

- Two-step approach:
  1. Generate all frequent itemsets (sets of items whose support ≥ minsup)
  2. Generate high confidence association rules from each frequent itemset
     - Each rule is a binary partitioning of a frequent itemset
- Frequent itemset generation is the more expensive operation

Sequential Pattern Discovery

- Given a set of objects, with each object associated with its own timeline of events, find rules that predict strong dependencies among different events.

(A B) (C) ➡ (D E)

- Examples:
  - In point-of-sale transaction sequences
    - (Intro_to_visual_C)(C++-Primer)→(Perl_for_dummies)(TCL_TK)
  - In Telecommunication alarm logs:
    - (Inverter_Problem Excessive_Line_Current) (Rectifier_Alarm) → (Fire_Alarm)

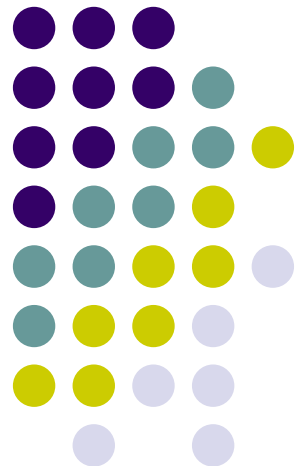
Regression

- Predict a value of a given continuous valued variable based on the values of other variables based on linear or non-linear model of dependency.
- Greatly studied in statistics and neural network fields
- Examples:
  - Predicting sales amount of a new product based on advertising expenses.
  - Time Series prediction of stock market indices

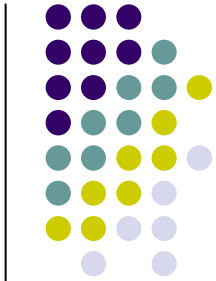
Deviation Detection

- Discovering most significant changes in data from previously measured or normative data.
- Usually categorized separately from other data mining tasks
  - Deviations are often infrequent.
- Modifications of classification, clustering and time series analysis can be used as means to achieve the goal.
- Outlier Detection in Statistics

Web Mining

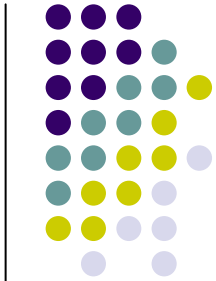


Web Mining



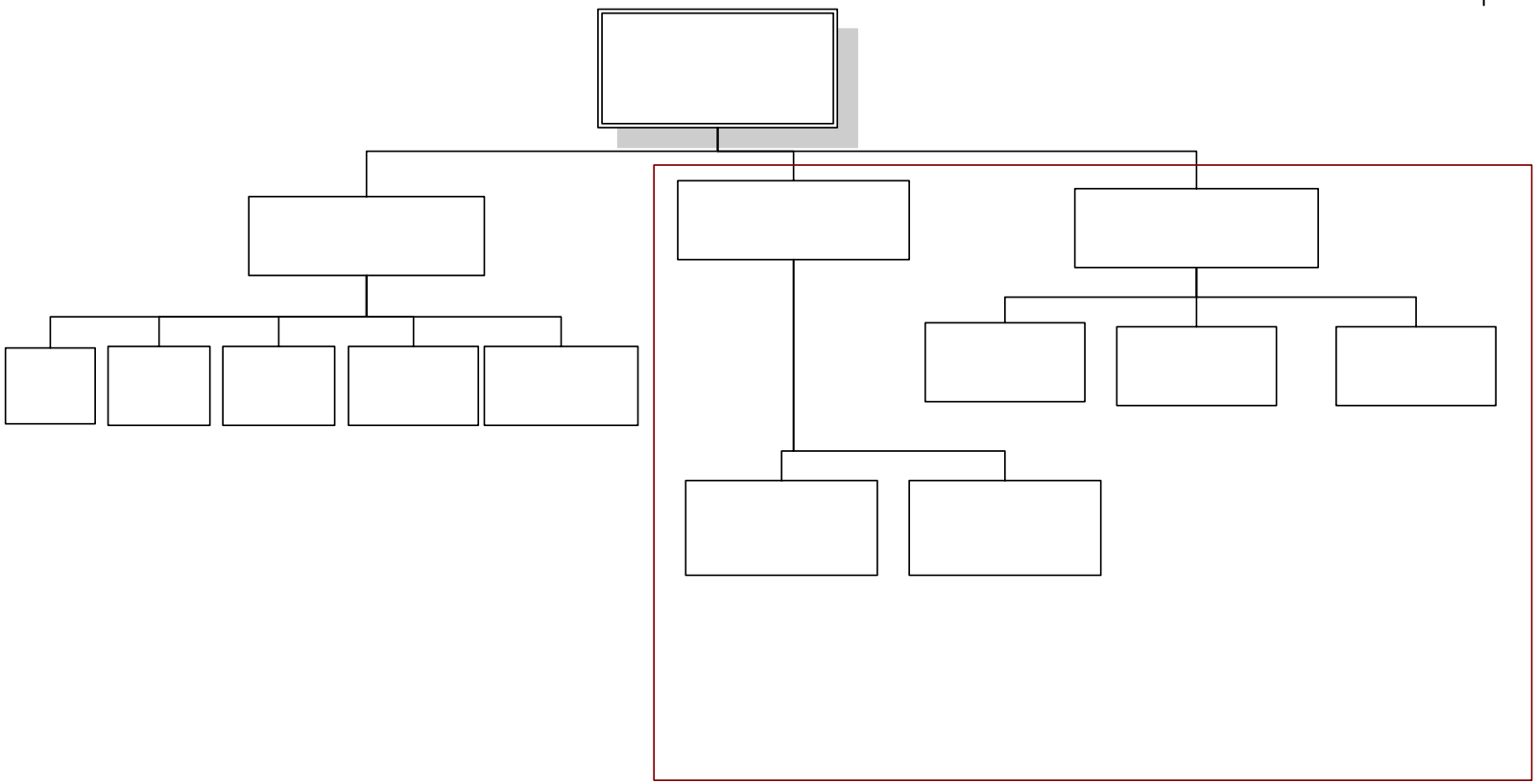
- *Web* is a collection of inter-related files on one or more *Web servers.*
- Web mining is
  - the application of data mining techniques to extract knowledge from Web data
- Web data is
  - Web content – text, image, records, etc.
  - Web structure – hyperlinks, tags, etc.
  - Web usage – http logs, app server logs, etc.

Web Mining – History



- Term first used in [E1996], defined in a ‘task oriented’ manner
- Alternate ‘data oriented’ definition given in [CMS1997]
- 1st panel discussion at ICTAI 1997 [SM1997]
- Continuing forum
  - WebKDD workshops with ACM SIGKDD, 1999, 2000, 2001, 2002, … ; 60 – 90 attendees
  - SIAM Web analytics workshop 2001, 2002, …
- Special issues of DMKD journal, SIGKDD Explorations
- Papers in various data mining conferences & journals
- Surveys [MBNL 1999, BL 1999, KB2000]

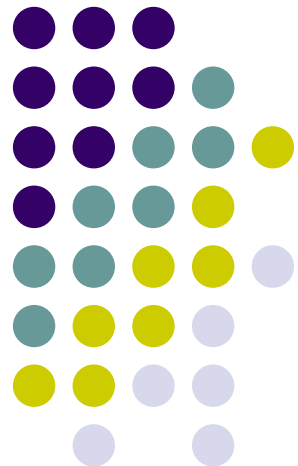
Web Mining Taxonomy



Pre-processing Web Data

- **Web Content**
  - Extract “snippets” from a Web document that represents the Web Document
- **Web Structure**
  - Identifying interesting graph patterns or pre-processing the whole web graph to come up with metrics such as PageRank
- **Web Usage**
  - User identification, session creation, robot detection and filtering, and extracting usage path patterns

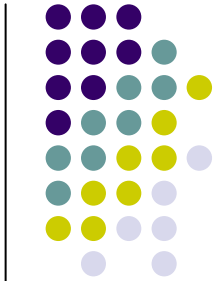
Web Content Mining



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) and natural language processing (NLP).

Pre-processing Content



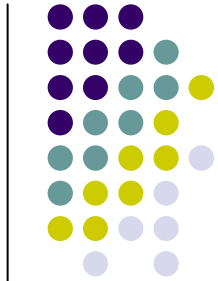
Content Preparation

- Extract text from HTML.
- Perform Stemming.
- Remove Stop Words.
- 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.

Common Mining Techniques



The more basic and popular data mining techniques include:

- Classification
- Clustering
- Associations

The other significant ideas:

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

Document Classification

- “Supervised” technique
- Categories are defined and documents are assigned to one or more existing categories
- The “definition” of a category is usually in the form of a term vector that is produced during a “training” phase
- Training is performed through the use of documents that have already been classified (often by hand) as belonging to a category

Document Clustering

- “Unsupervised” 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

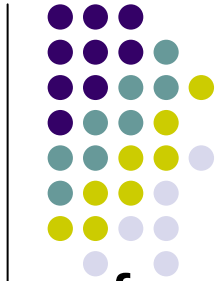
Topic Identification and Tracking

- 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

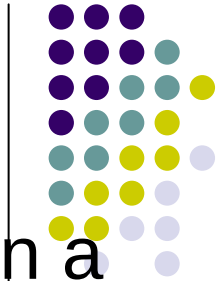
Relevance of Content



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

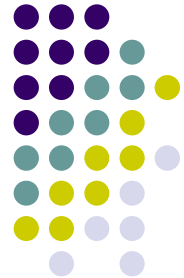
- ✓ Document
- ✓ Query based
- ✓ User Based
- ✓ Role/Task Based

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)

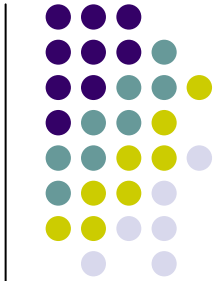
User Based Relevance

- 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

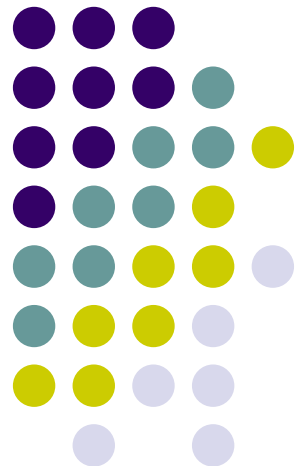
- Similar to User Based Relevance
- Profile is based on a particular role or task, instead of an individual
- Input to profile can come from multiple users

Web Content Mining Applications



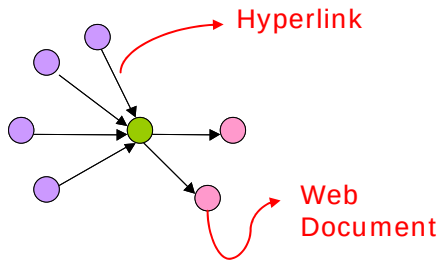
- Identify the topics represented by a Web Documents
- Categorize Web Documents
- Find Web Pages across different servers that are similar
- Applications related to relevance
  - Queries – Enhance standard Query Relevance with User, Role, and/or Task Based Relevance
  - Recommendations – List of top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

Web Structure Mining



What is Web Structure Mining?

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages

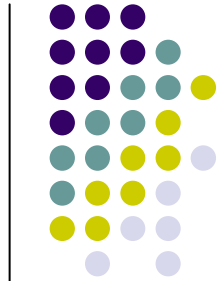


Web Graph Structure

Web Structure Mining can be is the process of discovering structure information from the Web

- This type of mining can be performed either at the (intra-page) document level or at the (inter-page) hyperlink level
- The research at the hyperlink level is also called *Hyperlink Analysis*

Motivation to study Hyperlink Structure



- Hyperlinks serve two main purposes.
  - ✓ Pure Navigation.
  - ✓ Point to pages with authority* on the same topic of the page containing the link.
- This can be used to retrieve useful information from the web.

*** - a set of ideas or statements supporting a topic**

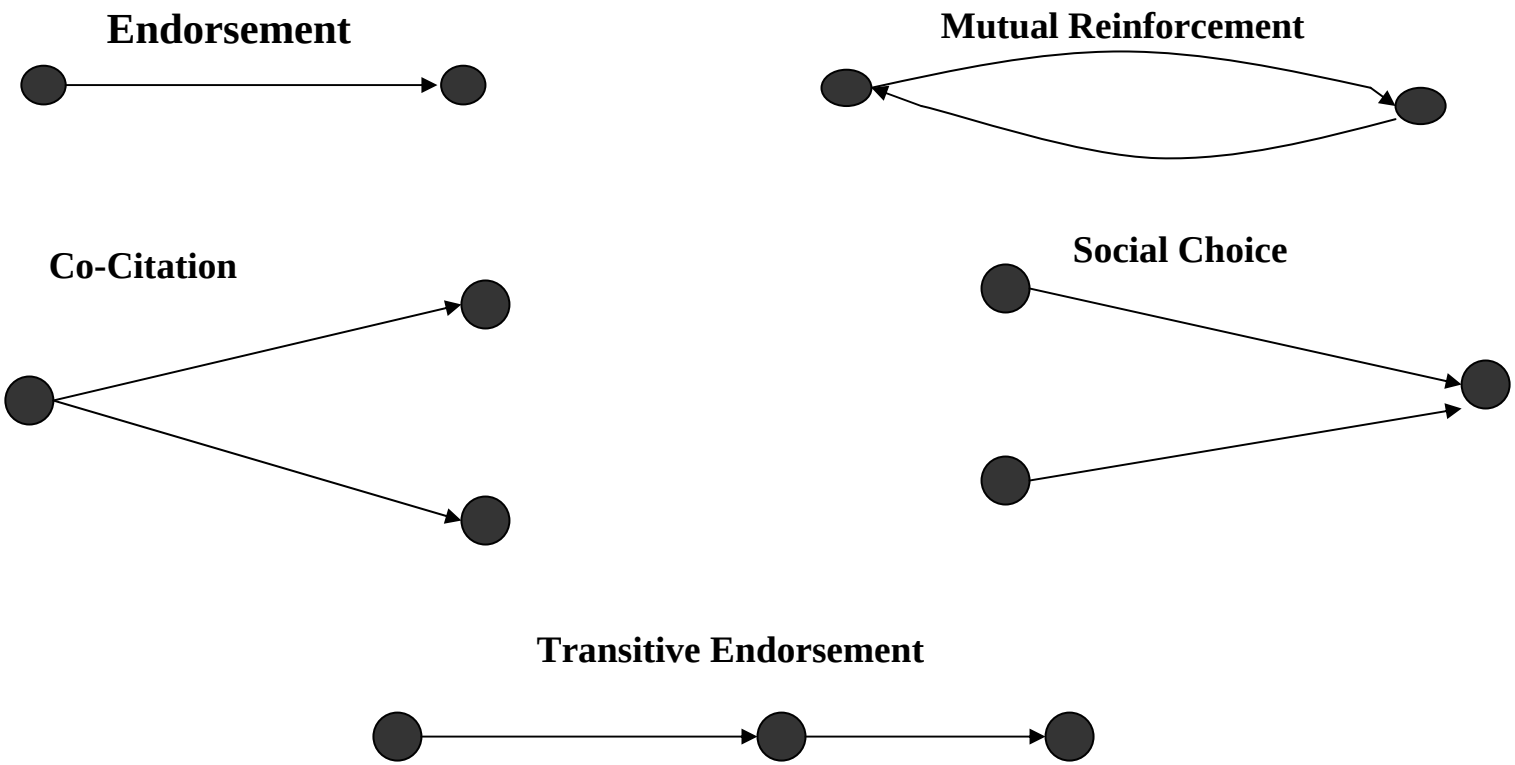
Web Structure Terminology(1)

- ***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.

