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

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.

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 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

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

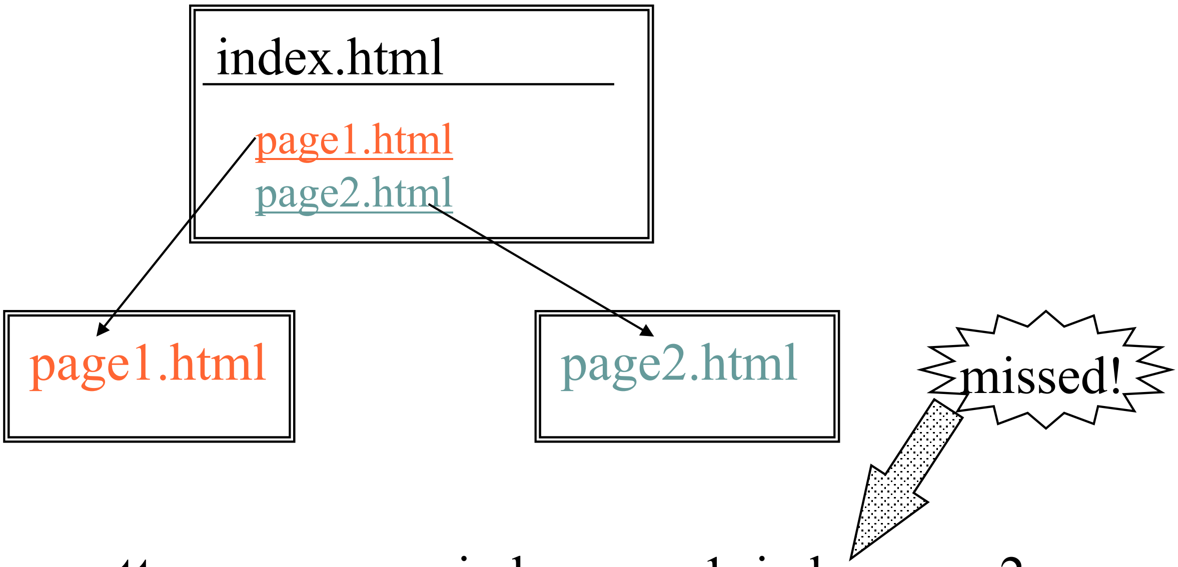
Session Identification Solutions

- 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

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

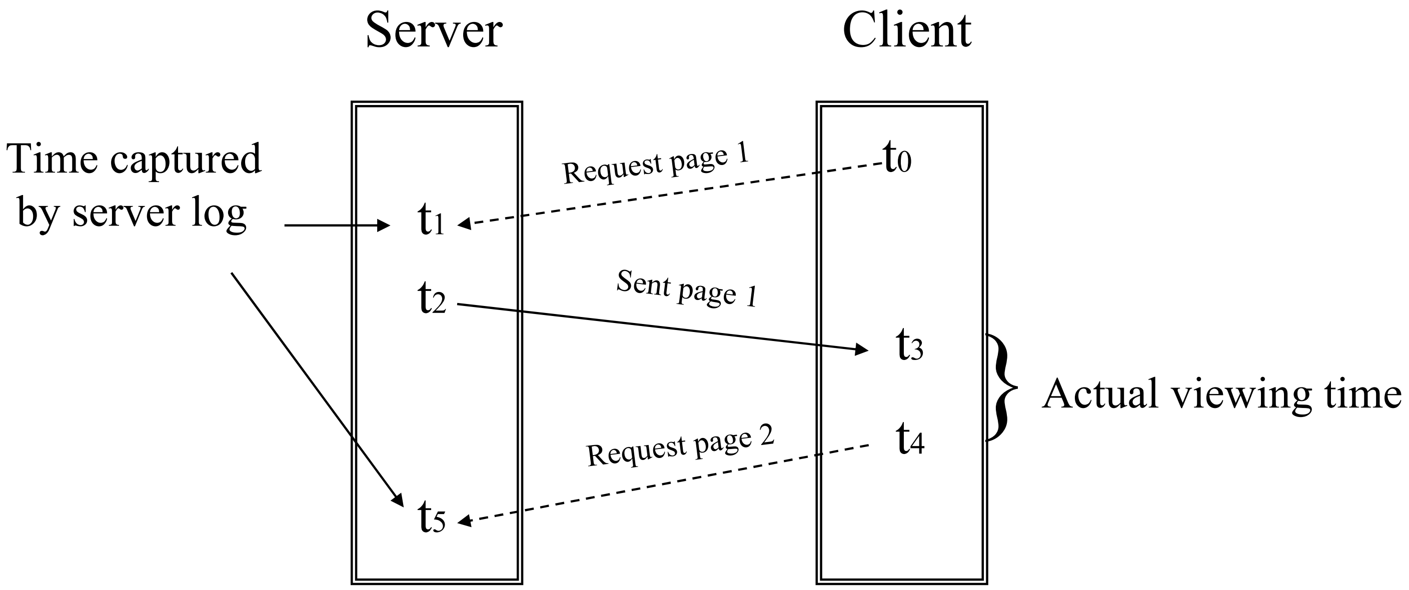
Server Log Incompleteness due to Caching



