

A Case Study of Behavior-driven Conjoint Analysis on Yahoo! Front Page Today Module

Wei Chu, Seung-Taek Park, Todd Beaupre, Nitin Motgi, Amit Phadke,
Seinjuti Chakraborty, Joe Zachariah

Yahoo! Inc.
701 First Avenue
Sunnyvale, CA 94089
{chuwei,parkst,tbeaupre,nmotgi,aphadke,seinjuti,joez}@yahoo-inc.com

ABSTRACT

Conjoint analysis is one of the most popular market research methodologies for assessing how customers with heterogeneous preferences appraise various objective characteristics in products or services, which provides critical inputs for many marketing decisions, e.g. optimal design of new products and target market selection. Nowadays it becomes practical in e-commercial applications to collect millions of samples quickly. However, the large-scale data sets make traditional conjoint analysis coupled with sophisticated Monte Carlo simulation for parameter estimation computationally prohibitive. In this paper, we report a successful large-scale case study of conjoint analysis on click through stream in a real-world application at Yahoo!. We consider identifying users' heterogeneous preferences from millions of click/view events and building predictive models to classify new users into segments of distinct behavior pattern. A scalable conjoint analysis technique, known as tensor segmentation, is developed by utilizing logistic tensor regression in standard partworth framework for solutions. In offline analysis on the samples collected from a random bucket of Yahoo! Front Page Today Module, we compare tensor segmentation against other segmentation schemes using demographic information, and study user preferences on article content within tensor segments. Our knowledge acquired in the segmentation results also provides assistance to editors in content management and user targeting. The usefulness of our approach is further verified by the observations in a bucket test launched in Dec. 2008.

Categories and Subject Descriptors

H.1.0 [Models and Principles]: General; H.3.3 [Information Search and Retrieval]: Information filtering; H.3.5 [Online Information Services]: Web-based services; H.4 [Information Systems Applications]: Miscellaneous

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

KDD'09, June 28–July 1, 2009, Paris, France.

Copyright 2009 ACM 978-1-60558-495-9/09/06 ...\$5.00.

General Terms

Algorithms, Experiments, Designs, Performance

Keywords

Conjoint Analysis, Classification, Clustering, Segmentation, Tensor Product, Logistic Regression

1. INTRODUCTION

Since the advent of conjoint methods in marketing research pioneered by Green and Rao [9], research on theoretical methodologies and pragmatic issues has thrived. Conjoint analysis is one of the most popular marketing research methodologies to assess users' preferences on various objective characteristics in products or services. Analysis of trade-offs, driven by heterogeneous preferences on benefits derived from product attributes, provides critical input for many marketing decisions, e.g. optimal design of new products, target market selection, and product pricing. It is also an analytical tool for predicting users' plausible reactions to new products or services.

In practice, a set of categorical or quantitative attributes is collected to represent products or services of interest, while a user's preference on a specific attribute is quantified by a utility function (also called partworth function). While there exist several ways to specify a conjoint model, additive models that linearly sum up individual partworth functions are the most popular selection.

As a measurement technique for quantifying users' preferences on product attributes (or partworths), conjoint analysis always consists of a series of steps, including stimulus representation, feedback collection and estimation methods. Stimulus representation involves development of stimuli based on a number of salient attributes (hypothetical profiles or choice sets) and presentation of stimuli to appropriate respondents. Based on the nature of users' response to the stimuli, popular conjoint analysis approaches are either choice-based or ratings-based. Recent developments of estimation methods comprise hierarchical Bayesian (HB) methods [15], polyhedral adaptive estimation [19], Support Vector Machines [2, 7] etc.

New challenges on conjoint analysis techniques arise from emerging personalized services on the Internet. A well-known Web site can easily attract millions of users daily. The Web content is intentionally programmed that caters to users'

needs. As a new kind of stimuli, the content could be a hyperlink with text or an image with caption. The traditional stimuli, such as questionnaires with itemized answers, are seldom applied on web pages, due to attention burden on the user side. Millions of responses, such as click or non-click, to content stimuli are collected at much lower cost compared to the feedback solicited in traditional practice, but the implicit feedback without incentive-compatibility constraints are potentially noisy and more difficult to interpret. Although indispensable elements in the traditional settings of conjoint analysis have been changed greatly, conjoint analysis is still particularly important in identifying the most appropriate users at the right time, and then optimizing available content to improve user satisfaction and retention.

We summarize three main differences between Web-based conjoint analysis and the traditional one in the following:

- The Web content may have various stimuli that potentially contain many psychologically related attributes, rather than predefined attributes of interest in traditional experimental design. Meanwhile, most of users are casual or new visitors who declare part or none of their personal information and interests. Since we have to extract attributes or discover latent features in profiling both content stimuli and users, parameter estimation methods become more challenging than that in the traditional situation;
- In feedback collection, most of respondents haven't experienced strong incentives to expend their cognitive resources on the prominent but unsolicited content. This issue causes a relatively high rate of false negative (false non-click);
- The sample size considered in traditional conjoint analysis is usually less than a thousand, whereas it is common in modern e-business applications to observe millions of responses in a short time, e.g. in a few hours. The large-scale data sets make the traditional conjoint analysis coupled with sophisticated Monte Carlo simulation for parameter estimation computationally prohibitive.