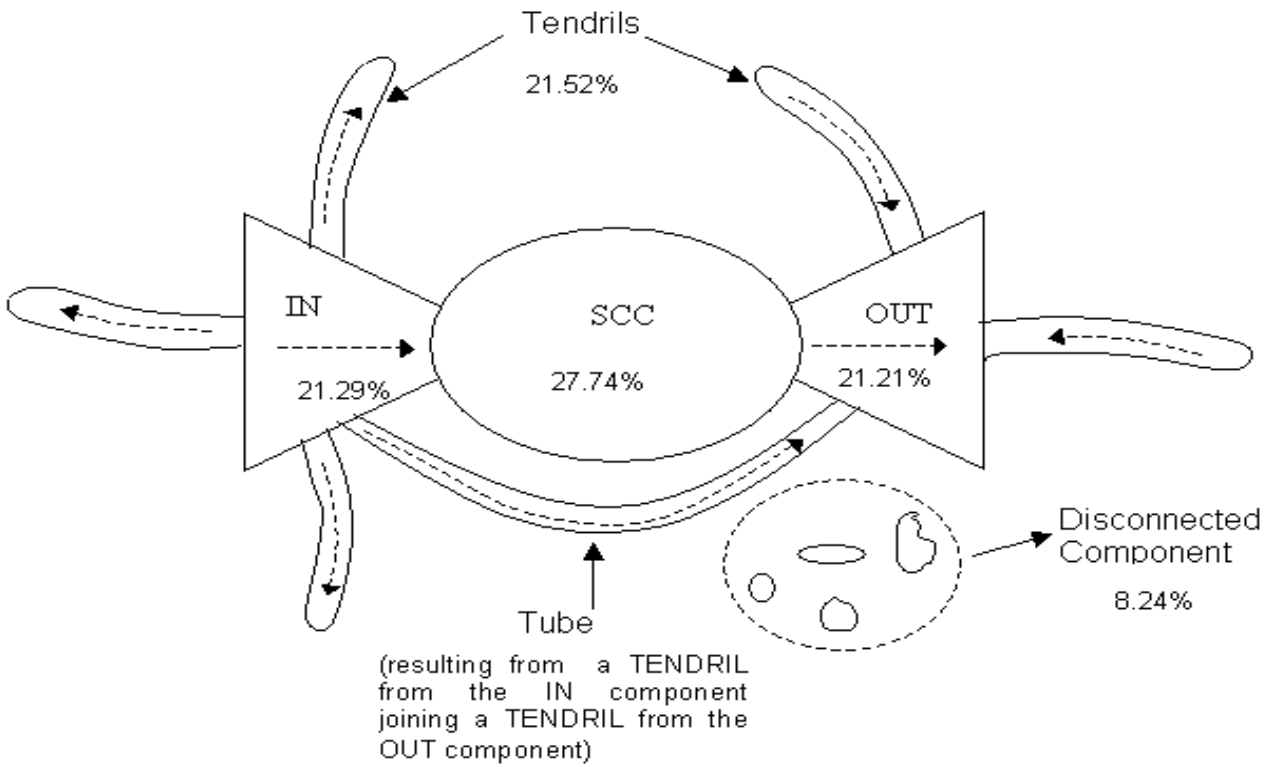
Web Structure Terminology(2)

- ***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 [ERC+2000]



The Bow-Tie Model of the Web [BKM+2000]



Hyperlink Analysis Techniques [DSKT2002]

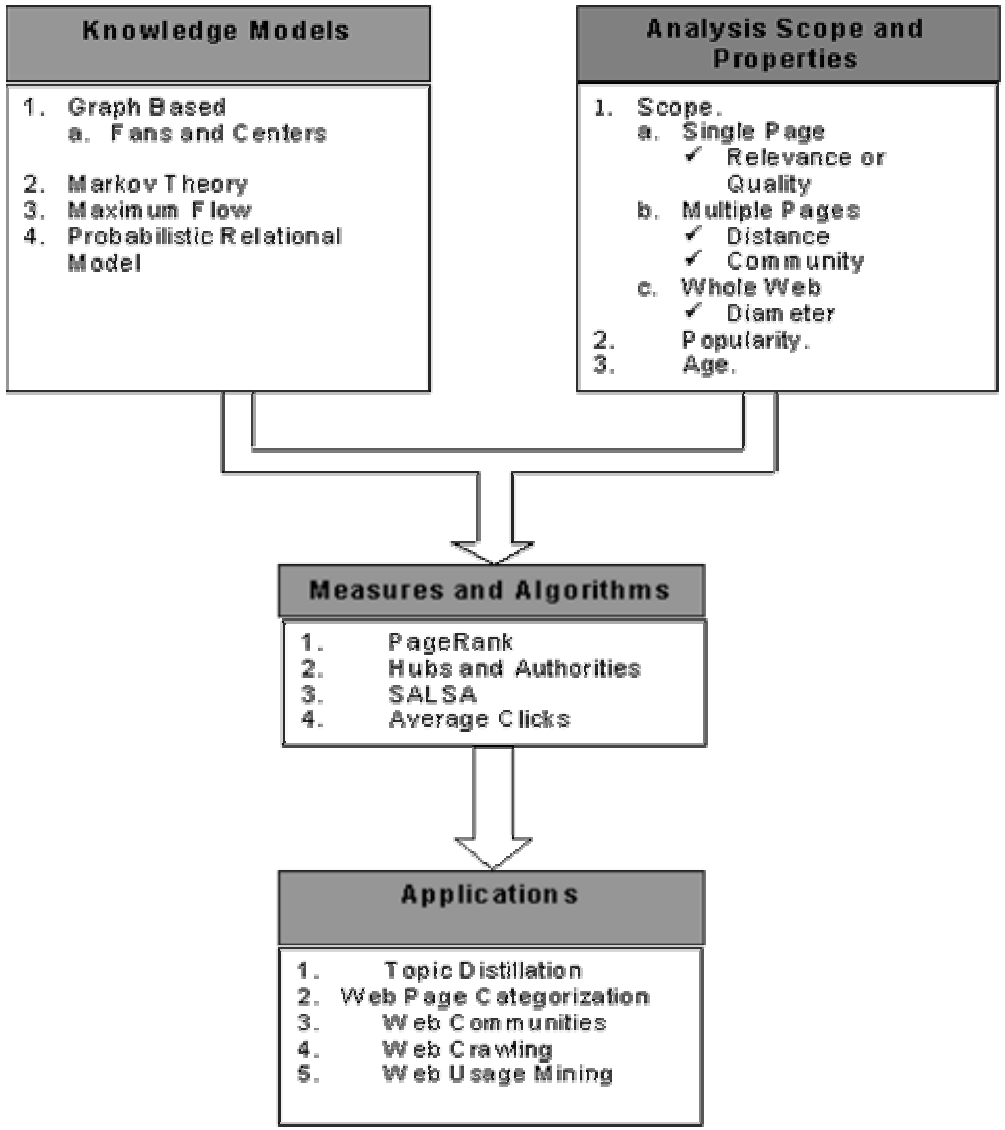
✓***Knowledge Models*:** The underlying representations that forms the basis to carry out the application specific task

✓***Analysis Scope and Properties*:** The scope of analysis specifies if the task is relevant to a single node or set of nodes or the entire graph. The properties are the characteristics of single node or the set of nodes or the entire web

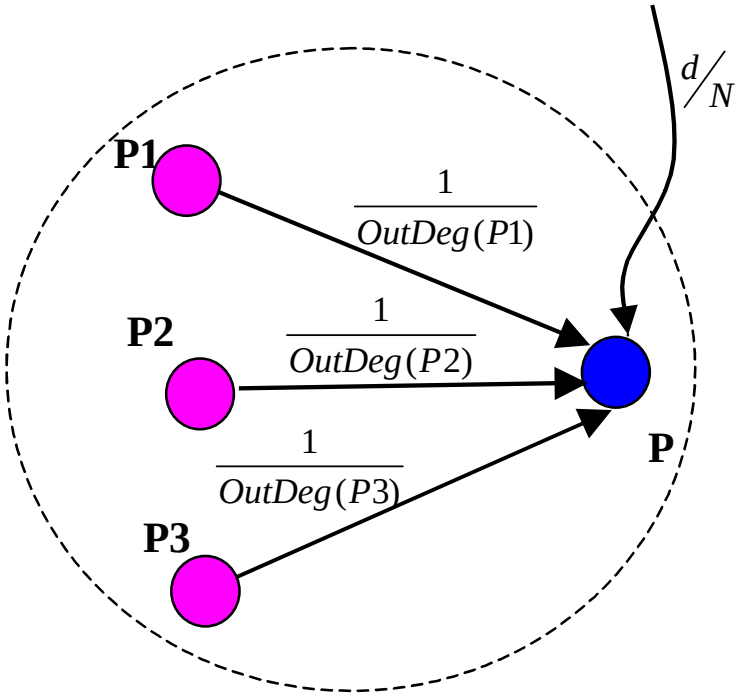
✓***Measures and Algorithms*:** The measures are the standards for the properties such as quality, relevance or distance between the nodes. Algorithms are designed to for efficient computation of these measures

These three areas form the fundamental blocks for building various ***Applications*** based on hyperlink analysis

Hyperlink Analysis Techniques



Google's PageRank [BP1998]



$$PR(P) = d/N + (1-d)\left(\frac{PR(P1)}{OutDeg(P1)} + \frac{PR(P2)}{OutDeg(P2)} + \frac{PR(P3)}{OutDeg(P3)}\right)$$

Key idea

Rank of a web page depends on the rank of the web pages pointing to it

The PageRank Algorithm [BP1998]

Set **PR** $\leftarrow [r_1, r_2, \ldots r_N]$, where $r\text{-}_i$ is some initial rank of page I, and N the number of Web pages in the graph;

$d \leftarrow 0.15$; **D** $\leftarrow [1/N \ldots\ldots 1/N]^T$;

A is the adjacency matrix as described above;

do

$\mathbf{PR_{i+1}} \leftarrow \mathbf{A^T * PR_i}$ **;**

$\mathbf{PR_{i+1}} \leftarrow (1\text{-}d) * \mathbf{PR_{i+1}} + d * \mathbf{D}$**;**

$\delta \leftarrow \| \mathbf{PR_{i+1}} - \mathbf{PR_i} \|_1$

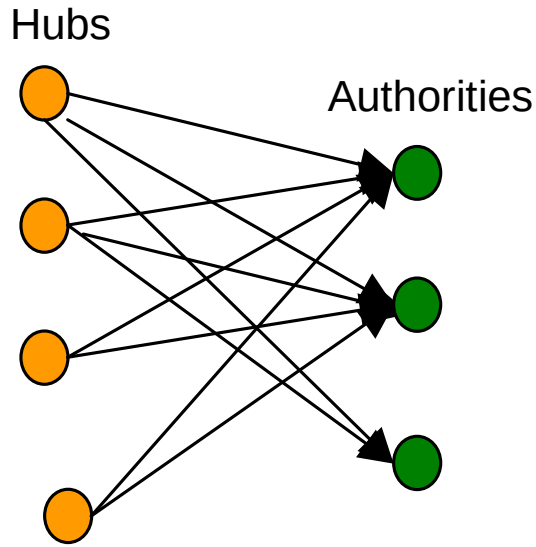
while $\delta < \varepsilon$, where ε is a small number indicating the convergence threshold

return PR.

Hubs and Authorities [K1998]

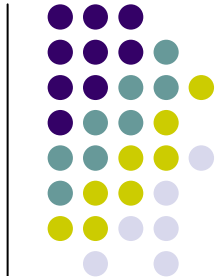
Key ideas

- *Hubs* and *authorities* are ‘fans’ and ‘centers’ in a bipartite core of a web graph
- A good hub page is one that points to many good authority pages
- A good authority page is one that is pointed to by many good hub pages



Bipartite Core

HITS Algorithm [K1998]



Let ***a*** is the vector of authority scores and ***h*** be the vector of hub scores

a=[1,1,….1], ***h*** = [1,1,…..1] ;

do

$a = A^T h$;

$h = Aa$;

Normalize ***a*** and ***h***;

while *a* and *h* do not converge(reach a convergence threshold)

$\mathbf{a}^* = a$;

$\mathbf{h}^* = h$;

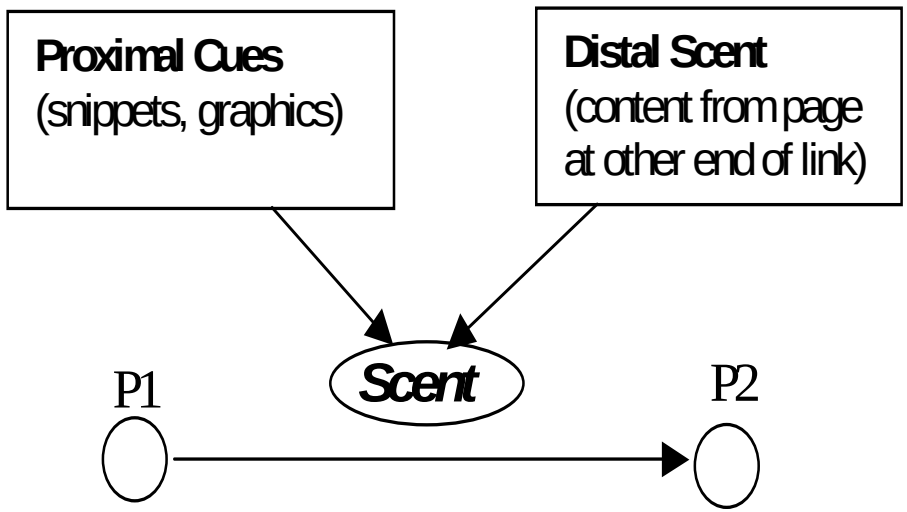
return $\mathbf{a}^*$, $\mathbf{h}^*$

The vectors **a*** and **h***represent the authority and hub weights

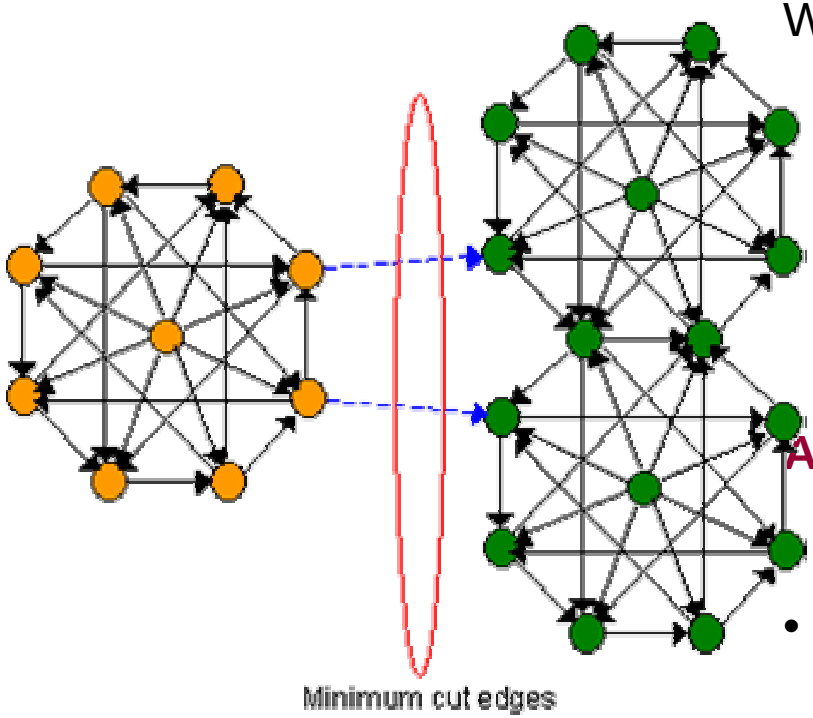
Information Scent [CPCP2001]

Key idea

- a user at a given page “foraging” for information would follow a link which “smells” of that information
- the probability of following a link depends on how strong the “scent” is on that link



Web Communities [FLG2000]



