Business Analytics for Non-profit Marketing and Online Advertising

by

Wei Chang

Bachelor of Science, Tsinghua University, 2003

Master of Science, University of Arizona, 2006

Submitted to the Graduate Faculty of

Katz Graduate School of Business in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

Katz Graduate School of Business

This dissertation was presented

by

Wei Chang

It was defended on

December 3rd , 2012

and approved by

R. Venkatesh, Professor, Katz Graduate School of Business

G.G. Hegde, Associate Professor, Katz Graduate School of Business

Jayant Rajgopal, Associate Professor, Swanson School of Engineering

Alan Montgomery, Associate Professor, Tepper School of Business

Dissertation Advisor:

Jennifer Shang, Professor, Katz Graduate School of Business

Business Analytics for Non-profit Marketing and Online Advertising

Wei Chang, PhD

University of Pittsburgh, 2013

Abstract

Business analytics is facing formidable challenges in the Internet era. Data collected from business website often contain hundreds of millions of records; the goal of analysis frequently involves predicting rare events; and substantial noise in the form of errors or unstructured text cannot be interpreted automatically. It is thus necessary to identify pertinent techniques or new method to tackle these difficulties. Learning–to-rank, an emerging approach in information retrieval research has attracted our attention for its superiority in handling noisy data with rare events. In this dissertation, we introduce this technique to the marketing science community, apply it to predict customers' responses to donation solicitations by the American Red Cross, and show that it outperforms traditional regression methods. We adapt the original learning-to-rank algorithm to better serve the needs of business applications relevant to such solicitations. The proposed algorithm is effective and efficient is predicting potential donors. Namely, through the adapted learning-to-rank algorithm, we are able to identify the most important 20% of potential donors, who would provide 80% of the actual donations.

The latter half of the dissertation is dedicated to the application of business analytics to online advertising. The goal is to model visitors' click-through probability on advertising video clips at a hedonic video website. We build a hierarchical linear model with latent variables and show its superiority in comparison to two other benchmark models. This research helps online business managers derive insights into the site visitors' characteristics that affect their click-through propensity, and recommends managerial actions to increase advertising effectiveness.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I am deeply grateful to my advisor, Dr. Jennifer Shang, for her inspiring guidance on my research in the last five years, for her mentorship on how to become a successful professional, and for her encouragement and patience when my research did not go well.

I want to thank all committee members for their time and attention on my dissertation. Dr. Venkatesh's course on quantitative marketing methods, Dr. Montgomery' research on clickstream data, Dr. Rajgopal's course on optimization and the GSA work with Dr. Hegde, all help to lead to the accomplishment of my dissertation.

In addition, the office of doctoral program provided strong support. Carrie Woods ensured me to meet all the requirements for Ph.D. degree. Dr. Galletta offered the final review on my dissertation and provided many detailed suggestions.

Most importantly, I want to express my deepest appreciation to my parents and my wife. After my son came into the world, they sacrificed their own time to help me take care of my son, so that I can focus on my dissertation. When I am frustrated with my research, their endless supports and undoubted faith on me kept me confident and motivated.

I have to say, I cannot receive the Ph.D. degree without the help from my family members, Katz faculty and staff members and colleagues in the Ph.D. program.

1.0 INTRODUCTION

1.1 CURRENT CHALLENGES FACED BY BUSINESS ANALYTICS

Many quantitative methods have been used in the marketing literature to predict the outcome of consumers' decisions. Choice models prevailed in the 1990s to predict consumers' decisions across product categories. In the early 20th century, as computational methods advanced, more complex statistical models were introduced into the marketing science community. In particular, the success of Markov Chain Monte Carlo (MCMC) estimation frees up the computational constraints for researchers to use complex econometrics models. Rossi and Allenby (2003) provide an excellent review of the application of Bayesian statistics in marketing research. In the second half of the 20th century, researchers advocated the use of powerful machine learning techniques. For example, Support Vector Machine (Cui and Curry (2005)), Neural Network (Kim, Street et al. (2005)) and Bayesian Network (Cui, Wong et al. (2006)), among others, are used to tackle the problem of identifying valuable customers. The general framework of this type of problems is described as follows: Given the training data set that contains the feature vector X and known outcome Y, we build a model f to describe the relationship between X and Y, $y = f(x)$. The model f can be a parametric model that reflects domain knowledge, as in econometric models; a semi-parametric model, as in Support Vector Machine; or even a

completely black box approach, as in Neural Network. All such approaches have reported successful prediction results at top notch journals.