Access pattern: index, page1, index, page2

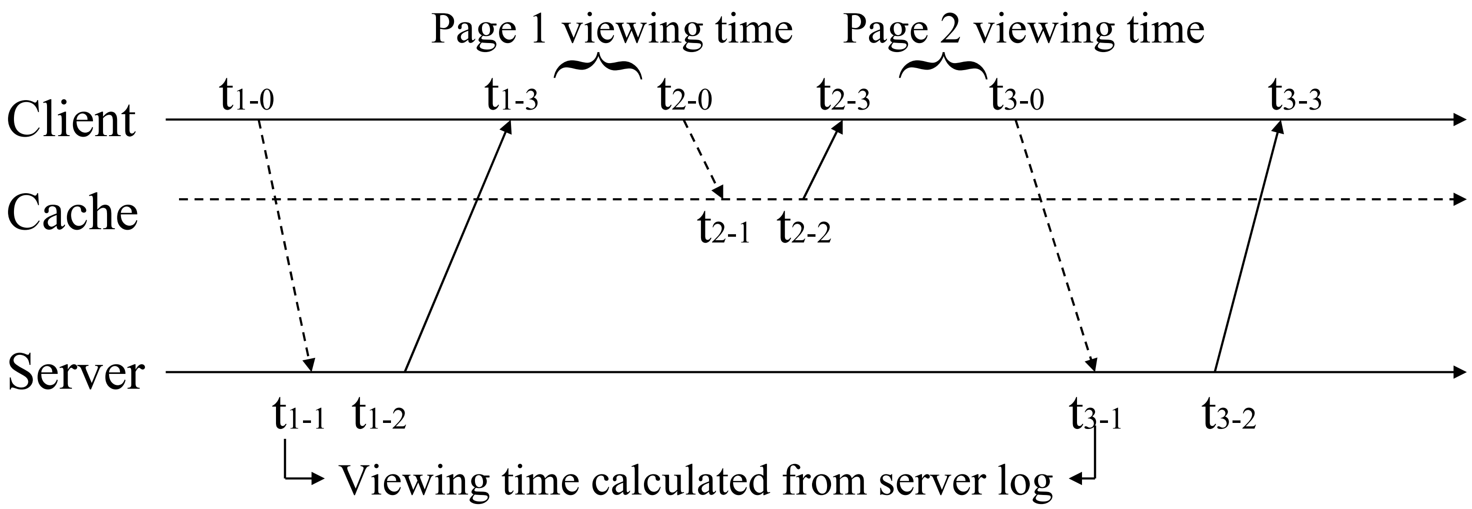
Record in server log:: index, page1, page2

Wrong Access Timings Recorded at Server



Missed Page Views at Server

- Viewing time for cached pages



© Jaideep Srivastava

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

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.

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

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

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