Web Communities

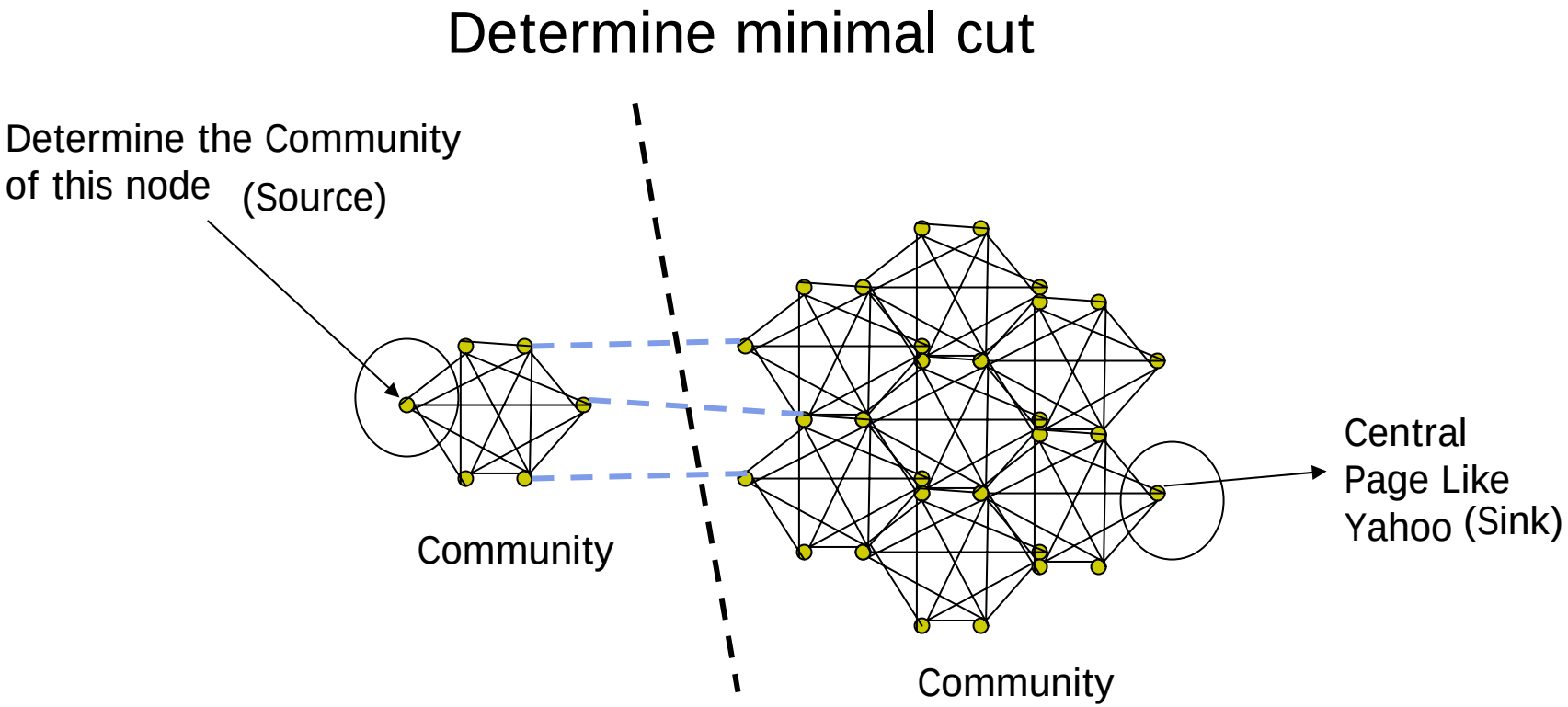
Definition

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.

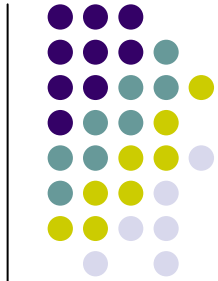
Approach

- Maximal-flow model
- Graph substructure identification

Max Flow- Min Cut Algorithm



Conclusions



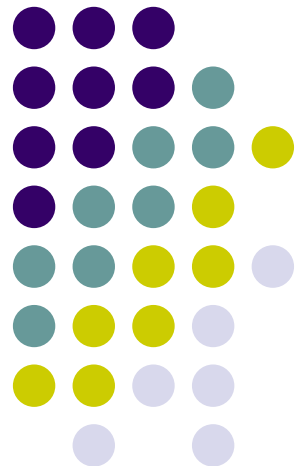
Web Structure is a useful source for extracting information such as

- Quality of Web Page
  - *The authority of a page on a topic*
  - *Ranking of web pages*
- Interesting Web Structures
  - *Graph patterns like Co-citation, Social choice, Complete bipartite graphs, etc.*
- Web Page Classification
  - *Classifying web pages according to various topics*

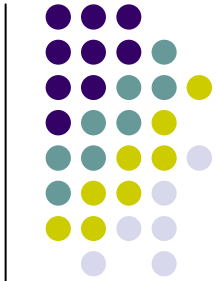
Conclusions (Cont…)

- Which pages to crawl
  - *Deciding which web pages to add to the collection of web pages*
- Finding Related Pages
  - *Given one relevant page, find all related pages*
- Detection of duplicated pages
  - *Detection of neared-mirror sites to eliminate duplication*

Web Usage Mining

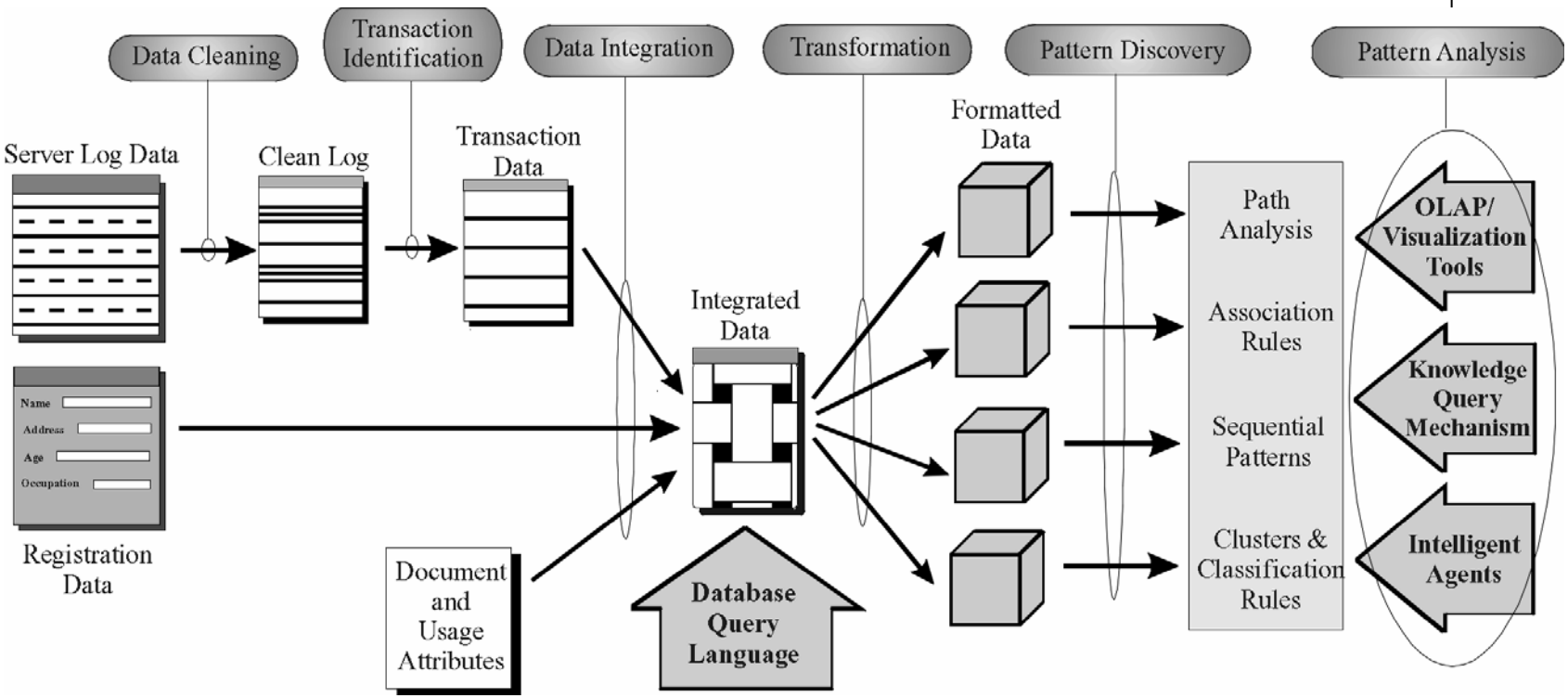


What is Web Usage Mining?

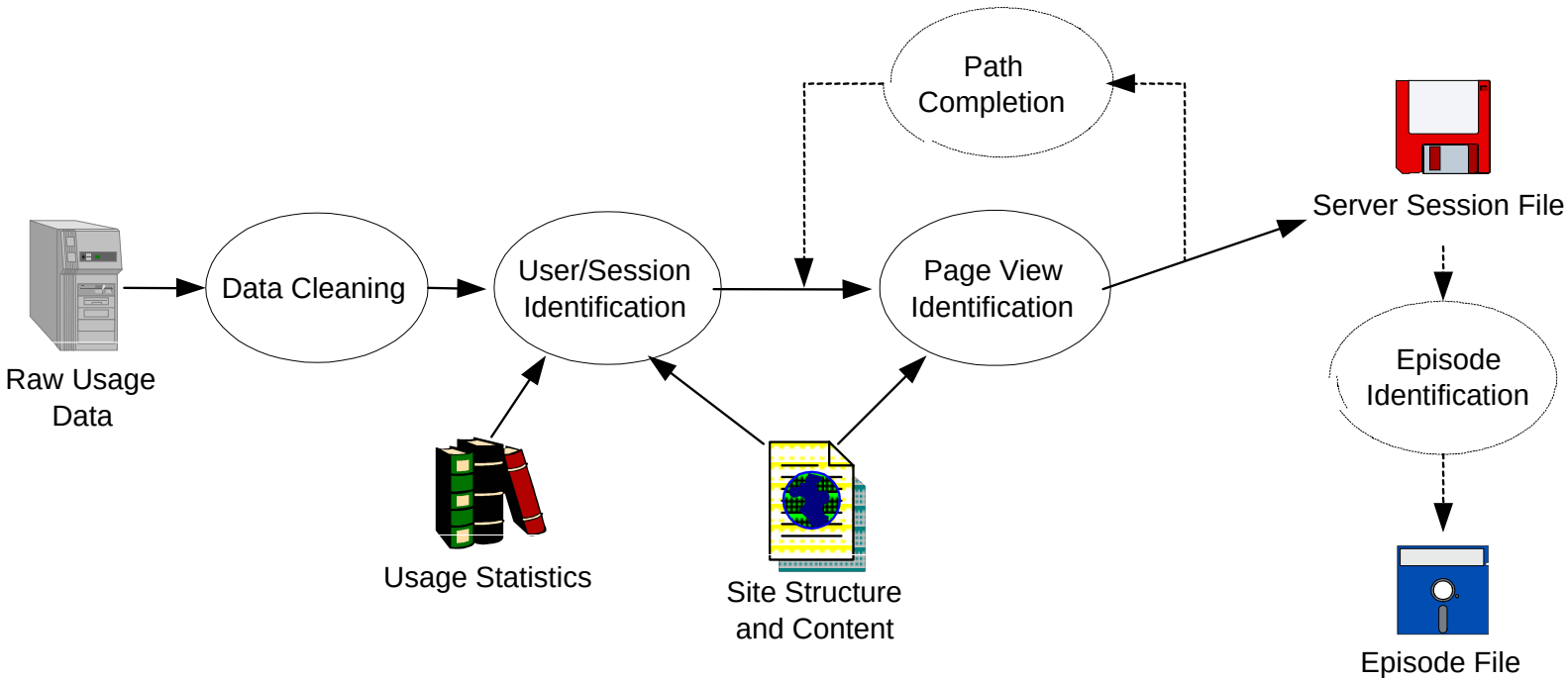


- A *Web* is a collection of inter-related files on one or more *Web servers*
- *Web Usage Mining*
  - ➔ Discovery of meaningful patterns from data generated by client-server transactions on one or more Web localities
- Typical Sources of Data
  - ➔ automatically generated data stored in server *access* logs, *referrer* logs, *agent* logs, and client-side *cookies*
  - ➔ user profiles
  - ➔ meta data: page attributes, content attributes, usage data

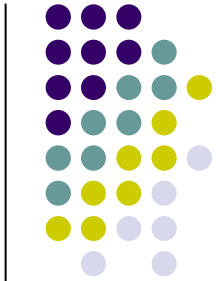
The Web Usage Mining Process



Preprocessing Architecture



ECLF Log File Format



IP Address	rfc931	authuser	Date and time of request	request	status	bytes	referer	user agent
128.101.35.92	-	-	[09/Mar/2002:00:03:18 -0600]	"GET /~harum/ HTTP/1.0"	200	3014	http://www.cs.umn.edu/	Mozilla/4.7 [en] (X11; I; SunOS 5.8 sun4u)

IP address: IP address of the remote host

Rfc931: the remote login name of the user

Authuser: the username as which the user has authenticated himself

Date: date and time of the request

Request:the request line exactly as it came from the client

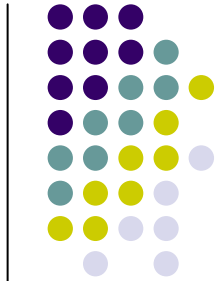
Status: the HTTP response code returned to the client

Bytes: The number of bytes transferred

Referer: The url the client was on before requesting your url

User_agent: The software the client claims to be using

Issues in Usage Data

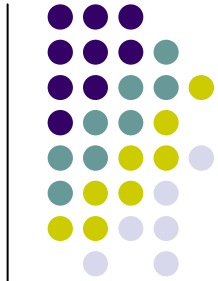


- Session Identification
- CGI Data
- Caching
- Dynamic Pages
- Robot Detection and Filtering
- Transaction Identification
  - ✓ Identify Unique Users
  - ✓ Identify Unique User transaction

Session Identification Problems

- "AOL Effect": Single IP Address/ Multiple Users
  - ISP Proxy Servers
  - Public Access Machines
- "WebTV Effect": Multiple IP Addresses/ Single Session
  - Rotating IP for load balancing
  - Privacy tools

Session Identification Solutions

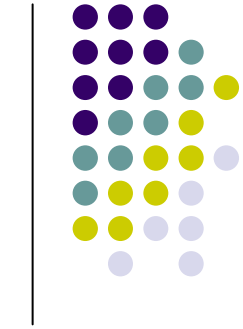


- Cookies - small piece of code that is saved on the client machine
- User Login – Require user to use login ID with password
- Embedded SessionID.
- IP+Agent.
- Client-side tracking

CGI Data

- Common Gateway Interface (CGI): Method used to pass variables and user entered data to Content Server
- Set of name/value pairs that are attached to end of a URI

Example URI



Base URI

/cgi-bin/templates

?BV_EngineID=falfiffkdgfbemmcfnnckcgl.0&BV_Operation=Dyn_RawSmartLink&BV_SessionID=2131083763.936854172&BV_ServiceName=MyStore&form%25destination=mysite/logo.tmpl

CGI Data

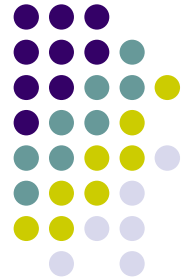
CGI Data Problems

- Hidden Values: POST requests have a “hidden” option that removes the name/value pairs from the URI
- Content Servers can maintain “state” in the form of session variables. The relevant data for determining what page was accessed may not be in the current CGI pairs

