Integrating Web Usage and Content Mining for More Effective Personalization

Authors:
Bamshad Mobasher, Hoghua Dai, Tao Luo, Yuqing Sun, and Jiang Zhu
(Depaul University, Chicago, Illinois, USA)

Presented by

Ateeq Abdul Rauf	**03020039**
Haroon Rashid	**04020064**

Agenda

- Introduction
- A Web Mining Framework for Personalization
  - (a) System Architecture
  - (b) Data Preparation for Usage and Content Mining
  - (c) Discovery of Aggregate Usage Profiles
  - (d) Discovery of Content Profiles
- Integrating Content and Usage Profiles for Personalization
- Experimental Results
- Conclusions and Future Work

What is web personalization?

- Web personalization can be defined as any action that tailors the Web experience to a particular user, or set of users.
- The personalized content can take the form of recommended links or products, targeted advertisements, or text and graphics tailored to the user's perceived preferences as determined by the matching usage and content profiles.

Web Usage Mining and Content Mining

Traditional Approach

- The type of input is subjective description of the users by the users themselves, and thus is prone to biases.

Web Usage Mining and Content Mining (continued)

The New Approach

- Web Usage Mining

  Utilizing User's History of Web Usage (browsing) to personalize

- Content Mining

  Using the contents of browsed pages to personalize

Web Usage Mining

The profiles are dynamically obtained from user patterns, and thus the system performance does not degrade over time as the profiles age. Web usage mining will reduce the need for obtaining subjective user ratings or registration-based personal preferences.

Web usage mining can also be used to enhance the effectiveness of collaborative filtering approaches.

Collaborative Filtering Approach

- Collaborative filtering is based on matching, in real-time, the current user's profile against similar records (nearest neighbors) obtained by the system over time from other users.

How Web Usage helps in Collaborative Filtering

- Problem:

  Recent studies show it becomes hard to scale collaborative filtering techniques to a large number of items, while maintaining reasonable prediction performance and accuracy.

- Solution:

  Cluster user records with similar characteristics, and focus the search for nearest neighbors only in the matching clusters. Clustering is done by records obtained through user's history (i.e. web usage)

Problem with Web Usage Mining Alone

- Problem: When little usage data is available pertaining to some objects or when the site content may change regularly.

- Solution: Integration of Usage and Content Mining

The Goal

- To create a uniform representation for both content and usage profiles that can be effectively used by the recommendation engine to perform real-time personalization.

Web Mining Framework for Personalization

- System Architecture
- Two main components
- (a) Offline component:
  Data preparation and specific Web mining tasks
- (b) Online component:
  A real-time recommendation which incorporates offline component