In this paper, we conduct a case study of conjoint analysis on click through stream to understand users' intentions. We construct features to represent the Web content, and collect user information across the Yahoo! network. The partworth function is optimized in tensor regression framework via gradient descent methods on large scale samples. In the partworth space, we apply clustering algorithms to identifying meaningful segments with distinct behavior pattern. These segments result in significant CTR lift over both the unsegmented baseline and two demographic segmentation methods in offline and online tests on the Yahoo! Front Page Today Module application. Also by analyzing characteristics of user segments, we obtain interesting insight of users' intention and behavior that could be applied for market campaigns and user targeting. The knowledge could be further utilized to help editors for content management.

The paper is organized as follows: In Section 2, we delineate the scenario under consideration by introducing the Today Module application at Yahoo! Front Page; In Section 3, we review related work in literature; In Section 4, we present tensor segmentation in detail. We report experimental results in Section 5 and conclude in Section 6.



Figure 1: A snapshot of the default “Featured” tab in the Today Module on Yahoo! Front Page, delineated by the rectangular. There are four articles displayed at footer positions, indexed by F1, F2, F3 and F4. One of the four articles is highlighted at the story position. At default, the article at F1 is highlighted at the story position.

2. PROBLEM SETTING

In this section, we first describe our problem domain and our motivations for this research work. Then we describe our data set and define some notations.

2.1 Today Module

Today Module is the most prominent panel on Yahoo! Front Page, which is also one of the most popular pages on the Internet, see a snapshot in Figure 1. The default “Featured” tab in Today Module highlights one of four high-quality articles selected from a daily-refreshed article pool curated by human editors. As illustrated in Figure 1, there are four articles at footer positions, indexed by F1, F2, F3 and F4 respectively. Each article is represented by a small picture and a title. One of the four articles is highlighted at the story position, which is featured by a large picture, a title and a short summary along with related links. *At default, the article at F1 is highlighted at the story position.* A user can click on the highlighted article at the story position to read more details if the user is interested in the article. The event is recorded as a “story click”. If a user is interested in the articles at F2~F4 positions, she can highlight the article at the story position by clicking on the footer position.

A pool of articles is maintained by human editors that incarnates the “voice” of the site, i.e., the desired nature and mix of various content. Fresh stories and breaking news are regularly acquired by the editors to replace out-of-date articles every a few hours. The life time of articles is short, usually just a few hours, and the popularity of articles, measured by their click-through rate (CTR)¹, changes over time. Yahoo! Front Page serves millions of visitors daily. Each visit generates a “view” event on the Today Module, though the visitor may not pay any attention to the Today Module. The users of the Today Module, a subset of the whole

¹CTR of an article is measured by the total number of clicks on the article divided by total number of views on the article in a certain time interval.

traffic, also generate a large amount of “story click” events by clicking at story position for more details of the stories they would like to read. Our setting is featured by dynamic characteristics of articles and users. Scalability is also an important requirement in our system design.

One of our goals is to increase user activities, measured by overall CTR, on the Today Module. To draw visitors’ attention and increase the number of clicks, we would like to rank available articles according to visitors’ interests, and to highlight the most attractive article at the F1 position. In our previous research [1] we developed an *Estimated Most Popular* algorithm (EMP), which estimates CTR of available articles in near real-time by a Kalman filter, and presents the article of the highest estimated CTR at the F1 position. Note that there is no personalized service in that system. i.e. the article shown at F1 is the same to all visitors at a given time. In this work we would like to further boost overall CTR by launching a partially personalized service. User segments which are determined by conjoint analysis will be served with different content according to segmental interests. Articles with the highest segmental CTR will be served to user segments respectively.

In addition to optimizing overall CTR, another goal of this study is to understand users’ intention and behavior to some extent for user targeting and market campaigns. Once we identify users who share similar interests in conjoint analysis, predictive models can be built to classify users (including new visitors) into segments. For example, if we find a user segment who like “Car & Transportation” articles, then we can target this user segment for Autos marketing campaigns. Furthermore, with knowledge of user interests in segments, we can provide assistance to the editors for content management. For example, if we know that most users in a segment who like articles of News visit us in the morning while most users in another segment who like articles about TV come in the evening. Then editors can target these two segments by simply programming more News articles in the morning and more TV-related articles in the evening.

Note that if we only consider the first goal of maximizing overall CTR, personalized recommender systems at individual level might be an appropriate approach to pursue. In our recent research [4], we observed that feature-based models yield higher CTR than conjoint models in offline analysis. However, conjoint models significantly outperform the “one-size-fits-all” EMP approach on the metric of CTR, and also provide actionable management on content and user targeting at segment level. The flexibility supplied by conjoint models is valuable to portals such as Yahoo!.