CGI Data Solutions

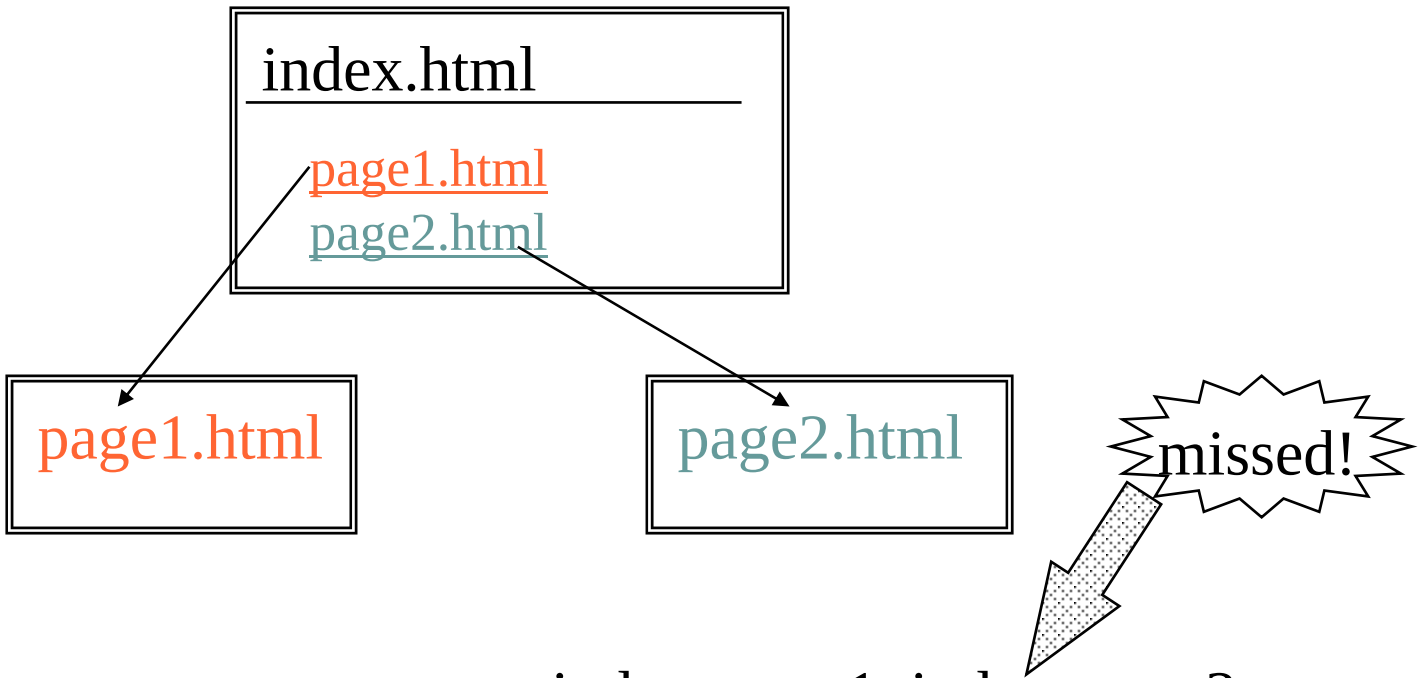
- Pull data directly from the HTTP traffic instead of the Server log
  - Advantages: Generic, works for any Web server/Content server configuration
  - Disadvantages: No access to secure data. No access to internal Content server variables
- Have Content server create an “access log”
  - Advantages: All relevant information is always available Clean log of page views instead of file accesses is created. No sessionID “first access” problems
  - Disadvantages: Content server performance may be degraded. Not automatic like Server logs

Caching Problems



- Clients and Proxy Servers save local copies of pages that have been accessed
- Uses of the “back” and “forward” buttons on a browser may access local copy instead of requesting a new one from the server

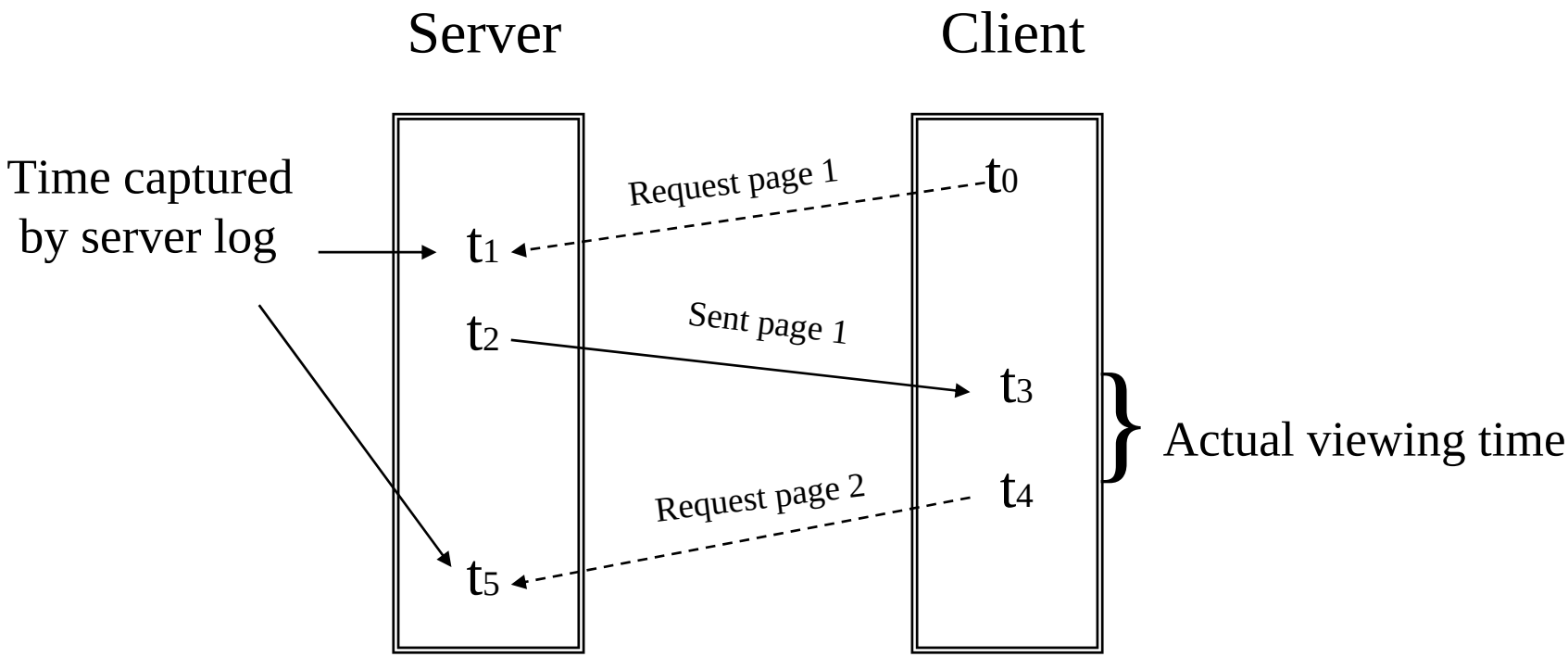
Server Log Incompleteness due to Caching



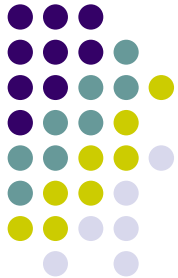
Access pattern: index, page1, index, page2

Record in server log:: index, page1, page2

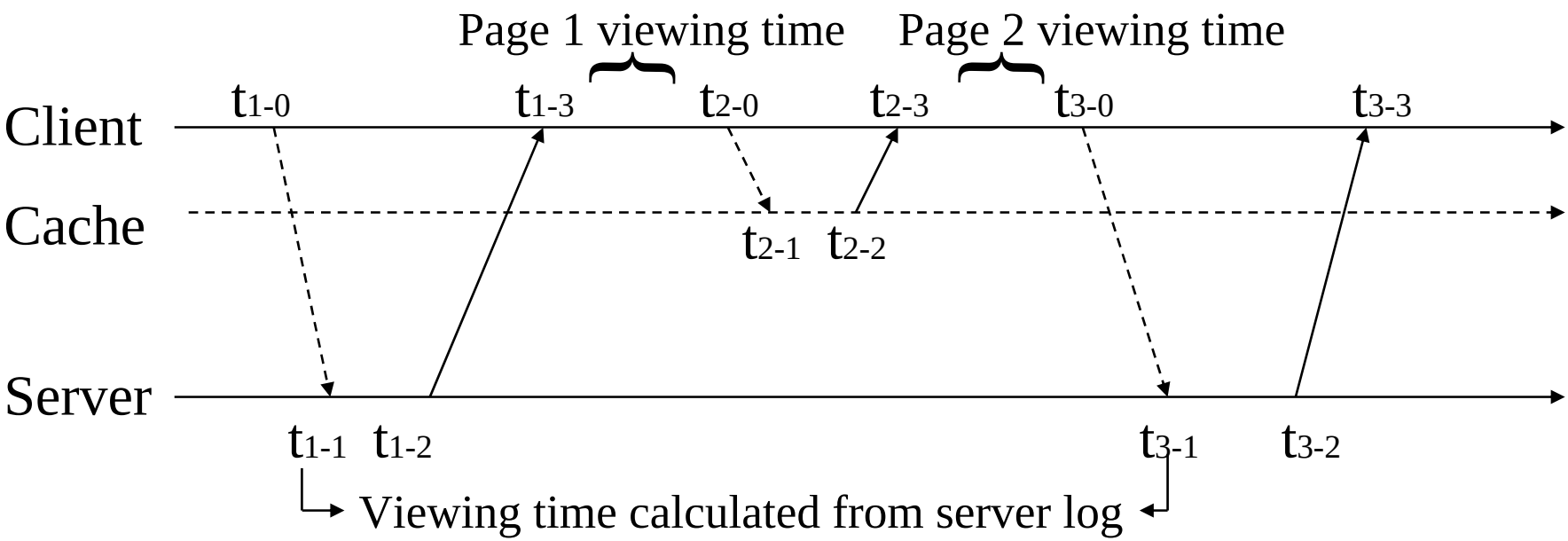
Wrong Access Timings Recorded at Server



Missed Page Views at Server



- Viewing time for cached pages



Caching Solutions

- Dynamic content greatly reduces the number of cached page accesses
  - Advantages: Fewer “missed” page views
  - Disadvantages: Increased Server traffic
- “Negative” expiration dates for pages force browsers to request a new version

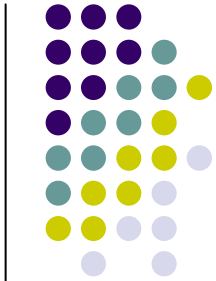
Robot Detection and Filtering [TK2002]

Web robots are software programs that automatically traverse the hyperlink structure of world wide web in order to locate and retrieve information

Motivation for distinguishing web robot visits from other users

- Unauthorized gathering of business information at e-commerce web sites
- Consumption of considerable network bandwidth
- Difficulty in performing click-stream analysis effectively on web data

Transaction Identification



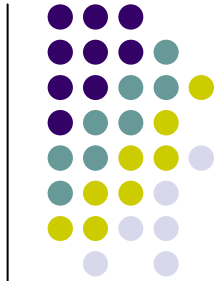
- Main Questions:
  - how to identify unique users
  - how to identify/define a user transaction
- Problems:
  - user ids are often suppressed due to security concerns
  - individual IP addresses are sometimes hidden behind proxy servers
  - client-side & proxy caching makes server log data less reliable
- Standard Solutions/Practices:
  - user registration
  - client-side cookies

  } not full-proof

  - *cache busting*

  } increases network traffic

Heuristics for Transaction Identification



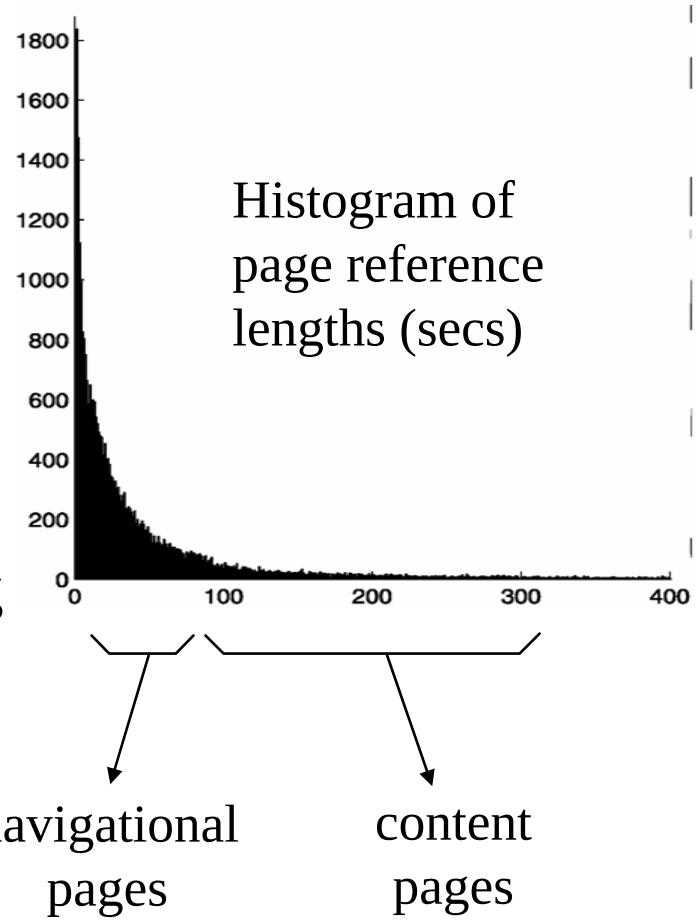
- Identifying User Sessions
  - use IP, agent, and OS fields as key attributes
  - use client-side cookies & unique user ids, if available
  - use session time-outs
  - use synchronized referrer log entries and time stamps to expand user paths belonging to a session
  - path completion to infer cached references

    EX: expanding a session A ==> B ==> C by an *access pair*

    (B ==> D) results in: A ==> B ==> C ==> B ==> D
  - to disambiguate paths, sessions are expanded based on page attributes (size, type), reference length, and no. of back references required to complete the path

Inferring User Transactions from Sessions

- Studies show that reference lengths follow an exponential distribution.
- Page types: *navigational*, *content*, *mixed*.
- Page types correlate with reference lengths.
- Can automatically classify pages as navigational or content using % of navigational pages (based on site topology) and a normal estimate of Chi-squared distribution.
- A transaction is an intra-session path ending in a content page.



Associations in Web Transactions

- Association Rules:
  - ➔ discovers affinities among sets of items across transactions

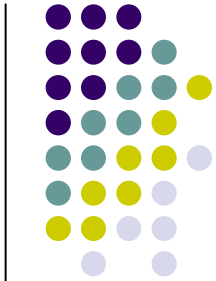
$$X \overset{\alpha,\ \sigma}{=\!=\!=\!=\!=\!=\!>} Y$$

where *X, Y* are sets of items, $\alpha =$ *confidence,* $\sigma =$ *support*

- Examples:
  - ➔ 60% of clients who accessed **`/products/`**, also accessed **`/products/software/webminer.htm.`**
  - ➔ 30% of clients who accessed **`/special-offer.html`**, placed an online order in **`/products/software/`**.
  - ➔ (Actual Example from IBM official Olympics Site)

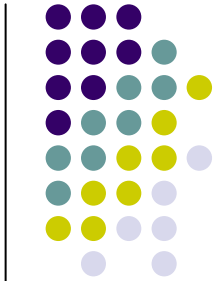
    {Badminton, Diving} ===> {Table Tennis} ($\alpha = 69.7\%,\ \sigma = 0.35\%$)

Other Patterns from Web Transactions



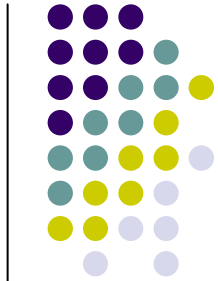
- Sequential Patterns:
  - 30% of clients who visited **`/products/software/`**, had done a search in **Yahoo** using the keyword “software” before their visit
  - 60% of clients who placed an online order for WEBMINER, placed another online order for software within 15 days
- Clustering and Classification
  - clients who often access **`/products/software/webminer.html`** tend to be from educational institutions.
  - clients who placed an online order for software tend to be students in the 20-25 age group and live in the United States.
  - 75% of clients who download software from /products/software/demos/ visit between 7:00 and 11:00 pm on weekends.

Path and Usage Pattern Discovery



- Types of Path/Usage Information
  - Most Frequent paths traversed by users
  - Entry and Exit Points
  - Distribution of user session durations / User Attrition
- Examples:
  - 60% of clients who accessed **`/home/products/file1.html`**, followed the path **`/home`** ==> **`/home/whatsnew`** ==> **`/home/products`** ==> **`/home/products/file1.html`**
  - (Olympics Web site) 30% of clients who accessed sport specific pages started from the *Sneakpeek* page.
  - 65% of clients left the site after 4 or less references.