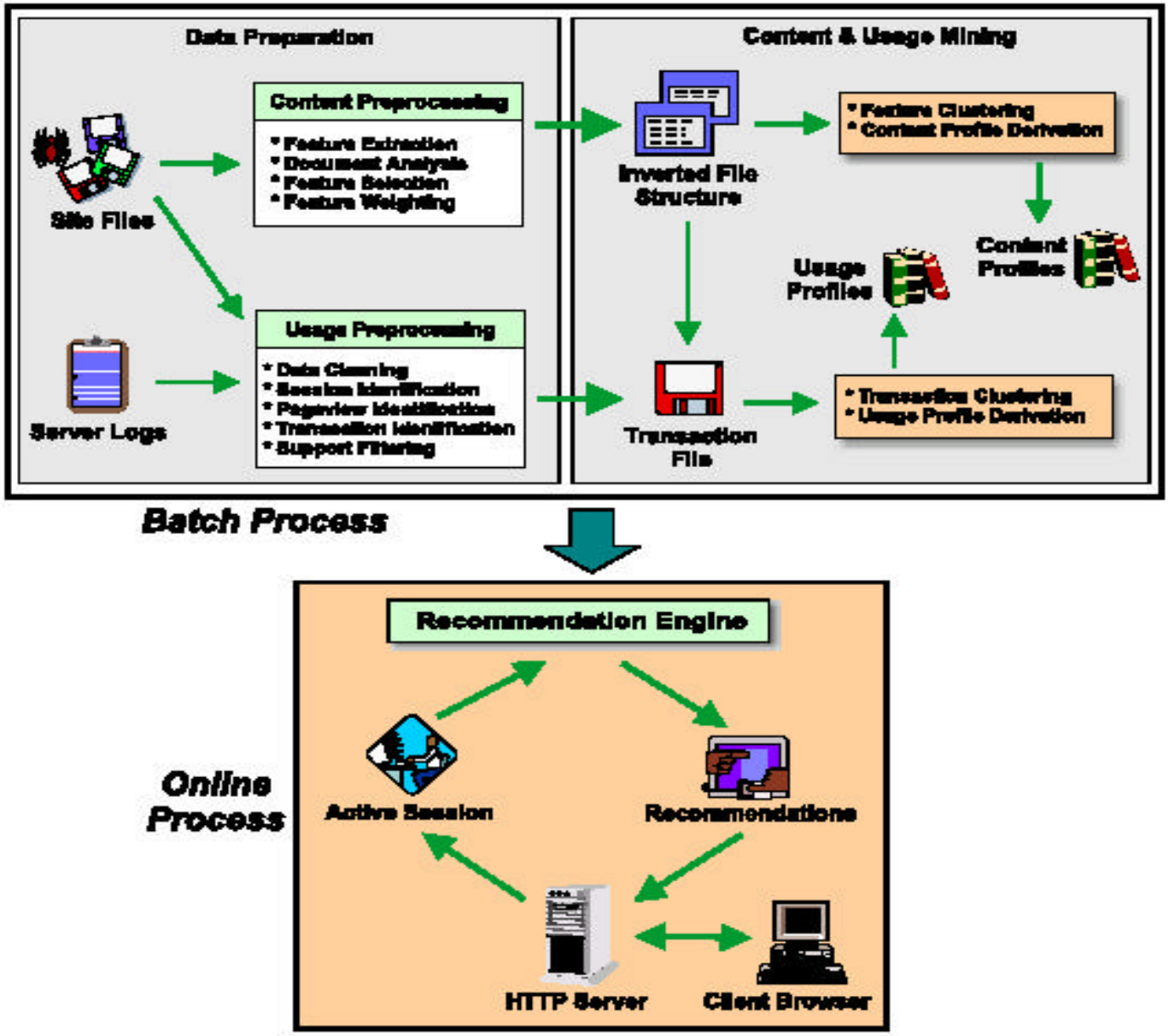

Fig. 1. A general framework for automatic personalization based on Web Mining

Offline Component

Web Usage Mining Components

- Involve the discovery of association rules, sequential patterns, pageview clusters, or transaction clusters

Content Mining Components

- Involve feature clustering (based on occurrence patterns of features in pageviews), pageview clustering based on content or meta-data attributes

Online component

The recommendation engine considers the active server session in conjunction with the discovered patterns and profiles to provide personalized content.

Data Preparation for Usage and Content Mining

- data cleaning,
- user identification,
- session identification*,
- pageview identification,
- path completion (client-side or proxy level caching)

Note: The author describes session Identification through the heuristics proposed in another paper to identify unique user sessions from anonymous usage data and to infer cached references.

Pageview Identification

- Pageview identification is the task of determining which page file accesses contribute to a single browser display.
- The signficance of a pageview may depend on usage, content and structural characteristics of the site, as well as prior domain knowledge specied by the site designer.

Pageview Attributes

Attributes include the pageview id (normally a URL uniquely representing the pageview), duration (for a given user session), static pageview type (e.g., content, navigational, product view, index page, etc.), and other meta-data.

Transaction Identification

- Final PreProcessing before Pattern Generation
- Filtered into 2 Types (noise)
  - V- Low Support
  - V-High Support
- Remove Pages with minimum knowledge
- Remove Patterns of non active users

Usage PreProcessing

Usage PreProcessing

- Weights assigned on following basis
  - Binary weights (product existance & vice versa)
  - Duration of the attatchment with the site.
  - As assigned by the annalyist.

Content PreProcessing

- Extraction of features form the text & meta data.
- Features Extracted from the meta data from
  - XML/HTML tags
  - Textual contents of pages
- Weights assigned to extracted material as given by the site designer.

Content PreProcessing

- Feature Page View Matrix, in which each column is a feature vector corrresponding to a pageView.

Aggregate Usage Profiles

- Need of Aggregate usage profiles
- Must capture overlapping interests
- Usage profiles as weighted collection of page view records

Aggregate Usage Profiles (cont…)

- Usage profiles divided into
  - Ordered profiles:

    if goal is to capture the navigational path
  - Unordered profiles:

    if focus is on capturing the associations among pages

- Helpful in integrating different profiles

Aggregate Usage Profiles (cont…)

- Cluster Formation using clustering algos

Aggregate Usage Profiles (cont…)

- Clusters :

  "a group of user pageviews with similar

  navigational patterns"

- Clusters as such not used for aggregate usage profiles so we make weighted clusters

Aggregate Content Profiles

- The same technique of weighted collections
- Pages with similar contents grouped
- Content profiles also capture overlapping information*.