2.2 Data Collection

We collected three sets of data, including content features, user profiles and interactive data between users and articles.

Each article is summarized by a set of features, such as topic categories, sub-topics, URL resources, etc. Each visitor is profiled by a set of attributes as well, e.g. age, gender, residential location, Yahoo! property usage, etc. Here we simply selected a set of informative attributes to represent users and articles. Gauch et al. [8] gave an extensive review on various profiling techniques.

The interaction data consists of two types of actions: view only or story click for each pair of a visitor and an article. One visitor can only view a small portion of available articles, and it is also possible that one article is shown to the

same visitor more than one time. It is difficult to detect whether the visitors have paid enough attention to the article at the story position. Thus in data collection a large amount of view events are false non-click events. In another words, there is a relatively high rate of false negative (false non-click) in our observations.

There are multiple treatments on users’ reactions in modeling the partworth utility, such as

- Choice-based responses: We only consider whether an article has been clicked by a visitor, while ignoring repeated views and clicks. In this case, an observed response is simply binary, click or not;
- Poisson-based responses: The number of clicks we observed on each article/user pair is considered as a realization from a Poisson distribution;
- Metric-based responses: We consider repeated views and clicks and treat CTR of articles by each user as target.

In the Today Module setting, Poisson-based and metric-based responses might be vulnerable by the high rate of false negative observations. Thus we follow the choice-based responses only in this work.

2.3 Notations

Let index the i -th user as \mathbf{x}_i , a $D \times 1$ vector of user features, and the j -th content item as \mathbf{z}_j , a $C \times 1$ vector of article features. We denote by r_{ij} the interaction between the user \mathbf{x}_i and the item \mathbf{z}_j , where $r_{ij} \in \{-1, +1\}$ for “view” event and “story click” event respectively. We only observe interactions on a small subset of all possible user/article pairs, and denote by \mathcal{O} the set of observations $\{r_{ij}\}$.

3. RELATED WORK

In very early studies [21], homogeneous groups of consumers are entailed by a priori segmentation. For example, consumers are assigned to groups on the basis of demographic and socioeconomic variables, and the conjoint models are estimated within each of those groups. Clearly, the criteria in the two steps are not necessarily related: one is the homogeneity of customers in terms of their descriptive variables and another is the conjoint preference within segments. However, this segmentation strategy is easy to implement, which is still widely applied in industry.

Traditionally, conjoint analysis procedures are of two-stage: 1) estimating a partworth utility function in terms of attributes which represents customers’ preference at individual-level, e.g. via ordinary least squares regression; 2) if segmentation is of interest to marketing, through hierarchical or non-hierarchical clustering algorithms, customers are grouped into segments where people share similar individual-level partworths.

Conjoint studies, especially on the partworth function, depend on designs of stimuli (e.g. product profiles or choices sets on questionnaires) and methods of data collection from respondents. One of challenges in the traditional conjoint analysis is to obtain sufficient data from respondents to estimate partworth utilities at individual level using relatively few questions. The theory of experimental design is adapted for constructing compact profiles to evaluate respondents’ opinion effectively. Kuhfeld et al. [14] studied orthogonal

designs for linear models. Huber and Zwerina [11] brought two additional properties, minimal level overlap and utility balance, into choice-based conjoint experiments. Sandor and Wedel [17] developed experimental designs by utilizing prior information. More references are reviewed by Rao [16] recently.

Hierarchical Bayesian (HB) methods [15] are developed to exploit partworth information of all respondents in modeling the partworth function. The HB models relate the variation in a subject's metric-based responses and the variation in the subjects' partworths over the population as follows:

$$r_{ij} = \beta_i^\top \mathbf{z}_j + \epsilon_{ij} \quad \forall i, j \quad (1)$$

$$\beta_i = \mathbf{W}^\top \mathbf{x}_i + \delta_i \quad \text{for } i = 1, \dots, n \quad (2)$$

where β_i is a C -dimensional partworth vector of the user i and \mathbf{W} is a matrix of regression coefficients that relates user attributes to partworths. The error terms $\{\epsilon_{ij}\}$ and $\{\delta_i\}$ in eq(1) and eq(2) are assumed to be mutually independent and Gaussian random variables with zero mean and covariance matrix $\{\sigma_j^2\}$ and $\mathbf{\Lambda}$ respectively, i.e. $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)$ and $\delta_i \sim \mathcal{N}(0, \mathbf{\Lambda})$.² Together with appropriate prior distributions over the remaining variables, the posterior analysis yields the HB estimator of the partworths as a convex function of an individual-level estimator and a pooled estimator, in which the weights mainly depend on the noise levels $\{\sigma_j^2\}$ and $\mathbf{\Lambda}$ in eq(1) and eq(2). The estimation on these noise levels usually involves the Monte Carlo simulation, such as the Gibbs sampling or Metropolis-Hastings algorithms [20], which is computationally infeasible for our applications with millions of samples.