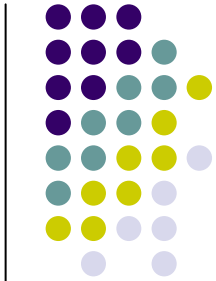
Pattern Analysis



- Pattern Analysis Tools/Techniques
  - Knowledge Query Mechanism
  - OLAP / Visualization Tools
  - Intelligent Agents / Expert Systems
- WEBMINER: SQL-like Knowledge Query Mechanism

```
SELECT association-rules (A*B*C*)
FROM "rules.out"
WHERE time >= 970101 AND domain = "edu" AND
      support >= 0.01 AND confidence >= .85
```

Implications of Web Usage Mining for E-commerce



- Electronic Commerce
  - ➔ determine lifetime value of clients
  - ➔ design cross marketing strategies across products
  - ➔ evaluate promotional campaigns
  - ➔ target electronic ads and coupons at user groups based on their access patterns
  - ➔ predict user behavior based on previously learned rules and users' profile
  - ➔ present dynamic information to users based on their interests and profiles

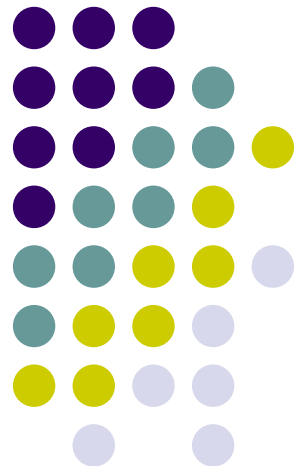
Implications for Other Applications

- Effective and Efficient Web Presence
  - determine the best way to structure the Web site
  - identify “weak links” for elimination or enhancement
  - A “site-specific” web design agent
  - Pre-fetch files that are most likely to be accessed
- Intra-Organizational Applications
  - enhance workgroup management & communication
  - evaluate Intranet effectiveness and identify structural needs & requirements

What's Round-the-Corner for WUM

- Data Integration and meta-level schemas
- Enhanced knowledge query mechanism, user interface, and visualization modules (OLAP)
- Intelligent agent to extract the most "interesting" rules from among the discovered rules
- Better models of user behavior (e.g. Information Foraging)
- Rule-based expert system to provide "suggestions" based on discovered rules

Related Concepts



Interestingness Measure [PT1998,C2000]

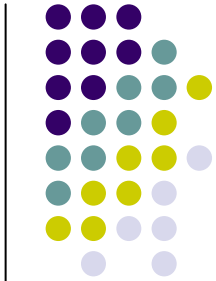
A measurement of patterns that are subjectively *different* from what is expected and above a certain support threshold

In the World Wide Web, there are two sources of information

- *Web Structure:* Reflects the author's viewpoint of browsing behavior
- *Web Usage:* Reflects the user's browsing behavior.

Any conflicting evidence from these sources of information would be termed "interesting"

User Behavior Profiles [MSSZ2002]



Why?

To understand the complex human decision making process.

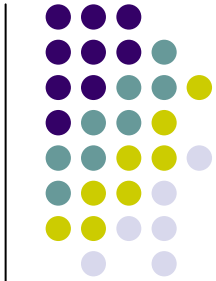
How?

- Record click-stream data.
- Gather other user information such as demographic, psychographic, etc data.

At what level?

- Within a web site e.g Amazon.Com [AMZNa].
- On the whole world wide web e.g Alexa research [ALEX] and DoubleClick [DCLKa].

Distributed Web Mining



Motivation: Data on the Web is huge and distributed across various sites

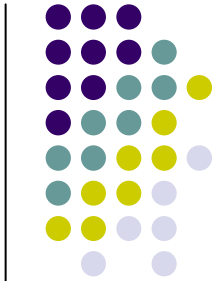
Traditional Approach: Integrate all data into one site and perform required analysis.

Problem: Time consuming and not scalable.

Solution: Analyze data locally at different locations and build an overall model

Application: Personalization of Web Sites depending on user's 'life on the web' (the users interests, locations and behavior across different sites).

Distributed Web Mining - Approaches



The approaches can be classified into two kinds

- **Surreptious**
  - ✓ User behavior across different web sites is tracked and integrated without the user having to explicitly submit any information.
- **Co-operative**
  - ✓ Behavior is reported to a central organization or database (e.g Network Attacks are reported to CERT)

Web Visualization

Motivation

Mining Web Data provides huge information that can be better understood using visualization tools than pure text representation.

Prominent tools developed

- WebViz
- WUM: Web Utlization Miner
- WEEV
- WebQuilt
- Naviz

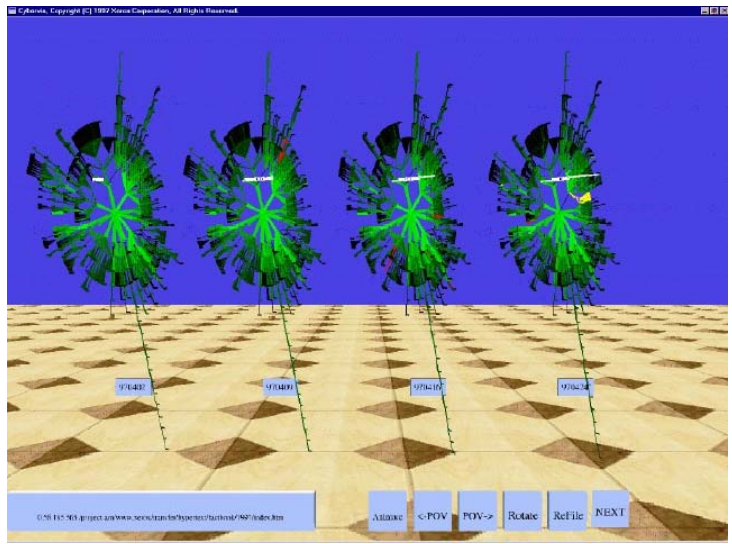
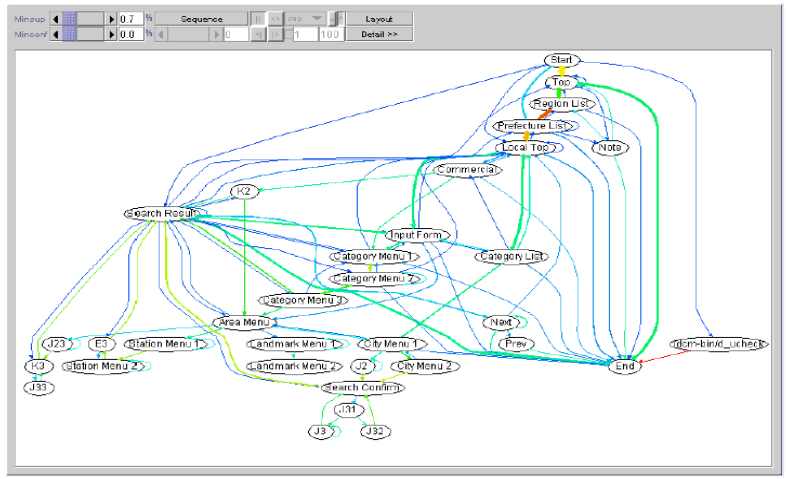


Figure: **WEEV**- *Time Tube* representing the evolution of Web Ecology over time

Naviz - User Behavior Visualization of Dynamic Page [PPT+2003]

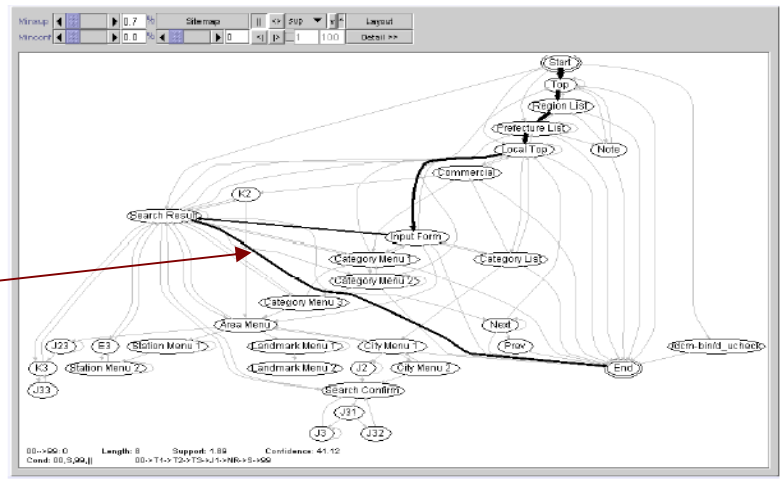
Naviz Features

- Two operation mode
  - ✓ Traversal diagram mode
  - ✓ Traversal path mode
- Thickness of edge
  - ✓ Represents support value
- Color of edge (range from blue to red)
  - ✓ Represents confidence degree (low to high)



NAVIZ: Traversal Diagram mode

Visitor Success path



NAVIZ: Traversal path mode

Topic Distillation

Definition: Identification of a set of documents or parts of documents that are most relevant to a query on a topic

Approaches:

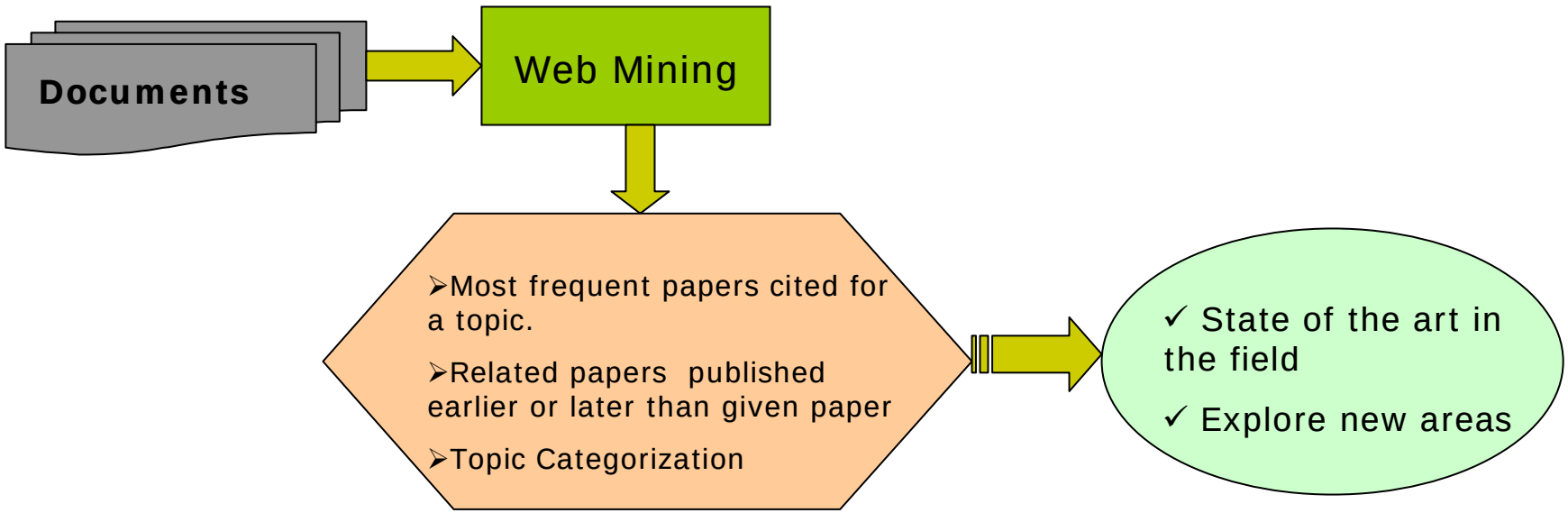
- ✓ Kleinberg's Hubs and Authorities
- ✓ The FOCUS project: Selectively seek out pages that are relevant to a pre-defined set of *topics*
- ✓ Integration of Document Object Model of a Web page and the hyperlink structure to build an extension of Hubs and Authorities model
- ✓ Web Page Reputations
- ✓ Topic Sensitive Pagerank

Online Bibiliometrics

Motivation:

- Articles that are online are more often cited than articles offline
- Interaction and exchange of information easier

Examples: SCI, ACM portal, CiteSeer, DBLP etc.



Web Page Categorization

Web page Categorization determines the category or class a web page belongs to, from a pre-determined set of categories or classes.

(*categories can be based on topics or other functionalities such as home page, research page, content pages etc.*)

Approaches:

- Pirolli et al. defined 8 categories and identified 7 features based on which they web pages can be classfied.
- Chakrabarti et al. used relaxation labeling technique and assigned categories based on neighboring documents that link to a given document or linked by a given document.
- Getoor et al used a Probabilistic Relational Model to specify probability distribution over document link databse and classify documents using belief popagation methods

Semantic Web Mining

Motivation:

- Automatic retrieval of documents from the unstructured form of data on the web is difficult.
- Search Engines are not precise with respect t the semantics of the documents retrieved by them.

Primary Idea of Semantic Web:

- To generate documents that have attached semantics.
- To develop techniques to mine information from such structured data with semantics.

Semantic Web Mining

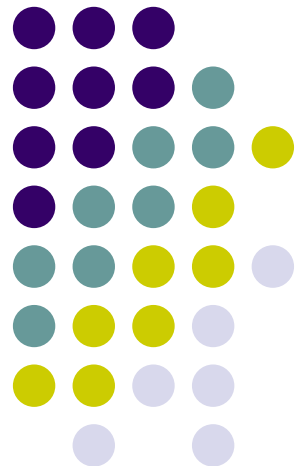
Semantic Web Formats:

- RDF:Nodes and attached attribute/value pairs that can be modeled as directed labeled graph.
- XML Topics Network of topics that can be formed using semantics oif the underlying data. It can be viewed as online versions of rinted indices and catalogs.

Tasks:

- Apply Web Mining techniques to understand ontologies from vast source if unstructured documents in the web.
- Define ontologies for existing and future documents to make search more precise.

Web Services & Web Mining



The slides in this section of the talk borrow heavily from the Web Services presentation at http://www.w3.org/2003/Talks/0818-msm-ws/

Definitions [SM2003, M2002]

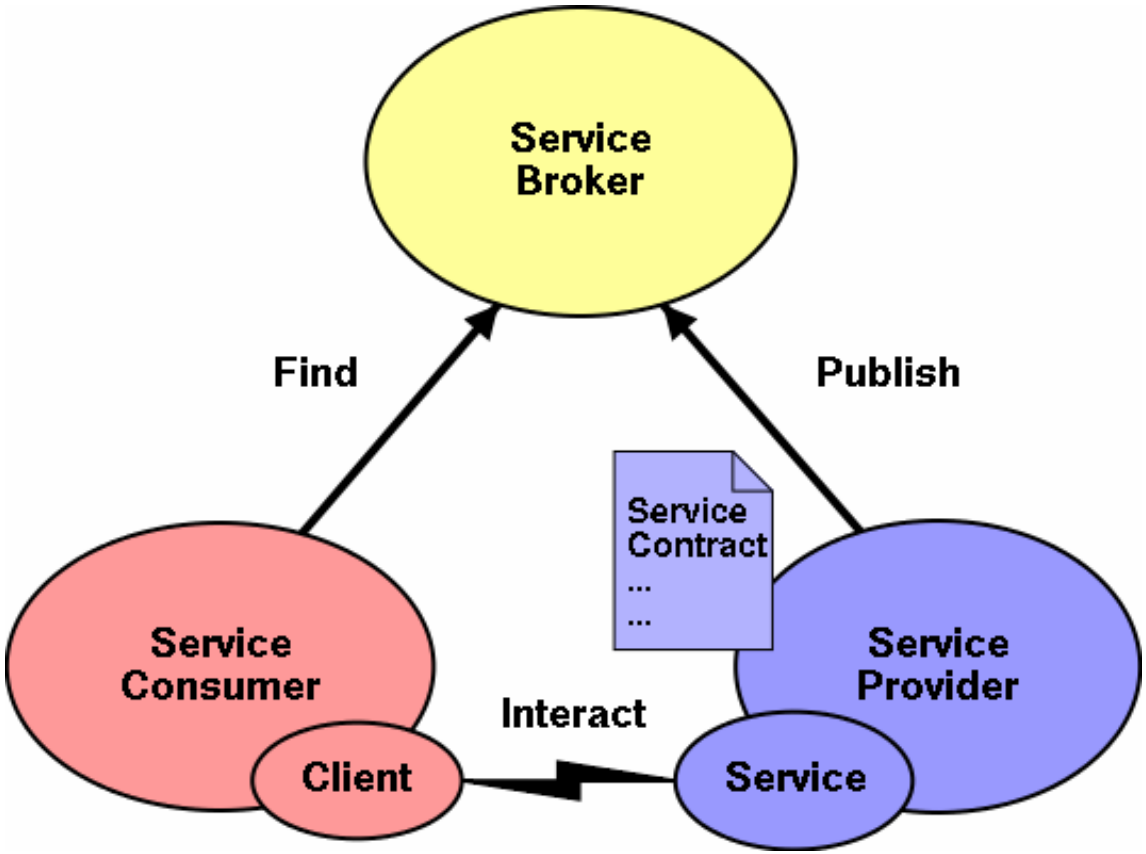
Web Services have been described in various ways

- ***Web Services*** are a means of allowing applications to talk to one another using XML (Extensible Markup Language) messages sent via the standard Web protocol of HTTP
- ***Web services*** are a new breed of web application that are self-contained, self-describing, modular applications and can be published, located, and invoked across the web
- ***Web services*** perform functions, which can be anything from simple requests to complicated business processes

Web Services – what they provide [SM2003]

- Once a web service is deployed, other applications (and other web services)can discover and invoke the deployed service.
- Web Services is also viewed as an important interoperability mechanism for the J2EE and Microsoft's .NET worlds to come together.
- Services that follows from this:
  - messaging (e.g. SOAP, XML)
  - description (e.g. WSDL, XML Schema)
  - discovery (e.g. UDDI)
  - security (e.g. TLS, SSL)

Service-Oriented Architecture [SM2003]



SOAP:Simple Object Access Protocol [SM2003]

SOAP is used as message structure, generic message-processing model, extensibility model

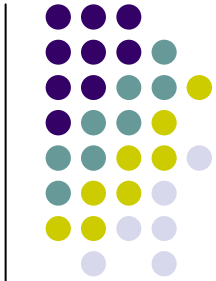
- ✓ message construction (envelope, header, body)
- ✓ message exchange patterns (MEP) and how to define more
- ✓ processing model for messaging: originator, intermediaries, destination
- ✓ extensibility mechanism and *mustUnderstand* attribute
- ✓ fault system
- ✓ bindings to transport protocols (HTTP, SMTP, ...)