As the recent development of information technology makes data easily available for business operations, business analytics has become the central theme at major venues of management science- and marketing science-related conferences. However, as the era of "Big Data" looms, traditional statistical or machine learning-based approaches encounter great challenges when tackling data sets of significant size.

One problem associated with the large data sets is tremendous noise. Many databases mechanically record all of the traces that customers leave during the interaction with the company without the need for any manual intervention or subjective discretion. These loosely connected data do not present strong patterns or relationship among them and make it hard to analyze using standard analytical approaches. This type of data calls for algorithms with the ability to handle noise better than traditional regression analysis.

Another problem is that the instances in which we are interested in predicting usually constitute a very small portion of the entire database. In Customer Relationship Management (CRM), the target is to identify a small percentage of customers (i.e., valuable customers) who can bring significant value to the companies. Moe and Fader (2004) reported that the typical conversion rate rarely exceeds 5% and is hard to predict. Chatterjee, Hoffman et al. (2003) claim that click-throughs on banner ads are extremely rare events, and logistic regression severely underestimates the probability of click-through rates. This is because many statistical methods used to predict

binary outcome, such as logistic regression, obtain parameter estimation largely by maximizing the likelihood function. However, statisticians have demonstrated that maximum likelihood can produce poor coefficient estimation in terms of *p*-value and confidence interval, when the percentage of positive cases in the data set is considerably low or large (King and Ryan (2002)). Some correction methods, including different estimation procedure and data resampling, have been proposed by statisticians. In this dissertation, we provide another alternative: "learning-to-rank" technique, which is not only powerful in predicting outcomes with imbalanced distribution and less prone to the noise, but also inherits some other merits from the recent development in machine learning disciplines. For example, ensemble/committee methods can model complex relationships between the predictors and outcome much more accurately than one single complicated model. The specific "learning-to-rank" algorithm implemented in our experiment is essentially an ensemble approach.

1.2 THE LEARNING-TO-RANK TECHINIQUES

A challenge faced by the marketing and management science communities exhibits vast similarities with the challenges faced by the information retrieval fields. The typical problem for information retrieval research is to find the most relevant documents given a user query. Many times researchers concern themselves only with returning the most relevant documents to a specific query. The size of the document corpus is enormous, and only a limited number of documents displayed in the first several pages are meaningful for a query user.

To overcome the deficiency of traditional statistical prediction approaches in predicting rare events in a large data set with compelling noise, the "learning-to-rank" technique has become an emerging research field in information retrieval in recent years. Simply put, a ranking algorithm assigns a score to each candidate document, and documents will feed into the result list in the order of their ranking scores. This was traditionally done by building heuristic scores for candidate documents. For example, Brin and Page (1998) introduced TF/IDF (term frequency/inverse document frequency) and cosine similarity scores based on document attributes. Cohen, Schapire et al. (1998) provided a short overview of traditional ranking strategies without involving the learning process. Within the past six years, many researchers have shown that statistic-based learning algorithms can significantly improve the performance of query results. The process of leveraging statistic learning in document ranking is depicted in Figure 1-1. A database is built to store the indices of a large corpus of documents. Given a user query, a ranking model outputs the top *k* retrievals and feeds the corresponding documents to the results page. The system will track which recommended documents have been clicked-though and use them as benchmarks to train and compare against the ranking scores.