Huber and Train [10] compared the estimates obtained from the HB methods with those from classical maximum simulated likelihood methods, and found the average of the expected partworths to be almost identical for the two methods in some applications on electricity suppliers. In the past decade, integrated conjoint analysis methods have emerged that simultaneously segment the market and estimate segment-level partworths, e.g. the finite mixture conjoint models [6]. The study conducted by [20] shows the finite mixture conjoint models performed well at segment-level, compared with the HB models.

Toubia et al. [19] developed a fast polyhedral adaptive conjoint analysis method that greatly reduces respondent burden in parameter estimation. They employ "interior point" mathematical programming to select salient questions to respondents that narrow the feasible region of the partworth values as fast as possible.

A recently developed technique on parameter estimation is based on the ideas from statistical learning and support vector machines. The choice-based data can be translated as a set of inequalities that compare the utilities between two selected items. Evgeniou et al. [7] formulated the partworth learning as a convex optimization problem in regularization framework. Chapelle and Harchaoui [2] proposed two machine learning algorithms that efficiently estimate conjoint models from pairwise preference data.

Jiang and Tuzhilin [12] experimentally demonstrated both 1-to-1 personalization and segmentation approaches significantly outperform aggregate modeling. Chu and Park [4] re-

cently proposed a feature-based model for personalized service at individual level. On the Today Module application, the 1-to-1 personalized model outperforms several segmentation models in offline analysis. However, conjoint models are still indispensable components in our system, because of the valuable insight on user intention and tangible control on both content and user sides at segment level.

4. TENSOR SEGMENTATION

In this section, we employ logistic tensor regression coupled with efficient gradient-descent methods to estimate the partworth function conjointly on large data sets. In the users' partworth space, we further apply clustering techniques to segmenting users. Note that we consider the cases of millions of users and thousands of articles. The number of observed interactions between user/article pairs could be tens of million.

4.1 Tensor Indicator

We first define an indicator as a parametric function of the tensor product of both article features \mathbf{z}_j and user attributes \mathbf{x}_i as follows:

$$s_{ij} = \sum_{a=1}^C \sum_{b=1}^D x_{i,b} z_{j,a} w_{ab}, \quad (3)$$

where D and C are the dimensionality of user and content features respectively, $z_{j,a}$ denotes the a -th feature of \mathbf{z}_j , and $x_{i,b}$ denotes the b -th feature of \mathbf{x}_i . The weight variable w_{ab} is independent of user and content features, which represents affinity of these two features $x_{i,b}$ and $z_{j,a}$ in interactions. In matrix form, eq(3) can be rewritten as

$$s_{ij} = \mathbf{x}_i^\top \mathbf{W} \mathbf{z}_j,$$

where \mathbf{W} denotes a $D \times C$ matrix with entries $\{w_{ab}\}$. The partworths of the user \mathbf{x}_i on article attributes is evaluated as $\mathbf{W}^\top \mathbf{x}_i$, denoted as $\tilde{\mathbf{x}}_i$, a vector of the same length of \mathbf{z}_j .

The tensor product above, also known as a bilinear model, can be regarded as a special case in the Tucker family [5], which have been extensively studied in literature and applications. For example, Tenenbaum and Freeman [18] developed a bilinear model for separating "style" and "content" in images, and recently Chu and Ghahramani [3] derived a probabilistic framework of the Tucker family for modeling structural dependency from partially observed high-dimensional array data.

The tensor indicator is closely related to the traditional HB models as in eq(1) and eq(2). Respondent heterogeneity is assumed either to be randomly distributed or to be constrained by attributes measured at individual level.

4.2 Logistic Regression

Conventionally the tensor indicator is related to an observed binary event by a logistic function. In our particular application, we found three additions in need:

- User-specific bias: Users' activity levels are quite different. Some are active clickers, while some might be casual users. We introduce a bias term for each user, denoted as μ_i for user i ;
- Article-specific bias: Articles have different popularity. We have a bias term for each article as well, denoted as γ_j for article j ;

² $\mathcal{N}(\tau, \varsigma^2)$ denotes a Gaussian distribution with mean τ and variance ς^2 .

- Global offset: Since the number of click events is much smaller than that of view events in our observations, the classification problem is heavily imbalanced. Thus we introduce a global offset ι to take this situation into account.

Based on the considerations above, the logistic function is defined as follows,

$$p(r_{ij}|\hat{s}_{ij}) = \frac{1}{1 + \exp(-r_{ij}\hat{s}_{ij} + \iota)}$$

where $\hat{s}_{ij} = s_{ij} - \mu_i - \gamma_j$ and s_{ij} is defined as in eq(3).