WSDL: Web Services Description Language [SM2003]

A WSDL document describes a service

- ✓ message(s) accepted and emitted: abstract description (XML Schema)
- ✓ network protocol(s) and message format(s)
- ✓ operation: exchange of messages
- ✓ port type: collection of operations
- ✓ port: implementation of a port type
- ✓ service: collection of ports

UDDI: Universal Description, Discovery, and Integration [SM2003]



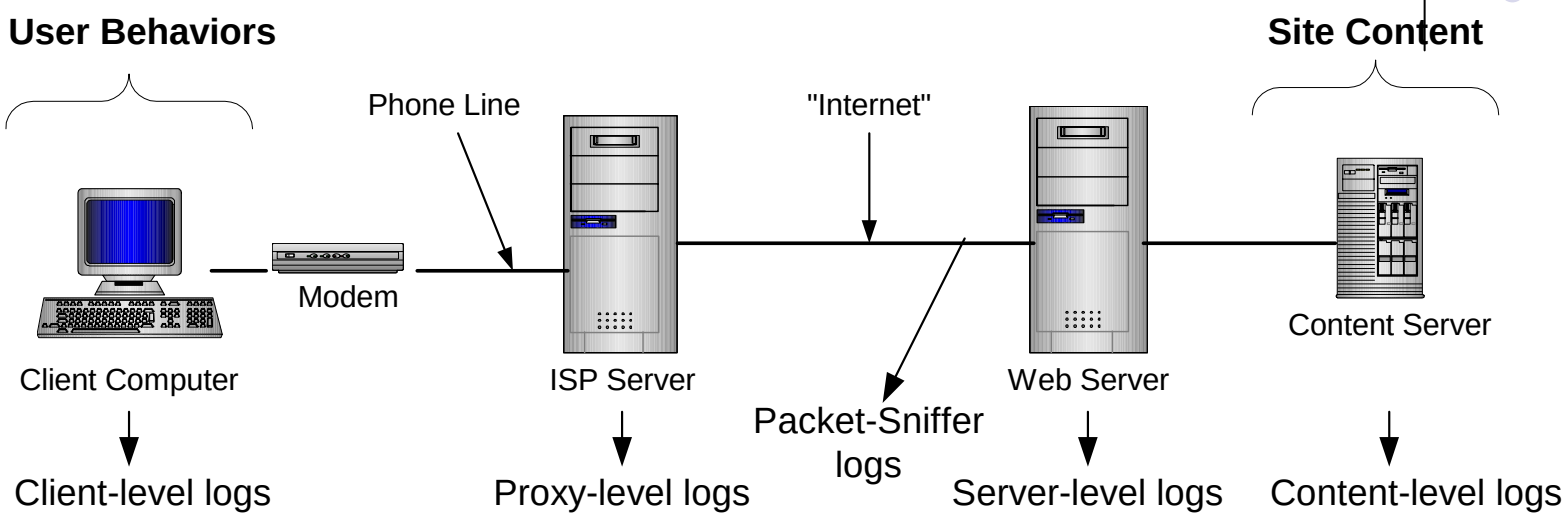
Universal description, discovery, and integration.

- ✓ registry system
- ✓ business entities, business services, specifications, service types
- ✓ standard taxonomies to describe businesses, services, and service types (?!)
- ✓ “The UDDI Business Registry is intended to serve as a global, all-inclusive listing of businesses and their services. The UDDI Business Registry does not contain detailed specifications about business services. It points to other sources that contain the service specifications.”
- ✓ private registries also possible

How Web Mining can help WS

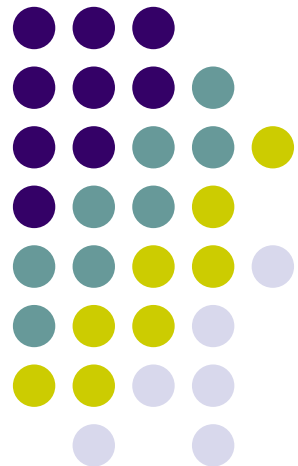
- Web Data collected at the client and server level can help in better performance and providing better features for Web Services
  - Understanding of client-server interactions
    - The data from the interactions can be mined for analyzing interesting patterns
  - Personalization of Web Services
    - The client level data can provide information to personalize Web services for the users
  - Fraud analysis

Web services optimization



- Various types of logs mined for
  - Improved caching and pre-fetching
  - Request routing
  - Congestion analysis

Web Mining Applications



Personalized experience in B2C e-commerce – Amazon.com

Use of Web mining

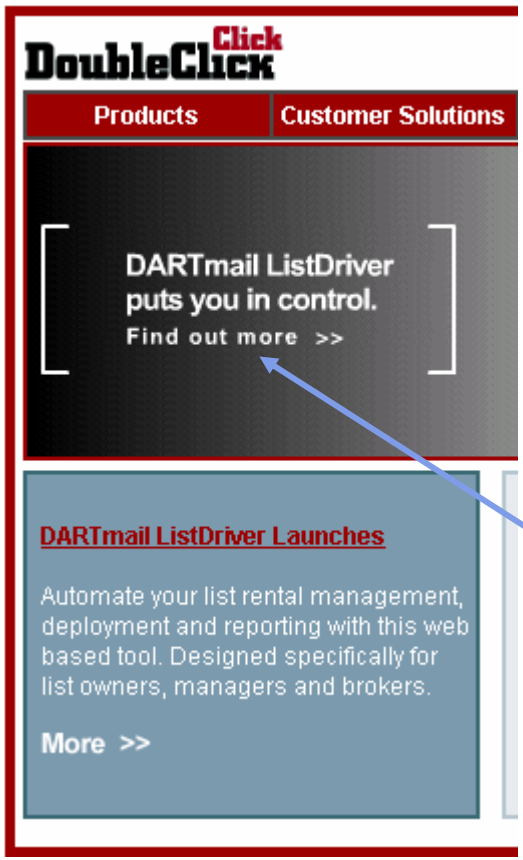
- cookies to identify user
- analysis of user's past behavior and 'peer group analysis' for
  - personalized messages
  - category recommendations
  - 'gold box' offers
- Use of clustering, association analysis, temporal sequence analysis, etc.

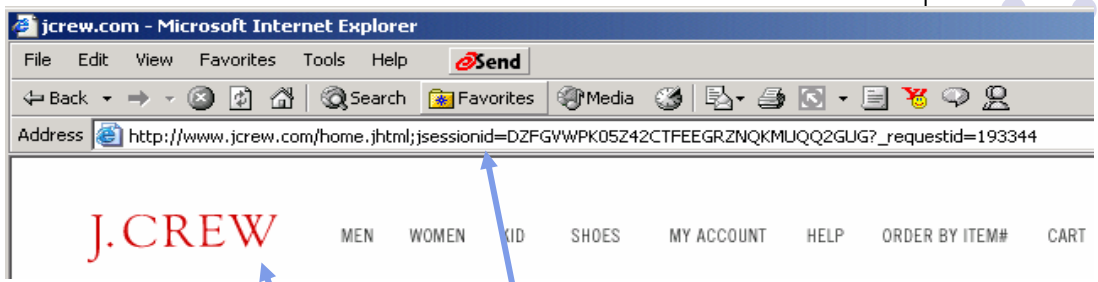
Web search - Google

Use of Web mining

- content analysis to determine relevant pages
- hyperlink analysis to rank the relevant pages based on their quality

Web-wide user tracking - DoubleClick





DoubleClick
- places its own cookie on the machine of its customer's users
- reads this cookie each time it serves an ad to this user through any customer in the DoubleClick network

Use of Web mining
- use of a special cookie to track user across multiple Web sites
- analysis of multi-site behavior
- ad serving using DART system

Understanding user communities - AOL

AOL groups can be
- sponsored (for a fee) by organizations interested in the behavior of group participants
- can have the orgn's representatives as participants

Web mining on group activity – usage & content
- interests and opinions of group members
- treat as a focus group
  - for new product/svc
  - for opinion on issue

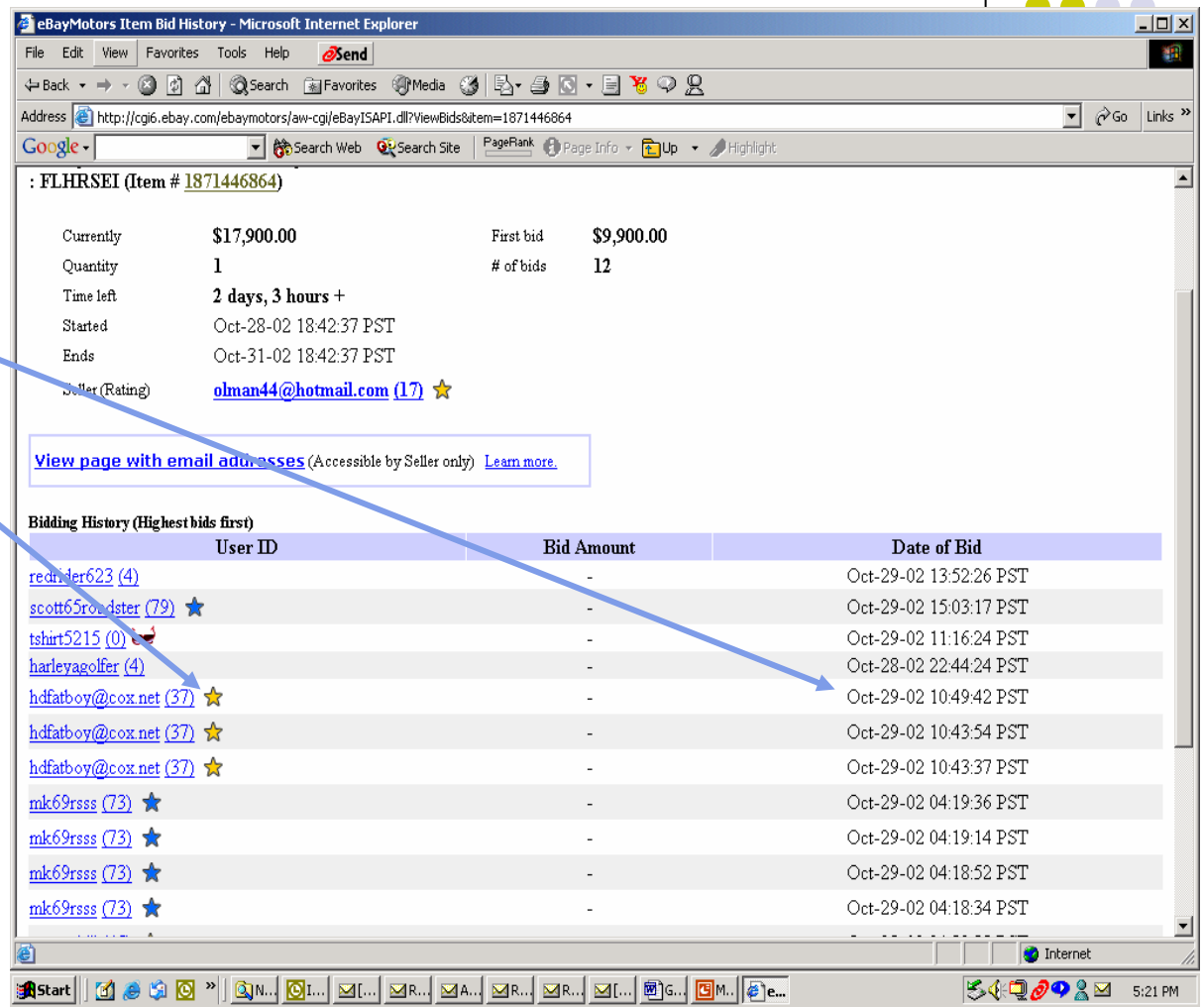
Understanding auction behavior - eBay

eBay has detailed data on

- bid history
- participant rating
- bid data
- usage data

Use of Web mining to

- categorize participants into various types
- classify auctions into various types
- determine fraudulent bids
- determine 'auction fixing'



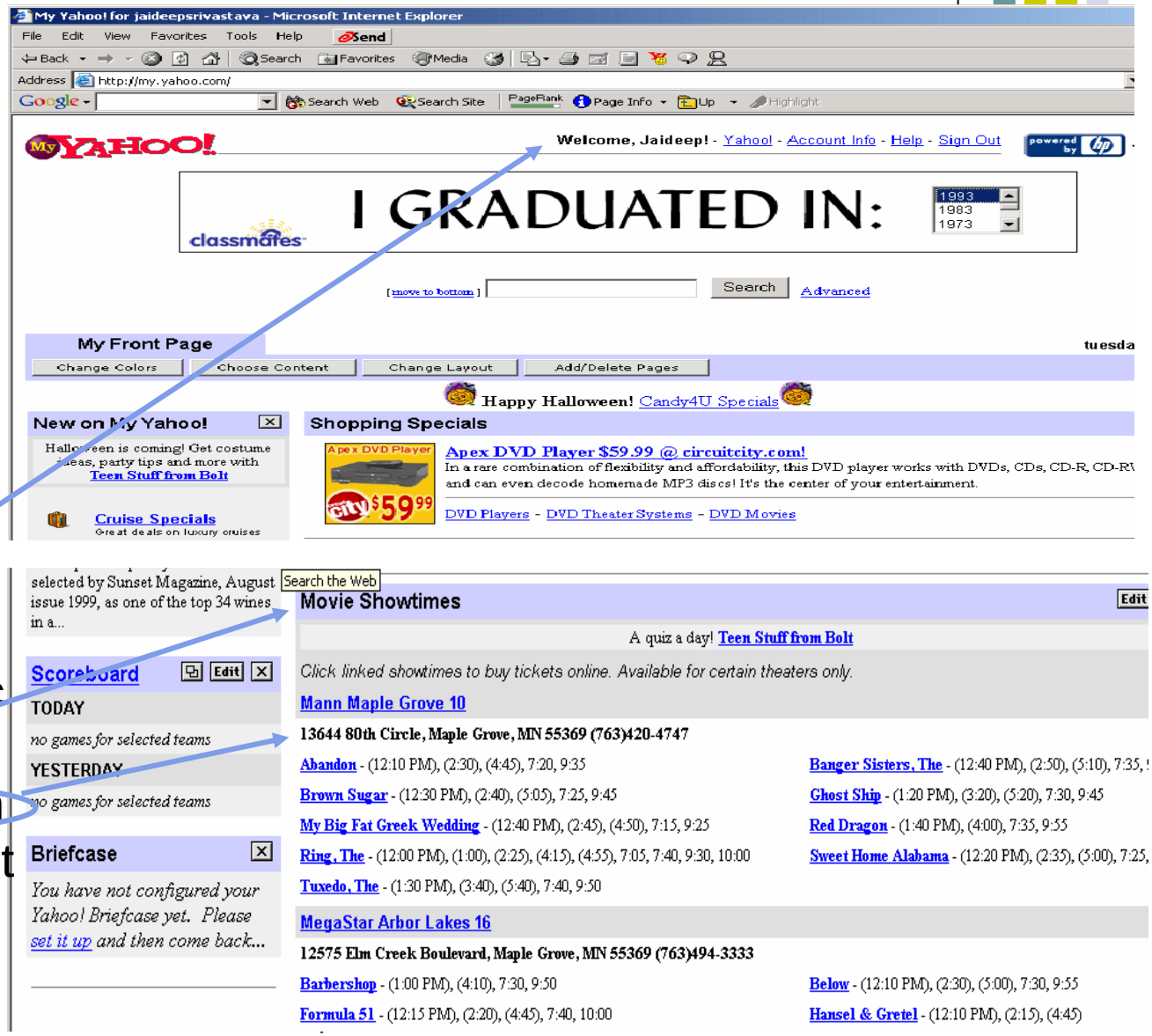
Personalized web portal - MyYahoo

MyYahoo has detailed Data on individual's

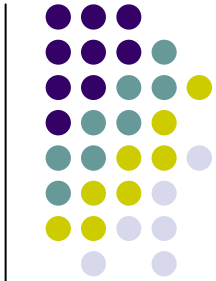
- demographic
- preferences
- media preferences
- usage patterns

Use of Web mining to

- create personalized messages
- recommend prod/svc based on preference & location
- deliver media content based on preference & usage (not shown)



CiteSeer – Online Bibliometrics



Search Topic

First paper returned according to the weighted citations

Papers that directly cite the given paper

Similar or Related Papers

CiteSeer Find: Web Mining Documents Citations

Searching for PHRASE **web mining**.