Aggregate content Profiles (cont…)

- Don’t cluster page views
- Only cluster page contents
- Normalize the values

Integrating Content and Usage Profiles for Personalization

Recommendation Engine

- Online component of the Web personalization system
- Computes a recommendation set of objects that match the current user profile
- Uses history depth
- Sliding Window --- last n pages influence recommendation value
- Structural Characteristics

Calculating Recommendation Scores

- Both content and usage profiles represented as sets of pageview-weight pairs
- Both active session profile and history profile expressed as vectors
- Formulae for C and S
- Rec Score
- U Rec and C Rec

Recommendation Scores

- Take maximal recommendation score of URec and CRec
- Allows either profile to contribute
- Other methods

Experimental Results

- 62 pageviews
- 28 feature clusters
- 566 significant features
- 18430 user transactions
- threshold of 0.5

Experimental Results

Weight	Pageview ID	Significant Features (stems)
1.00	CFP: One World One Market	world challeng busi co manag global
0.69	CFP: Int'l Conf. on Marketing & Development	challeng co contact develop intern
0.35	CFP: Journal of Global Marketing	busi global
0.32	CFP: Journal of Consumer Psychology	busi manag global
Weight	**Pageview ID**	**Significant Features (stems)**
1.00	CFP: Journal of Psych. & Marketing	psychologi consum special market
1.00	CFP: Journal of Consumer Psychology I	psychologi journal consum special market
0.72	CFP: Journal of Global Marketing	journal special market
0.61	CFP: Journal of Consumer Psychology II	psychologi journal consum special
0.50	CFP: Society for Consumer Psychology	psychologi consum special
0.50	CFP: Conf. on Gender, Market., Consumer Behavior	journal consum market

Fig. 2. Two Overlapping Content Profiles

Experimental Results

Pages In Active Session Window	Recommendations	Score
* ACR Board of Directors Meeting	NO RECOMMENDATIONS	
* ACR Board of Directors Meeting	ACR 1999 Annual Conference	0.57
* Conference Update	CFP: ACR'99 Asia-Pacific Conf.	0.53
	CFP: Int'l Conf. on Marketing & Development	0.52
	CFP: ACR'99 European Conf.	0.51
	President's Column - December, 1998	0.50
* Conference Update	ACR News Special Topics	0.64
* CFP: Journal of Psych. & Marketing	ACR 1999 Annual Conference	0.57
	CFP: Int'l Conf. on Marketing & Development	0.53
	CFP: ACR'99 European Conf.	0.53
	CFP: Winter 2000 SCP Conference	0.52
	CFP: Int'l Research Seminar in Marketing	0.50
* CFP: Journal of Psych. & Marketing	ACR News Special Topics	0.69
* CFP: Journal of Global Marketing		

Fig. 3. Recommendations Based on Usage Profiles

Experimental Results

Pages in Active Session Window	Recommendations	Score
* ACR Board of Directors Meeting	Special Topics - ACR Letters and Special Topics	0.70
	Special Topics - ACR Board of Directors Agenda	0.65
	Special Topics - ACR Appointments	0.62
* ACR Board of Directors Meeting * Conference Update	Call for Papers	0.67
	ACR News Updates	0.54
	ACR News Special Topics	0.51
* Conference Update * CFP: Journal of Psych. & Marketing	Call for Papers	0.67
	CFP: Journal of Consumer Psych. I	0.69
	CFP: Journal of Psych. & Marketing	0.66
	CFP: Journal of Consumer Psych. II	0.51
	CFP: Journal of Global Marketing	0.50
* CFP: Journal of Psych. & Marketing * CFP: Journal of Global Marketing	CFP: Journal of Consumer Psych. I	0.77
	CFP: Journal of Psych. & Marketing	0.73
	CFP: Journal of Consumer Psych. II	0.59
	CFP: Marketing & Public Policy Conf.	0.58
	CFP: Conf. on Gender, Marketing , Consumer Behavior	0.54

Fig. 4. Recommendations Based on Content Profiles

Conclusions and Future Work

- The method eliminates subjectivity from profile data and keeps it up-to-date
- Increases usefulness and accuracy of recommendations
- The approach is promising

Question and Answers