Together with a standard Gaussian prior on the coefficients $\{w_{ab}\}$, i.e. $w_{ab} \sim \mathcal{N}(0, c)$,³ the maximum a posteriori estimate of $\{w_{ab}\}$ is obtained by solving the following optimization problem,

$$\min_{\mathbf{W}, \mu, \gamma} \frac{1}{2c} \sum_{a=1}^D \sum_{b=1}^C w_{ab}^2 - \sum_{ij \in \mathcal{O}} \log p(r_{ij}|\hat{s}_{ij}).$$

where μ denotes $\{\mu_i\}$ and γ denotes $\{\gamma_j\}$. We employ a gradient descent package to find the optimal solution. The gradient with respect to w_{ab} is given by

$$\frac{1}{c} w_{ab} - \sum_{ij \in \mathcal{O}} r_{ij} x_{i,b} z_{j,a} (1 - p(r_{ij}|\hat{s}_{ij})),$$

and the gradient with respect to the bias terms can be derived similarly. The model parameter c is determined by cross validation.

4.3 Clustering

With the optimal coefficients \mathbf{W} in hand, we compute the partworths for each training user by $\tilde{\mathbf{x}}_i = \mathbf{W}^\top \mathbf{x}_i$. The vector $\tilde{\mathbf{x}}_i$ represents the user's preferences on article attributes.

In the partworth space spanned by $\{\tilde{\mathbf{x}}_i\}$, we further apply a clustering technique, e.g. K-means [13], to classify training users having similar preferences into segments. The number of clusters can be determined by validation in offline analysis.

For an existing or new user, we can predict her partworths by $\tilde{\mathbf{x}}_t = \mathbf{W}^\top \mathbf{x}_t$, where \mathbf{x}_t is the vector of user features. Then her segment membership can be determined by the shortest distance between the partworth vector and the centroids of clusters, i.e.

$$\arg \min_k \|\tilde{\mathbf{x}}_t - \mathbf{o}_k\| \quad (4)$$

where $\{\mathbf{o}_k\}$ denote the centroids obtained in clustering.

5. EXPERIMENTS

We collected events from a random bucket in July, 2008 for training and validation. In the random bucket, articles are randomly selected from a content pool to serve users. An event records a user's action on the article at the F1 position, which is either "view" or "story click" encoded as -1 and +1 respectively. We also collected events from a random bucket in September 2008 for test.

Note that a user may view or click on the same article multiple times but at different time stamps. In our training, repeated events were treated as a single event. The distinct events were indexed in triplet format, i.e. (user_id, article_id, click/view).

³Appropriate priors can be specified to bias terms too.

We split the July random bucket data by a time stamp threshold for training and validation. There are 37.806 million click and view events generated by 4.635 million users before the time stamp threshold, and 0.604 million click events happened after the time stamp for validation. In the September test data, there are about 1.784 million click events.

The features of users and items were selected by "support". The "support" of a feature means the number of samples having the feature. We only selected the features of high support above a prefixed threshold, e.g. 5% of the population. Each user is represented by a vector of 1192 categorical features, which include:

- Demographic information: gender (2 classes) and age discretized into 10 classes;
- Geographical features: 191 locations of countries and US States;
- Behavioral categories: 989 binary categories that summarize consumption behavior of a user within Yahoo! properties.

Each article is profiled by a vector of 81 static features. The 81 static features include:

- URL categories: 43 classes inferred from the URL of the article resource;
- Editor categories: 38 topics tagged by human editors to summarize the article content.

Categorical features are encoded as binary vectors with non-zero indicators. For example, "gender" is translated into two binary features, i.e., "male" is encoded as [0, 1], "female" is encoded as [1, 0] and "unknown" is [0, 0]. As the number of non-zero entries in a binary feature vector varies, we further normalized each vector into unit length, i.e., non-zero entries in the normalized vector are replaced by $1/\sqrt{k}$, where k is the number of non-zero entries. For article features, we normalized URL and Editor categories together. For user features, we normalized behavioral categories and the remaining features (age, gender and location) separately, due to the variable length of behavioral categories per user. Following conventional treatment, we also augmented each feature vector with a constant attribute 1. Each content item is finally represented by a feature vector of 82 entries, while each user is represented by a feature vector of 1193 entries.

5.1 Offline Analysis

For each user in test, we computed her membership first as in eq(4), and sorted all available articles in descending order according to their CTR in the test user's segment at the time stamp of the event. On click events, we measured the rank position of the article being clicked by the user. The performance metric we used in offline analysis is the number of clicks in top four rank positions. A good predictive model should have more clicks on top-ranked positions. We computed the click portion at each of the top four rank positions in predictive ranking, i.e. $\frac{\# \text{ of clicks at a position}}{\# \text{ of all clicks}}$, and the metric "lift" over the baseline model, which is defined as $\frac{\text{click portion} - \text{baseline click portion}}{\text{baseline click portion}}$.