Restrict to: Header Title Order by: Citations Hubs Usage Date Try: Amazon B&N Go

233 documents found. **Order: citations weighted by year.**

Web Usage Mining: Discovery and Applications of.. - Srivastava.. (2000) (Correct) (28 citations)

Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data (2000)

Corrections) (31 citations)

Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan

SIGKDD Explorations

Cited by: More

i-Miner: A Web Usage Mining Framework Using.. - Ajith Abraham And (Correct)

Mining Web Logs for Personalized Site Maps - Toolan, Kushmerick (2002) (Correct)

Predicting Next Page Access By Time Length Reference In The.. - Yalçinkaya (2002) (Correct)

Similar documents (at the sentence level):

31.5%: Web Usage Mining: Discovery and Application of Interestin.. - Cooley (2000) (Correct)

Similar documents based on text: More All

0.7: Some Experiences on Large Scale Web Mining - Kitsuregawa, Pramudiono, Ohura, .. (Correct)

0.7: Blockmodeling Techniques for Web Mining - Schoier (Correct)

0.5: Game Usage Mining: Information Gathering for Knowledge.. - Tveit, Tveit (2002) (Correct)

Related documents from co-citation: More All

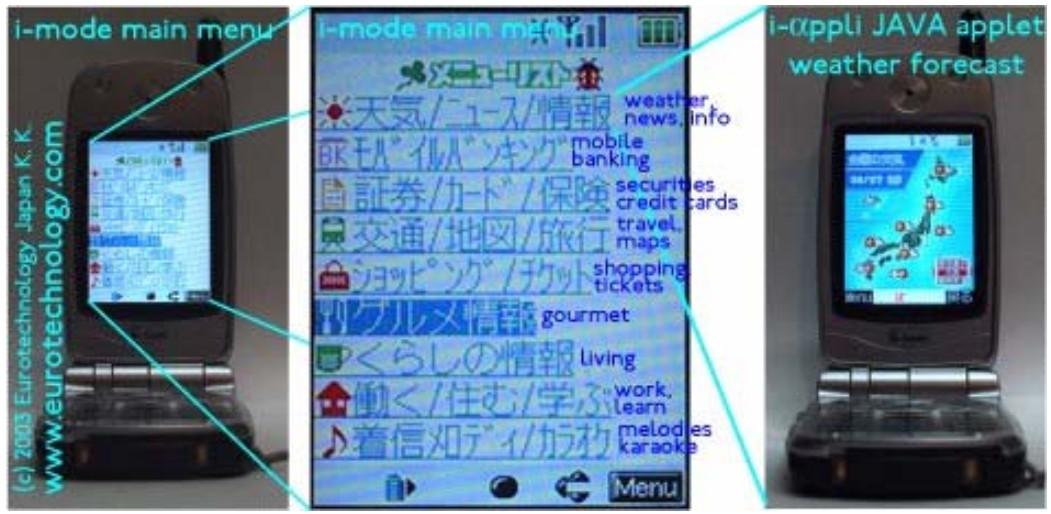
14: Data preparation for mining world wide web browsing patterns - Cooley, Mobasher et al. - 1999

10: Web Mining: Information and Pattern Discovery on the World Wide Web - COOLEY, SRIVASTAVA et al. - 1997

9: Fast Algorithms for Mining Association Rules - Agrawal, Srikant - 1994

i-Mode – NTT D0C0Mo's mobile internet access system

- 40 million users who access internet from their cell-phones.
- Users can:
  - ✓Receive, send e-mail
  - ✓Do online shopping or banking
  - ✓Receive traffic news and weather forecasts
  - ✓Search for local restaurants and other things.

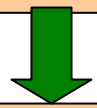


i-Mode: Internet access through mobile system

Mining information from i-MODE

i-Mode has its own semantics, structure and usage:

- It uses its own Markup Language: cHTML (compact HTML).
- Content of web pages are also restricted.(5 Kbytes max)
- Usage data is available at an individual level.

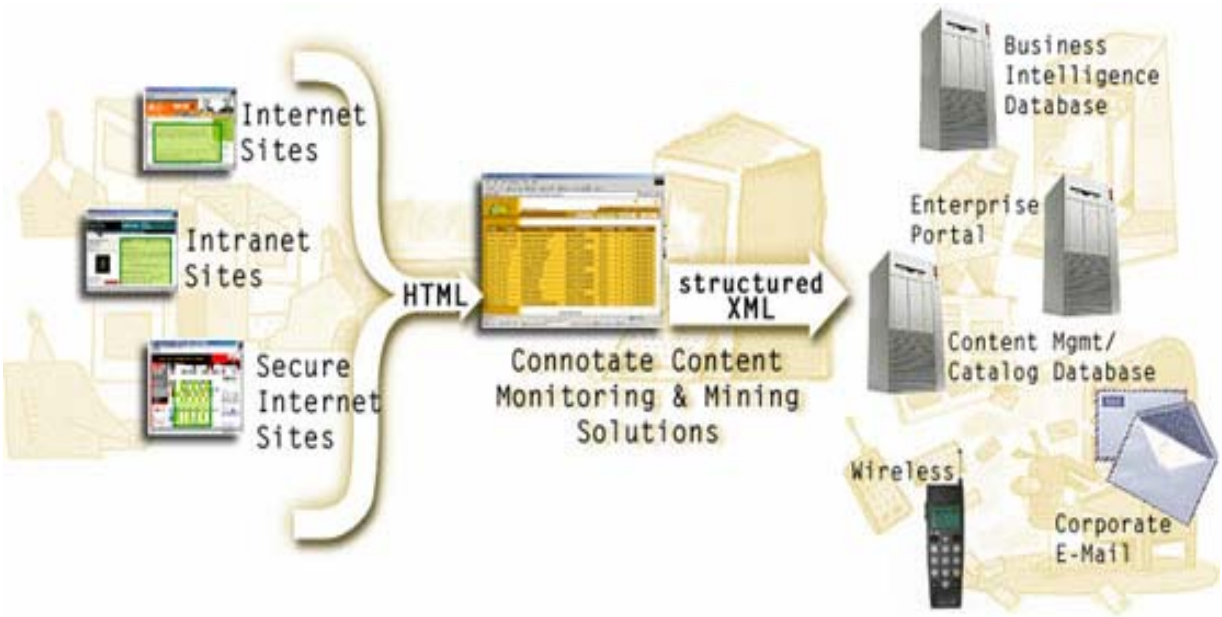


✓ Techniques for mining information for this kind of data.

✓Personalization at an individual level including geographical preferences.

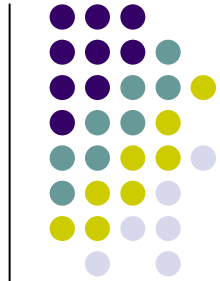
V-TAG Web Mining Server- Connotate Technologies

The Web Mining Server supports information agents that monitor, extract and summarize information from web sources.



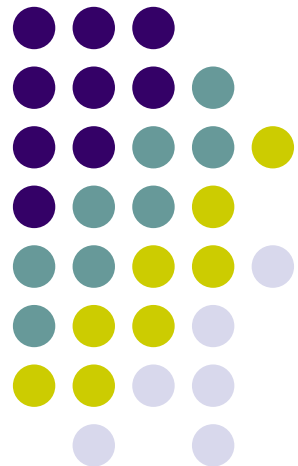
v-Tag Web Mining Server Architecture

Features of v-Tag



- Easy to set up graphical user interface
- Automation of tracking and summarizing helps businesses and enterprises to analyze the various processes easily
- Content can be converted to more structured format like XML
  - ✓ Can be used for business intelligence, supply chain integration
  - ✓ Converted content can also be sent as an e-mail or message to an user on his mobile

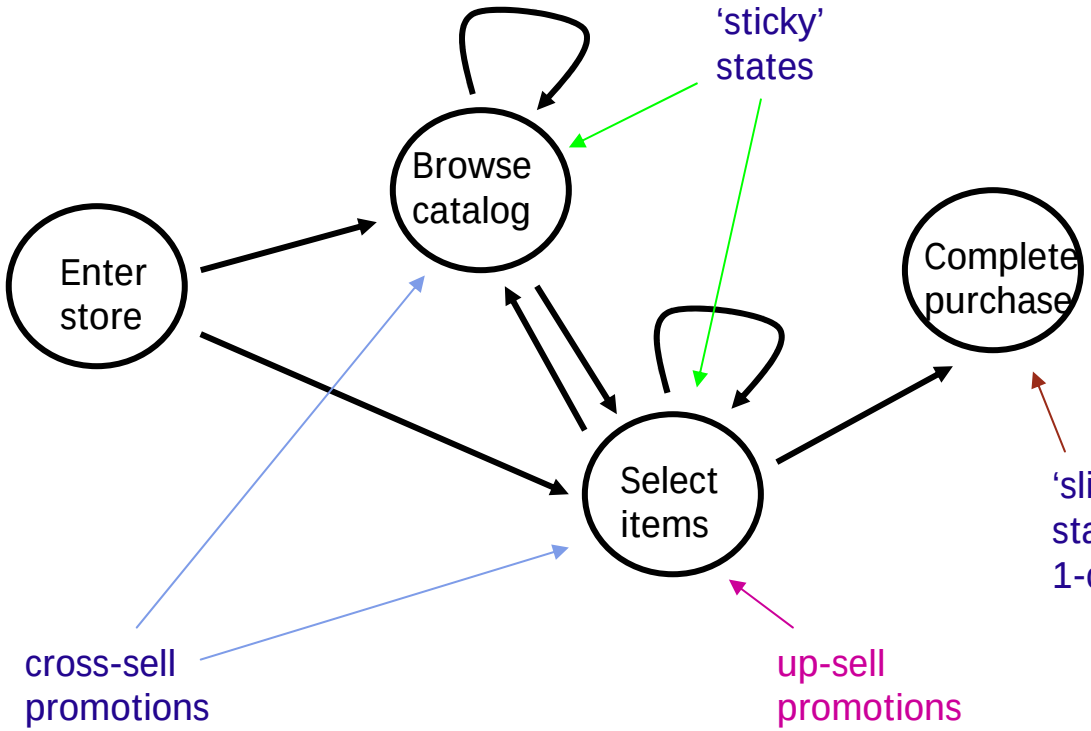
Future Directions



Web metrics and measurements

- Web as an apparatus for behavior experiments – e.g. Amazon's WebLab
  - Very large sample size – 10K to 100K
  - No 'testing bias' on part of subjects
  - No 'peer-influence bias' on subjects
- Issues
  - Design of useful metrics – what matters to the application
  - Techniques for efficient instrumentation and collection of measurements related to these metrics

Process mining example – Shopping pipeline analysis



Overall goal:
- **Maximize probability of reaching final state**
- **Maximize expected sales from each visit**

- Shopping pipeline modeled as state transition diagram
- Sensitivity analysis of state transition probabilities
- Promotion opportunities identified
- E-metrics and ROI used to measure effectiveness

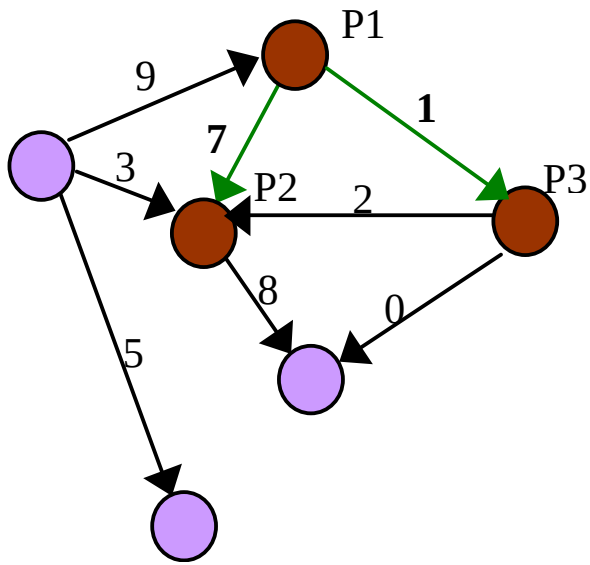
Process mining – Issues

- Analyze Web data (usage and structure) to extract process models
- Analyze ‘process outcome’ data to understand the value of various parts (e.g. states) of the process model – e.g. impact of various states on the probability of desired/undesired outcomes
- Provide (quantitative) input to help develop strategies for increasing (decreasing) the probabilities of desired (undesired) outcomes

Combining Web Usage With Web Structure

Number of traversals (*Web Usage*) on each link (*Web Structure*) is used to estimate the transition probabilities that can be used for

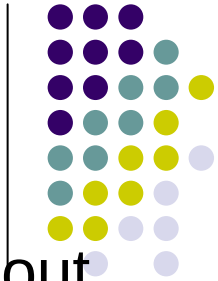
- Link Prediction in Adaptive Web sites
- Determining quality of Web pages



Starting from Page P1, probability to traverse:

Link (P1-P2) $= \frac{7}{(7+1)} = \frac{7}{8} >$ Link (P1-P3) $= \frac{1}{(7+1)} = \frac{1}{8}$

Temporal Evolution of the Web



- The Internet Archive is a valuable source of data about the (largely structural aspects) of the Web's evolution www.thewaybackmachine.org
- Usage data history is available at individual sites
- Issues to be investigated
  - effect of Web structure on Web usage
  - metrics of evolution
    - structural properties that change/are invariant
    - rate of change
- Mining "interesting" usage patterns over time

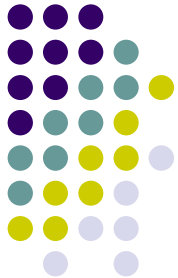
Mining Information from E-mails

Kind of Data available: Content, Usage, evolving Network

Applications

- Target Marketing
  - Source for multi-channel purchases
  - Tracking user interests and purchasing behavior.
  - Increase level of personalization, (e.g. women are found to be more receptive to promotions and discounts
- Social Networks
  - Identifying communities and their interests

Fraud at E-tailer A.com



- The Setup
  - A.com is known for it's attention to customer service
  - A.com decides to create a 'marketplace' where small vendors can sell their wares
  - Customer concern is addressed by A.com agreeing to provide up to $250 cash back if service by partner is not satisfactory
- The Sting
  - Perpetrator P signs up as vendor P.com, and advertises he has 10 VCRs to sell
  - P also signs up as 10 customers C1, C2, ... who all 'buy' from P
  - 7 of the 'customers' complain to A.com that they did not receive their VCRs
  - A.com pays out $250 each to 4 of the customers before discovering the sting

Fraud at On-line Auctioneer e.com

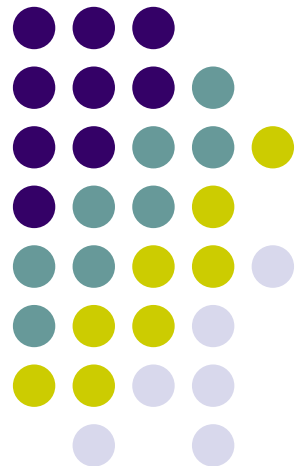
- Auctioneer e.com creates the ultimate ‘virtual flea market’
- Gains immense traction
  - Participation in large numbers
  - People spend large amounts of time
  - Popular for similar reasons as gambling and game shows

- Enter perpetrator P whose
  - Core competencies are product catalog & expediting payment
  - But NOT product delivery
- Buyers complain to e.com, who lowers ‘reputation rating’ of P
- P changes identity to Q

Other Threats

- Identity theft
- Defamation
- Industrial espionage
- Ransom
  - Hacker threat to CD Universe in March 2000
- Vandalism
- Market manipulation through ‘hot stock tips’
  - ‘analyst reports’ don’t seem to be much better either ☺

Web mining and privacy



Public attitude to privacy

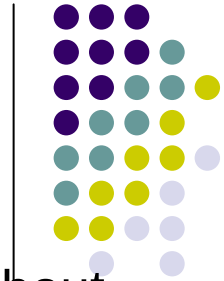
- A (self-professed) non scientific study carried out by a USA Today reporter
- Asked 10 people the following two questions
  - Are you concerned about privacy? 8 said YES
  - If I buy you a Big Mac, can I keep the wrapper (to get fingerprints)? 8 said YES
- ACM E-Commerce 2001 paper [Spiekermann et al]
- Most people willing to answer fairly personal questions to anthropomorphic web-bot, even though not relevant to the task at hand
- Different privacy policies had no impact on behavior
- Study carried out in Europe, where privacy consciousness is (presumably) higher

Public Attitude (contd.)

