CS454 Topics in Advanced Computer Science

Web Usage Mining

Chengyu Sun
California State University, Los Angeles

Web Data

- Content
- Structure
- Usage
- User profile

Web Usage Mining

- Application of data mining techniques to discover usage patterns from web data

What Can Web Usage Mining Do?

- Statistical analysis
- Recommendation
- Caching
- Improve web site design
- Identify user groups and interests
- Provide market intelligence
- …

How Does Web Usage Mining Do it?

- Data collection
- Preprocessing
- Pattern discovery
- Pattern analysis

Data Collected

- User interaction with a web site
  - Page requested, request parameter, IP address, time stamp …
- User interaction with a web page
  - Mouse clicks, keyboard input, window resizing and scrolling …

Data Sources

- Client
  - JavaScript embedded in web pages
  - Browser modification
- Server
  - Server log
  - Packet sniffer
- Proxy
  - Proxy cache
  - Specialized proxy

Sample HTTP Server Log

74.6.22.167 - - [21/Jun/2009:08:38:33 -0700]
"GET /csns/download.html?fileId=2082676 HTTP/1.0"
200 399223 "-"
"Mozilla/5.0 (compatible; Yahoo! Slurp/3.0; http://help.yahoo.com/help/us/ysearch/slurp)"

- Client IP
- Time stamp
- Response
  - Code
  - Length
- Request
  - Method
  - URI
  - Protocol
  - Headers
    - User-Agent

Preprocessing

- Data filtering
- Page views vs. page requests
- Identify users
- Identify sessions
- Add content and/or structural information
- Data formatting

Pattern Discovery – Association Rules

$\{P_1,P_2\} \Rightarrow P_3$

Users who visited page P_1 and P_2 are likely to visit P_3.

- Typical applications
  - Recommendation
  - Caching

Pattern Discovery – Sequential Pattern

To get to page P_3 from page P_1, users usually take the path $P_1 \rightarrow P_4 \rightarrow P_5 \rightarrow P_3$ instead of $P_1 \rightarrow P_2 \rightarrow P_3$.

- Typical applications
  - Improve web site design

Pattern Discovery – Classification

Users who visited page P_1 and P_2 but not P_3 are likely to be female in the 18-25 age group.

- Typical applications
  - User profiling
  - Market intelligence

Pattern Discovery – Clustering

User clusters: users who demonstrated similar web browsing patterns.
Page clusters: pages that have related content.

- Typical applications
  - Identify user groups and interests
  - Recommendation
  - Content analysis

Pattern Discovery – Probabilistic Modeling

At page P_1, the probability of a user going to visit P_2 is 75%, and the probability of visiting P_3 is 25%.

- Typical applications
  - User action prediction
  - Web traffic prediction
  - Simulation

Pattern Analysis

- Interpret patterns
- Visualize patterns
- Efficient storage, query, and analysis of patterns (like a data warehouse for patterns)

Web Usage Mining in Action

- *Discovery of Significant Usage Patterns from Clusters of Clickstream Data*, by Lin Lu, Margaret Dunham, and Yu Meng

Data

- jcpenny.com's web log on 10/5/2003
- 1,463,180 sessions
- 593,223 user sessions
- 4000 sessions used in experiments
  - 2000 sessions with purchase
  - 2000 sessions without purchase

Frequent Navigation Patterns – The Naïve Approach

- Preprocessing web log
  - Remove entries generated by web crawlers
  - Group page requests into sessions
    - E.g. (p_1,p_2,p_3,p_4), (p_2,p_4), (p_2,p_5,p_4) …
- Pattern discovery
  - Apply a frequent sequential pattern discovery algorithm

Problems with the Naïve Approach …

<u>Pages</u>

p_0: placing order page
p_1: list of all CPU products
p_2: product description of Intel P4 processor
p_3: list of all video cards
p_4: product description of Nvidia 260 video card
P_5: product description of ATI 4860 video card

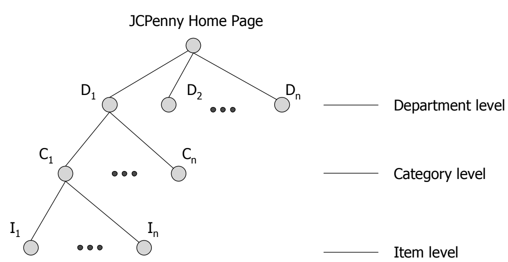
<u>Sessions</u>

s_1: (p_1, p_2, p_0)
s_2: (p_3, p_4, p_0)
s_3: $(p_3, p_4, p_3, p_5, p_0)$
s_4: $(p_1, p_2, p_3, p_4, p_3, p_5)$

… Problems with the Naïve Approach

- Should s_1 and s_2 be consider the same?
- Should s_2 and s_3 be consider similar?
- How do we define session *similarity*?
- Should s_4 be consider together with the other sessions?