We trained our logistic regression models on the training data with different values of the trade-off parameter c where

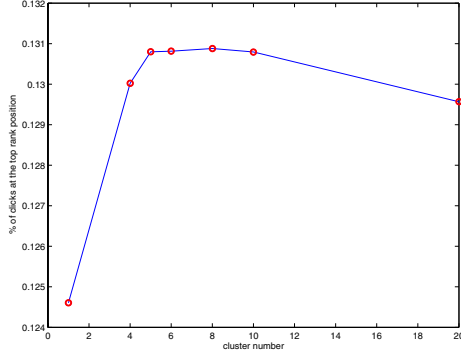


Figure 2: Click portion at the top rank position with different cluster number on the July validation data.

$c \in \{0.01, 0.1, 1, 10, 100\}$, and examined their performance (click portion at the top position) on the validation data set. We found that $c = 1$ gives the best validation performance.

Using the model with $c = 1$, we run K-means clustering [13] on the partworth vectors of training users computed as in eq(4) to group users with similar preferences into clusters. We varied the number of clusters from 1 to 20, and presented the corresponding results of the click portion at the top rank position in Figure 2. Note that the CTR estimation within segments suffers from low-traffic issues when the number of segments is large. We observed the best validation performance at 8 clusters, but the difference compared with that at 5 clusters is not statistically significant. Thus we selected 5 clusters in our application.

To verify the stability of the clusters we found in the July data, we further tested on the random bucket data collected in September 2008. The EMP approach, described as in Section 2.1, was utilized as the baseline model. We also implemented two demographic segmentations for comparison purpose:

- Gender: 3 clusters defined by users’ gender, {“male”, “female”, “unknown”};
- AgeGender: 11 clusters defined as {“<17, male”, “17~24, male”, “25~34, male”, “35~49, male”, “>49, male”, “<17, female”, “17~24, female”, “25~34, female”, “35~49, female”, “>49, female”, “unknown”}.

We estimated the article CTR within segments by the same technique [1] used in the EMP approach. A user will be served with the most popular article in the segment which she belongs to.

We computed the lifts over the EMP baseline approach and presented the results of the top 4 rank positions in Figure 3. All segmentation approaches outperform the baseline model, the EMP unsegmented approach. AgeGender segmentation having 11 clusters works better than Gender segmentation with 3 clusters in our study. Tensor segmentation with 5 clusters consistently gives more lift than Gender and AgeGender at all the top 4 positions.

5.1.1 Segment Analysis

We collected some characteristics in the 5 segments we discovered. On the September data, we identified cluster

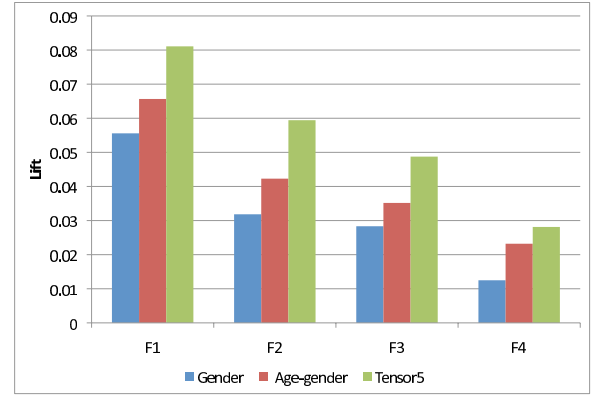


Figure 3: A comparison on lift at top 4 rank positions in the offline analysis on the September data. The baseline model is the EMP unsegmented approach.

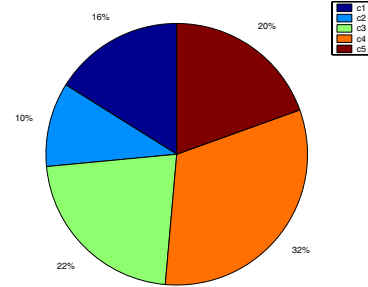


Figure 4: Population distribution in the 5 segments.

membership for all users, and plotted the population distribution in the 5 segments as a pie chart in Figure 4. The largest cluster takes 32% of users, while the smallest cluster contains 10% of users. We further studied the user composition of the 5 clusters with popular demographic categories, and presented the results in Figure 5 as a Hinton graph. We found that

- Cluster c1 is of mostly female users under age 34;
- Cluster c2 is of mostly male users under age 44;
- Cluster c3 is for female users above age 30;
- Cluster c4 is of mainly male users above age 35;
- Cluster c5 is predominantly non-U.S. users.

We also observed that c1 and c2 contains a small portion of users above age 55, and c3 has some young female users as well. Here, cluster membership is not solely determined by demographic information, though the demographic information gives a very strong signal. It is users’ behavior (click pattern) that reveals users’ interest on article topics.