- Amazon.com (and practically every commercial site) uses cookies to identify and track visitors
  - 97.6% of Amazon.com customers accepted cookies
- Airline frequent flier programs with cross promotions
  - We willingly agree to be tracked
  - Get upset if the tracking fails!
- Over 2 million people have trusted financial information aggregation services with the names and passwords of their financial accounts (bank accounts, 401K accounts, etc.) in less than 3 years months
- Adoption rate has been over 3 times the most optimistic projections
- Imagine the exposure!

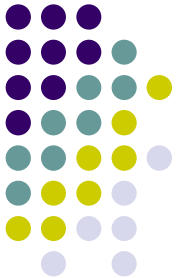
Medical data is (perhaps) an exception to this

Why this attitude? – some guesses



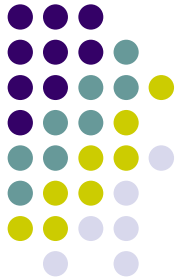
- People don't even know that so much data is being collected about them – e.g. approx. 30GB/day of click-stream data per day at Amazon.com two years ago
- Even if they knew, most people cannot even begin to comprehend the implications of modern day data collection and analysis – 'Database Nation' by Stimson
- Blissful ignorance, or 'what you don't know won't hurt you'
- 'It can't happen to me; it only happens to stupid people' attitude
- Not sufficient prosecution of crimes
  - Some degree of 'David vs. Goliath' syndrome towards cyber crimes
  - 'Computer Capers' by Don B. Parker
  - Though this is changing rapidly

What if people understood the implications? – another BIG guess



- Some people will not share any kind of private data at any cost – the ‘paranoids’
- Some people will share any data for returns – the ‘Jerry Springerites’
- The vast majority in the middle wants
  - a reasonable level of comfort that private data about them will NOT be misused
  - Tangible and compelling benefits in return for sharing their private data – Big Mac example, frequent flier programs

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
  - …

Conclusion

- 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

References

- [ALEX] Alexa Research, http://www.alexa.com/.
- [AMZNa] Amazon.com, www.amazon.com.
- [AOLa] America Online, www.aol.com.
- [BEA] BEA Weblogic Server, http://www.bea.com/products/weblogic/server/index.shtml.
- [BKM+2000] A. Broder et al, Graph Structure in the Web. In the Proc. 9th WWW Conference 2000.
- [BL1999] J. Borges, M. Levene, "Mining Association Rules in Hypertext Databases", in Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), New York City, 1998.
- [BP1998] S. Brin, L. Page, "The anatomy of a large-scale hyper-textual Web search engine". In the 7th International World Wide Web Conference, Brisbane, Australia, 1998.
- [BV] Broadvision 1-to-1 portal, http://www.bvportal.com/.
- [C2001] D. Clark, Shopbots become agents for business change, IEEE Computer, 18-21.
- [C2000] R. Cooley, "Web Usage Mining: Discovery and Usage of Interesting Patterns from Web Data", Ph.D. Thesis, University of Minnesota, Computer Science & Engineering, 2000.
- [C2002] E. Colet, "Using Data Mining to Detect Fraud in Auctions", DSStar, 2002.
- [CMS1997] R. Cooley, B. Mobasher, J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web", in Proceedings of the 9th IEEE International Conference on Tools With Artificial Intelligence (ICTAI '97), Newport Beach, CA, 1997.

References

- [CWLD2002] T. Spring, “Google Launches News Service”, PC World, September 2002, http://www.computerworld.com/developmenttopics/websitemgmt/story/0,10801,74470,00.html.
- [CMS1999] R. Cooley, B. Mobasher, J. Srivastava, “Data Preparation for Mining World Wide Web Browsing Patterns”, Knowledge and Information Systems, 1(1), 1999.
- [CPCP2001] E.H. Chi, P. Pirolli, K. Chen and J. Pitkow, “Using Information Scent to Model User Information Needs and Actions on the Web”. In the Proc. Of ACM CHI 2001. Conference on Human Factors in computing systems, pp490-497. ACM Press April 2001, Seattle WA.
- [D1999] W. Dong, “ “, M.S. Thesis, University of Minnesota, Computer Science & Engineering, 1999.
- [DCLKa] DoubleClick’s DART Technology, http://www.doubleclick.com/dartinfo/.
- [DCLKb] DoubleClick’s Lawsuit, http://www.wired.com/news/business/0,1367,36434,00.html.
- [DCLKc] DoubleClick’s DART for Advertisers: Best Practices in Ad Management and Ad Serving for Agencies and Advertisers (Insight 2002) , http://www.doubleclick.com/us/knowledge/documents/best_practices/bp_adserving_dfa_insight_0205.pdf
- [DG2000] C. Dembeck, P. A. Greenberg, “Amazon: Caught Between a Rock and a Hard Place”, E-Commerce Times, Spetember 8, 2000, http://www.ecommercetimes.com/perl/story/2467.html.
- [DSKT2002] P. Desikan, J. Srivastava, V. Kumar, P.-N. Tan, “Hyperlink Analysis – Techniques & Applications”, Army High Performance Computing Center Technical Report, 2002.

References

- [E1995] D.Eichmann, Ethical Web Agents, Computer Networks and ISDN Systems, 28(1), Elsevier Science.
- [E1996] O. Etzioni, “The World Wide Web: Quagmire or Gold Mine”, in Communications of the ACM, 39(11):65-68, 1996.
- [ERC+2000] Kemal Efe, Vijay Raghavan, C. Henry Chu, Adrienne L. Broadwater, Levent Bolelli, Seyda Ertekin (2000), The Shape of the Web and Its Implications for Searching the Web, International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet- Proceedings at http://www.ssgrr.it/en/ssgrr2000/proceedings.htm, Rome. Italy, Jul.-Aug. 2000.
- [EBAYa] eBay Inc., www.ebay.com.
- [FF1956] L.R. Ford Jr and D.R. Fulkerson, “Maximal Flow through a network.” Canadian J. Math.,8:399-404, 1956.
- [FLG2000] G.W. Flake, S. Lawrence, C.L. Giles, “Efficient identification of Web Communities”. Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2000. pp 150-160.
- [G2000] L. Graham, Keep Your Bots to Yourself, IEEE Software, 17(6):106-107.
- [GOOGa] Google Inc. http://www.google.com/
- [GOOGb] http://www.google.com/press/pressrel/b2b.html
- [GOOGc] http://news.google.com/
- [GS2001] J. Ghosh, J. Srivastava, ed. Proceedings of “Workshop on Web Mining”, Chicago, IL, 2001, http://www.lans.ece.utexas.edu/workshop_index.htm.
- [GS2002] J. Ghosh, J. Srivastava, ed. Proceedings of “Workshop on Web Analytics”, Arlington, VA, 2002, http://www.lans.ece.utexas.edu/workshop_index2.htm.

References

- [IA] The Internet Archive Project, http://www.archive.org/.
- [KLM1996] M. Koster, “Robot Exclusion Standard Revisited”,, http://www.kollar.com/robots.html
- [K1995] M. Koster, “Robots in the Web: Threat or Treat”, ConneXions, 9(4), 1995.
- [K1998] J.M. Kleinberg “Authoritative Sources in Hyperlinked Environment”, In Proc. Of Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, 1998.
- [K2001] R. Kohavi, “Mining E-Commerce Data: The Good, the Bad, the Ugly”, Invited Industrial presentation at the ACM SIGKDD Conference, San Francisco, CA, 2001, http://robotics.stanford.edu/users/ronnyk/kddITrackTalk.pdf.
- [K2002] R.H. Katz, “Pervasive Computing: It’s All About Network Services”, Keynote Address, Pervasive 2002, Zurich, Switzerland, 2002, http://www.cs.berkeley.edu/~randy/Talks/Pervasive2002.ppt.
- [KB2000] R. Kosala, H. Blockeel, “Web Mining Research: A Survey”, in SIGKDD Explorations 2(1), ACM, July 2000.
- [KMSS2001] R. Kohavi, B. Masand, M. Spiliopoulou, J. Srivastava, ed. Proceedings of “WebKDD2001 – Mining Log Data Across All Customer Touchpoints”, San Francisco, CA, 2001, http://robotics.stanford.edu/~ronnyk/WEBKDD2001/.
- [KSS2000] R. Kohavi, M. Spiliopoulou, J. Srivastava, ed. Proceedings of “WebKDD2000 – Web Mining for E-Commerce – Challenges & Opportunities”, Boston, MA, 2000, http://robotics.stanford.edu/~ronnyk/WEBKDD2000/.
- [LDEKST2002] A. Lazarevic, P. Dokas, L. Ertoz, V. Kumar, J. Srivastava, P.N. Tan, “Data Mining for Network Intrusion Detection”, NSF Workshop on Next Generation Data Mining, Baltimore, MD, 2002.

References

- [M2002] C.Mohan, "Dynamic e-Business: Trends in Web Services", Invited Talk at the 3rd VLDB Workshop on Technologies for E-Services (TES'02), Hong Kong, China.
- [M2001] E. Morphy, “Amazon Pushes ‘Personalized Store for Every Customer’ “, E-Commerce Times, September 28, 2001, http://www.ecommercetimes.com/perl/story/13821.html
- [MBNL1999] S. Madria, S.S. Bhowmick, W.K. Ng, E.-P. Lim, “Research Issues in Web Data Mining”, in Proceedings of Data Warehousing and Knowledge Discovery, First International Conference, DaWaK 1999, pp 303-312.
- [MLN2000] Chuang-Hue Moh, Ee-Peng Lim, Wee Keong Ng, “DTD-Miner: A Tool for Mining DTD from XML Documents”, WECWIS 2000: 144-151.
- [MMETa] Jupiter Media Metrix, “Top 50 US Web and Digital Properties”, April 2002, http://www.jmm.com/xp/jmm/press/mediaMetrixTop50.xml.
- [MS1999] B. Masand, M. Spiliopoulou, ed. Proceedings of “WebKDD1999 – Workshop on Web Usage Analysis and User Profiling”, San Diego, CA 1999, http://www.wiwi.hu-berlin.de/~myra/WEBKDD99/.
- [MSB2001] B. Mobasher, M. Spiliopoulou, B. Berendt, “ “, Proceedings of the SIAM Web Analytics Workshop, Chicago, IL, 2001.
- [MSSZ2002] B. Masand, M. Spiliopoulou, J. Srivastava, O. Zaiane, ed. Proceedings of “WebKDD2002 – Web Mining for Usage Patterns and User Profiles”, Edmonton, CA, 2002, http://db.cs.ualberta.ca/webkdd02/.
- [ONL2002] Kok-Leong Ong, Wee Keong Ng, Ee-Peng Lim, “Mining Relationship Graphs for Effective Business Objectives”, PAKDD 2002: 561-566.

References

- [PNM1995] M. Pazzani, L. Bguyen, S. Mantik “Learning from Hotlists and Coldlists – Towards a WWW Information Filtering and Seeking Agent”, in Proceedings of the IEEE International Conference on Tools with AI, 1995.
- [PMB1996] M. Pazzani, J. Muramatsu, D. Billsus, “Syskill and Webert: Identifying Interesting Web Sites”, in Proceedings of AAAI/IAAI Symposium, 1996.
- [PBMW1998] L. Page, S. Brin, R. Motwani and T. Winograd “The PageRank Citation Ranking: Bringing Order to the Web” Stanford Digital Library Technologes, 1999-0120, January 1998.
- [PE1999] M. Perkowitz, O. Etzioni, “Adaptive Web Sites: Conceptual Cluster Mining”, in Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, Sweden, 1999.
- [PPT+2002] B.Prasetyo, et. al., “Naviz: User Behavior Visualization of Dynamic Page”, Pacific-Asia Conference on Knowledge Discovery and Data Mining 2002, Taipei, Taiwan
- [PSS2001] A. Pandey, J. Srivastava, S. Shekhar, “A Web Intelligent Prefetcher for Dynamic Pages Using Association Rules – A Summary of Results, SIAM Workshop on Web Mining, 2001.
- [PT1998] B. Padmanabhan, A. Tuzhilin, “A Belief-Driven Method for Discovering Unexpected Patterns”, in Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, New York, NY, 1998.
- [S1999] M. Spiliopoulou, “Data Mining for the Web”, Proceedings of the Symposium on Principles of Knowledge Discovery in Databases (PKDD), 1999.
- [S2000] D. Scarponi, “Blackmailer Reveals Stolen Internet Credit Card Data”, Associated Press, January 10, 2000, http://abcnews.go.com/sections/world/DailyNews/internet000110.html.

References

- [SM2003] C.M.Sperberg-McQueen, “Web Services and W3C”,Aug2003 http://w3c.dstc.edu.au/presentations/2003-08-21-web-services-interop/msm-ws.html
- [S2002] T. Springer, “Google LaunchesNews Service”, PC World, September 23, 2002, http://www.computerworld.com/developmenttopics/websitemgmt/story/0,10801,74470,00.html.
- [SGB2001] Spiekermann, S., Grossklags, J., & Berendt, B., “E-privacy in 2nd generation E-Commerce: privacy preferences versus actual behavior”, in Proceedings of the ACM Conference on Electronic Commerce (EC'01), Tampa, FL, 14-17 October 2001.
- [SCDT2000] J. Srivastava, R. Cooley, M. Deshpande and P-N. Tan. “Web Usage Mining: Discovery and Applications of usage patterns from Web Data”, SIGKDD Explorations, Vol 1, Issue 2, 2000.
- [SM1997] J. Srivastava, B. Mobasher, Panel discussion on “Web Mining: Hype or Reality?” at the 9th IEEE International Conference on Tools With Artificial Intelligence (ICTAI ’97), Newport Beach, CA, 1997.
- [TK2000] Pang-Ning Tan, Vipin Kumar, Modeling of Web Robot Navigational Patterns, WebKDD 2000: Web Mining for E-Commerce, Boston, MA, August 20 (2000).
- [TK2002] Pang-Ning Tan, Vipin Kumar, Discovery of Web Robot Sessions based on their Navigational Patterns, DMKD, 6(1): 9-35 (2002).
- [U2000] P. Underhill, “Why We Buy: The Science of Shopping”, Touchstone Books, New York, 2000.
- [USDoJ2002] United States Department of Justice Press Release, “Man Sentenced for eBay Auction Fraud of Certain Rare Baseball and Basketball Card Sets”, March 4, 2002, http://www.cybercrime.gov/wildmanSent.htm.
- [VIGN] Vignette StoryServer, http://www.cio.com/sponsors/110199_vignette_story2.html.

References

- [WHN2002] Web Host News, “Hosting Firm Reports Continued Growth”, May 2002, http://thewhir.com/marketwatch/ser053102.cfm.
- [WL1998] K. Wang and H. Lui, “Discovering Typical Structures of Documents: A Road Map Approach”, in Proceedings of the ACM SIGIR Symposium on Information Retrieval, 1998.
- [YHOOa] Yahoo!, Inc. www.yahoo.com.
- [YODLa] Yodlee, Inc. www.yodlee.com