Session Abstraction I – Concept Hierarchy



Session Abstraction I – Abstracted Session

Example of an abstracted session:

D0|C875|I, D0|C875|I, P27593, P27592, P28, -507169015

- Item IDs are ignored
- General pages that do not belong to the concept hierarchy are abstracted as `P`

Sequence Similarity – Edit Distance

```
b r i t n e y
      ⇕
b r i t t a n i
```

- The minimum number of operations (insertion, deletion, and substitution) needed to transform one sequence to the other

Sequence Similarity – Sequence Alignment

```
b r i t - - n e y

b r i t t a n - i
```

- The alignment score is a weighted sum of the *similarity* of matching parts

Page Similarity

Page 1: D0|C875|I weight=6+1+4+1+2
Page 2: D0|C875 weight=6+1+4+1

Similarity=12/14=0.857

- The similarity of two web pages is the ratio of the sum of the weights of the matching parts to the total weight

Needleman-Wunsch Alignment Algorithm

- Consider two sequences $X_1 \ldots X_i$ and $Y_1 \ldots Y_j$, the optimal alignment score $A(i,j)$ is the maximum of the following
  - $A(i-1,j-1)+s(X_i,Y_j)$
  - $A(i-1,j)+d$
  - $A(I,j-1)+d$

$s(X_i,Y_j)$ is the similarity between X_i and Y_j, and d is the score of aligning X_i or Y_j with a gap.

Compute Optimal Alignment Score

		Y_1	…	Y_{j-1}	Y_j	…	Y_n
	0	d	…	(j-1)*d	j*d	…	n*d
X_1	d						
…	…						
X_{i-1}	(i-1)*d			A(i-1,j-1)	A(i-1,j)		
X_i	i*d			A(i,j-1)	A(i,j)		
…	…						
X_m	m*d						A(m,n)

Optimal Alignment Computation Example

S_1: P47104, D0|C0|I, D469|C469, D2652|C2652
S_2: D469|C16758|I, D0|C0|I, D469|C469

		P47104	D0|C0|I	D469|C469	D2652|C2652
	0	-10	-20	-30	-40
D469|C16758|I	-10				
D0|C0|I	-20				
D469|C469	-30				

d = -10

$s(X_i,Y_j) = -10 + 30 \times \text{Page\_Similarity}$

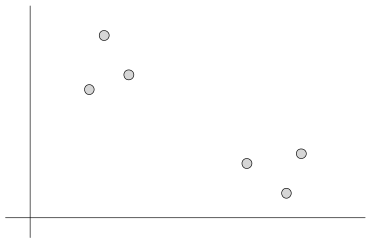
Session Similarity

$$\text{Session Similarity} = \frac{\text{Optimal alignment score}}{\text{Length of the longer session}}$$

Session Clustering

- Nearest Neighbor Clustering Algorithm
  - Given sessions $\{s_1,s_2,\ldots,s_n\}$ and a distance threshold
  - Start with $\{s_1\}$ as a cluster
  - For each remaining session
    - Find the shortest distance to a session that is already in a cluster
    - If the distance is less than or equal to the distance threshold, merge into the cluster; otherwise create a new cluster

Nearest Neighbor Clustering Example



Session Abstraction II

D7107|C7121, D7107|C7126|I076, D7107|C7121, P96, P27

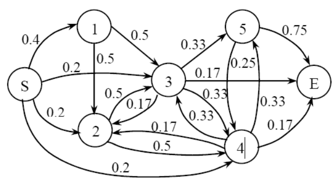
⇩

C1, I1, C1, P1, P2

- Keep only the lowest concept level
- Each page is assigned a locally (i.e. within session) unique id

Markov Model

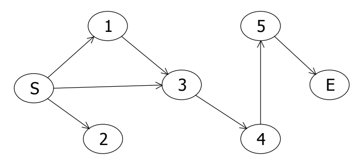
(a) 1,2,3,5,4
(b) 2,4,3,5
(c) 3,2,4,5
(d) 1,3,4,3
(e) 4,2,3,4,5