We utilized the centroid of each cluster as a representative to illustrate users’ preferences on article topics. The centroids in the space of article topics are presented in Figure 6 as a heatmap graph. The gray level indicates users’ preference, from like (white) to dislike (black). We found the following favorite and unattractive topics by comparing the representatives’ scores across segments:

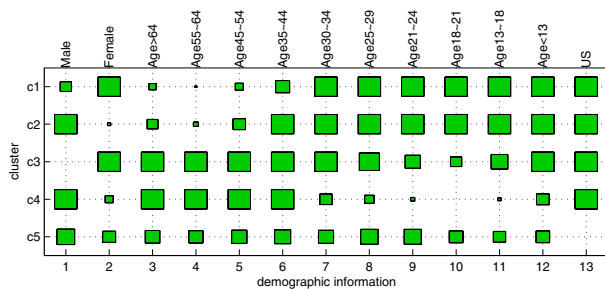


Figure 5: Distributions of users with a specific demographic category in the 5 clusters. In the Hinton graph, each square’s area represents user percentage in a cluster. Each column represents users having a particular demographic category, while each row is of a cluster.

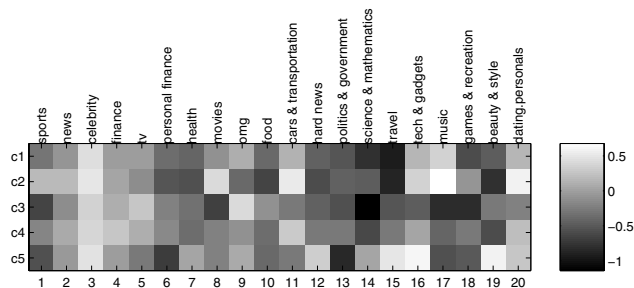


Figure 6: Users’ preferences on selected article topics in the 5 clusters. Each square’s gray level indicates the preference of a segment on the corresponding article topic, from white (like) to black (dislike).

- Users in c1 don’t have strong opinion on article topics, while relatively like Music more;
- Users in c2 greatly like Sports, Movies, Cars & Transportation, Tech & Gadgets and Dating/Personals, while dislike TV and Food;
- Users in c3 like TV, OMG,⁴ Food, while dislike Sports, Movies, Science & Mathematics, Games & Recreation etc.;
- Users in c4 are active readers who like Finance, Cars & Transportation and Politics & Government most;
- Users in c5 like Travel, Hard News and Beauty & Style, while are not interested in Personal Finance and Politics & Government.

Some topics are preferable for any user, such as Celebrity, but the discrepancy on interests is significant on most of topics. The discrepancy between clusters can be exploited by editors in content management to enhance user engagement.

5.1.2 Editorial Assistant

One interesting finding in our conjoint analysis is that visiting patterns of some segments are quite different from the

⁴<http://omg.yahoo.com/>, a web site of celebrity gossip, news, photos, etc.

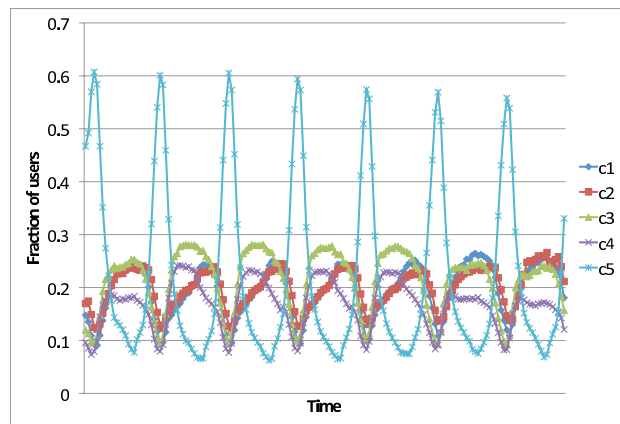


Figure 7: Fraction of views in each user segment over a week. The first day is a US holiday and the sixth and seventh day are Saturday and Sunday.

Table 1: Bucket test results of three segmentation methods, Gender, AgeGender and Tensor-5. All buckets have served the almost same amount of page views.

Segmentation	Lift on Story CTR
Gender	1.49
AgeGender	2.45
Tensor-5	3.24

others, as shown in Figure 7. Yahoo! Front Page is visited by more older users (c3 and c4) in the morning while by more younger users (c1 and c2) in the late afternoon. Most users around the midnight are international users (c5). Portion of traffic from older male users significantly decreases during weekend/holiday while traffic from other segments remains almost the same level for the entire week. This finding suggests some tips for content management, such as programming more articles related with News, Politics and Finance in the morning of weekdays, more entertainment articles of Sports and Music in the late afternoon, and more articles relevant to international users around midnight.

We can also monitor user activities within segments and remind editors to target underperformed segments when the CTR within the segments runs below their average level.

5.2 Online Bucket Test

To validate the tensor segmentation we proposed, we launched a bucket test in December 2008. Three segmentation methods, Gender, AgeGender and Tensor-5, were implemented in our production. From 8:00am 12 December to 0:00am 15 December, each of the three schemes and the control (EMP) bucket served about several million page views respectively. The numbers of page views in these four buckets are almost the same in the three days. We computed the story CTR and reported the corresponding lifts over the EMP control bucket for the three segmentation schemes in Table 1. The tensor segmentation with 5 clusters yields the most lift in the bucket test. We also observed the AgeGender segmentation outperforms the Gender segmentation. The observations in the online bucket test are consistent with our results in of-

fine analysis. Although the bucket test only lasted about 3 days around a weekend, the CTR lift we observed in significant amount of traffic gives another strong empirical evidence.

6. CONCLUSIONS

In this study, we executed conjoint analysis on a large-scale click through stream of Yahoo! Front Page Today Module. We validated the segments discovered in conjoint analysis by conducting offline and online tests. We analyzed characteristics of users in segments and also found different visiting patterns of segments. The insight on user intention at segment level we found in this study could be exploited to enhance user engagement on the Today Module by assisting editors on article content management. In this study, a user can belong to only one segment. We would like to exploit other clustering techniques, such as Gaussian mixture models, that allow for multiple membership, and then a user's preference might be determined by a weighted sum of several segmental preferences. We plan to pursue this direction in the future.

7. ACKNOWLEDGMENTS

We thank Raghu Ramakrishnan, Scott Roy, Deepak Agarwal, Bee-Chung Chen, Pradheep Elango, and Ajoy Sojan for many discussions and helps on data collection.

8. REFERENCES

- [1] D. Agarwal, B. Chen, P. Elango, N. Motgi, S. Park, R. Ramakrishnan, S. Roy, and J. Zachariah. Online models for content optimization. In *Advances in Neural Information Processing Systems 21*, 2009.
- [2] O. Chapelle and Z. Harchaoui. A machine learning approach to conjoint analysis. In *Advances in Neural Information Processing Systems 17*. MIT Press, 2005.
- [3] W. Chu and Z. Ghahramani. Probabilistic models for incomplete multi-dimensional arrays. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, 2009.
- [4] W. Chu and S.-T. Park. Personalized recommendation on dynamic content using predictive bilinear models. In *Proc. of the 18th International World Wide Web Conference*, 2009.
- [5] R. Coppi and S. Bolasco, editors. *Multiway data analysis*. North-Holland Publishing Co., Amsterdam, The Netherlands, The Netherlands, 1989.
- [6] W. S. DeSardo, M. Wedel, M. Vriens, and V. Ramaswamy. Latent class metric conjoint analysis. *Marketing Letters*, 3(3):273–288, 1992.
- [7] T. Evgeniou, C. Boussios, and G. Zacharia. Generalized robust conjoint estimation. *Marketing Science*, 24(3):415–429, 2005.
- [8] S. Gauch, M. Speratta, A. Chandranouli, and A. Micarelli. User profiles for personalized information access. In P. Brusilovsky, A. Kobsa, and W. Nejdl, editors, *The Adaptive Web — Methods and Strategies of Web Personalization*. Springer Berlin / Heidelberg, 2007.
- [9] P. E. Green and V. R. Rao. Conjoint measurement for quantifying judgmental data. *Journal of Marketing Research*, 8:355–363, 1971.
- [10] J. Huber and K. Train. On the similarity of classical and Bayesian estimates of individual mean partworths. *Marketing Letters*, 12(3):259–269, 2001.
- [11] J. Huber and K. Zwerina. On the importance of utility balance in efficient designs. *Journal of Marketing Research*, 33:307–317, 1996.
- [12] T. Jiang and A. Tuzhilin. Segmenting customers from population to individuals: does 1-to-1 keep your customers forever? *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1297–1311, 2006.
- [13] T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu. A local search approximation algorithm for k-means clustering. *Computational Geometry: Theory and Applications*, 28:89–112, 2004.
- [14] W. F. Kuhfeld, R. D. Tobias, and M. Garratt. Efficient experimental designs with marketing research applications. *Journal of Marketing Research*, 31:545–557, 1994.
- [15] P. J. Lenk, W. S. DeSardo, P. E. Green, and M. R. Young. Hierarchical bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science*, 15(2):173–191, 1996.
- [16] V. R. Rao. Developments in conjoint analysis. Technical report, Cornell University, July 2007.
- [17] Z. Sandor and M. Wedel. Designing conjoint choice experiments using managers' beliefs. *Journal of Marketing Research*, 38:455–475, 2001.
- [18] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12:1247–1283, 2000.
- [19] O. Toubia, J. R. Hauser, and D. I. Simester. Polyhedral methods for adaptive conjoint analysis. *Journal of Marketing Research*, 42:116–131, 2003.
- [20] M. Vriens, M. Wedel, and T. Wilms. Metric conjoint segmentation methods: A Monte Carlo comparison. *Journal of Marketing Research*, 33:73–85, 1996.
- [21] Y. Wind. Issue and advances in segmentation research. *Journal of Marketing Research*, 15:317–337, 1978.