- Each page is considered a state
- Add a start and an end state
- Calculate transition probability

Markov Model Construction Example

(a) 1,2,3,5,4
(b) 2,4,3,5
(c) 3,2,4,5
(d) 1,3,4,3
(e) 4,2,3,4,5



Significant Usage Patterns (SUP)

Path: $S_1 \rightarrow S_2 \rightarrow \ldots \rightarrow S_n$

Probability of a path: $P = \prod_{i=1}^{n-1} P_{t_i}$

Normalized Probability of a path: $P_N = \left(\prod_{i=1}^{n-1} P_{t_i} \right)^{\frac{1}{n-1}}$

- A SUP is a path that may have a specific beginning and/or end state, and its normalized probability is greater than a given threshold

Experimental Results …

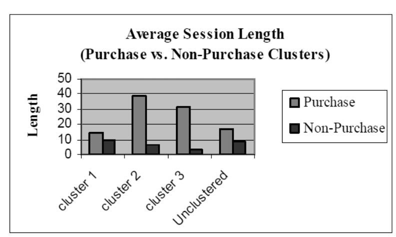


Fig 6. Average session length

… Experimental Results …

Table 3. SUPs in non-purchase cluster

Cluster No.	No. of Sessions	Threshold (θ)	Average Session Length	No. of States	SUPs
1	1746	0.3	9.6	98	1. $S-C_1-C_1-C_2-C_3-C_4-C_5-C_6-C_7-E$ 2. $S-C_1-C_1-C_2-C_3-C_4-C_5-E$ 3. $S-C_1-C_1-C_2-C_3-E$ 4. $S-C_1-C_2-C_3-C_3-C_4-C_5-C_6-C_7-E$ 5. $S-C_1-C_2-C_3-C_4-C_4-C_5-C_6-C_7-E$ 6. $S-C_1-C_2-C_3-C_4-C_5-C_5-C_6-C_7-E$ 7. $S-C_1-C_2-C_3-C_4-C_5-C_6-C_6-C_7-E$ 8. $S-C_1-C_2-C_3-C_4-C_5-C_6-C_7-C_7-E$ 9. $S-C_1-C_2-C_3-C_4-C_5-C_6-C_7-C_8-E$ 10. $S-C_1-C_2-C_3-C_4-C_5-C_6-C_7-E$ 11. $S-C_1-C_2-C_3-C_4-C_5-C_6-E$ 12. $S-C_1-C_2-C_3-C_4-C_5-E$ 13. $S-C_1-C_2-C_3-C_4-E$ 14. $S-C_1-C_2-C_3-E$

… Experimental Results

2	241	0.37	6.6	38	1. $S-P_1-P_2-P_3-P_3-E$ 2. $S-P_1-P_2-P_3-P_4-P_4-P_5-E$ 3. $S-P_1-P_2-P_3-P_4-P_4-E$ 4. $S-P_1-P_2-P_3-P_4-P_5-P_4-E$ 5. $S-P_1-P_2-P_3-P_4-P_5-P_5-E$ 6. $S-P_1-P_2-P_3-P_4-P_5-P_6-C_1-E$ 7. $S-P_1-P_2-P_3-P_4-P_5-P_6-P_7-E$ 8. $S-P_1-P_2-P_3-P_4-P_5-P_6-E$ 9. $S-P_1-P_2-P_3-P_4-P_5-E$ 10. $S-P_1-P_2-P_3-P_4-C_1-E$ 11. $S-P_1-P_2-P_3-P_4-E$ 12. $S-P_1-P_2-P_3-C_1-E$ 13. $S-P_1-P_2-P_3-E$ 14. $S-P_1-P_2-E$
3	13	0.3	3.0	6	1. $S-C_1-P_1-P_1-P_2-E$ 2. $S-C_1-P_1-P_1-E$ 3. $S-C_1-P_1-P_2-E$ 4. $S-C_1-P_1-E$ 5. $S-I_1-P_1-P_1-P_2-E$ 6. $S-I_1-P_1-P_1-E$ 7. $S-I_1-P_1-P_2-E$ 8. $S-I_1-P_1-E$

Summary

- Session abstraction I
- Similarity measure: sequence alignment
- Clustering: nearest neighbor
- Session abstraction II
- Markov model construction (per cluster)
- Significant Usage Pattern