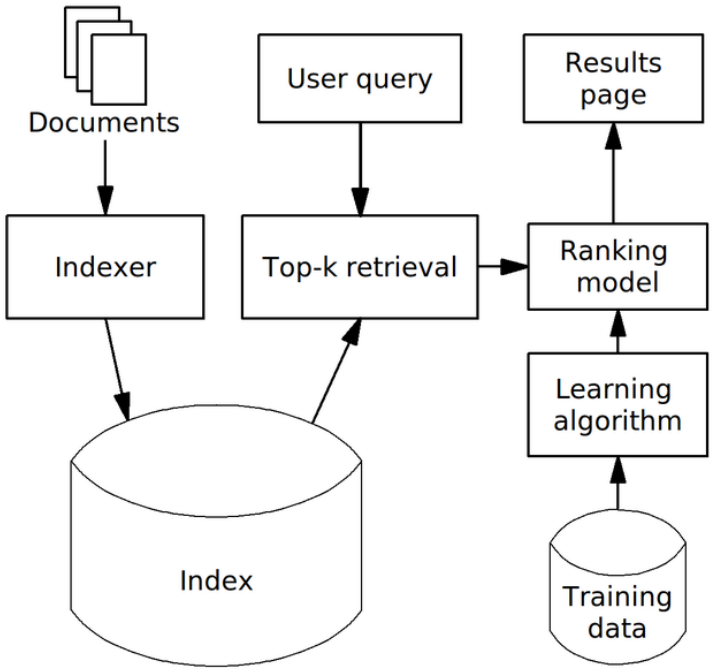


Figure 1-1 The process of leveraging statistic learning in document ranking

Learning-to-rank has achieved great success not only in information retrieval, but also in the fields of computational biology (Duh and Kirchhoff (2008)) and proteomics (Henneges, Hinselmann et al. (2009)). It has been shown to be especially more effective when the data have a skewed distribution in both positive and negative cases. For instance, when the data are dominated by negative cases, the prediction accuracy using traditional regression will be greatly compromised by the negative portion of the data. Interestingly, ranking algorithms can even improve the performance of classic logistic regression when they are combined in use. Sculley (2010) added ranking loss function (defined on pairs of documents) to regression loss function (defined on single documents), and used stochastic gradient descent to optimize the combined loss function. He claims that the combined approach with ranking components can improve the regression prediction in the case of rare events or skewed distribution.

1.3 DISSERTATION OUTLINE AND CONTRIBUTIONS

To the best of our knowledge, learning-to-rank techniques have not been used in the area of direct marketing. Given their superior performance over traditional statistical methods, we feel impelled to introduce learning-to-rank techniques to the marketing science community so as to help tackle the above-mentioned big challenges in business analytics.

Not only will we introduce the learning-to-rank techniques to the marketing community by providing an extensive review of these methods, but we also aim to modify and enhance them in order to better serve our needs in specific business analytics applications. Despite the many similarities between information retrieval and direct marketing problems, they also have many distinctions that are worth noting.

Firstly, although in low percentage, the absolute size of high-value customers in direct marketing applications is usually much larger than in information retrieval. For example, a company typically contacts over 5,000 customers in a marketing campaign while an information retrieval engine only concerns the top 10 results displayed on the first page.

Secondly, in information retrieval, the order of documents in the returned set matters. However, in a direct marketing problem, it does not. For instance, the American Red Cross contacts about 5,000 customers every month. Whether a customer is the most valuable or the 100^{th} most valuable makes no significant difference as long as he or she shows up in the top 5,000 recommendation list.

Thirdly, the importance of the ranking accuracy decreases dramatically in information retrieval problems, which is not the case in a direct marketing problem.

In Chapter 2, we give an extensive review of existing learning-to-rank approaches developed in the information retrieval field and aim to provide quality resources for marketing researchers interested in such techniques. We emphasize the advantages of a learning-to-rank algorithm against traditional statistical prediction methods. We also discuss ways to adapt the learning-to-rank algorithm to better serve the needs in the applications of direct marketing, or, more broadly, in customer relationship management.

Chapter 3 presents the experimental results of applying adapted learning-to-rank techniques on data from a non-profit organization, the American Red Cross, and shows that it outperforms traditional regression methods in predicting the donation tendencies of its customers.

In chapter 4, we present results from a different business analytics application: We collect click stream data from a hedonic video website and build models to predict visitors' click-through probability on advertising video clips. We build a hierarchical linear model with latent variables, and show its superiority in comparison to two other benchmark models. Using results from a hierarchical linear model, we apply a ranking algorithm to prioritize site visitors. This research helps web site managers gain insights into factors that affect visitors' click-through probability and suggests managerial actions to increase click-through rates.

Chapter 5 summarizes the key results of two applications in business analytics and suggests directions for future work.

2.0 PROPOSAL OF LEARNING-TO-RANK ALGORITHMS FOR DIRECT MARKETING APPLICATIONS

In this chapter, we first give an extensive review of the learning-to-rank techniques, aiming to provide a variety of scholarly resources for interested readers. Then we discuss how we select and adapt certain learning-to-rank techniques to make them better suited for direct marketing applications.

2.1 REVIEW OF LEARNING-TO-RANK TECHNIQUES

Learning-to-rank algorithms have strong roots in data mining/machine learning. Liu (2011) gave an excellent review of and tutorial on the recent developments in the ranking algorithms in information retrieval research. Here, we review several popular ranking algorithms following Liu's categorization: pointwise, pairwise and listwise.

2.1.1 Pointwise approaches

Pointwise approaches find the mapping function between a candidate document and its corresponding score $f(d_i) \rightarrow S_i$. This type of approach relies on knowledge of the "ground truth labels/scores" to work. The model can be trained only if the ground truth scores of relevance

(such as "exact match," "most relevant," etc.) for the candidate documents are known. Such information may be based on subjective opinions acquired from domain experts.

Traditional regression or classification methods belong to pointwise approaches. One can treat customers' value or document relevance as either continuous variables or multiple-ordered categories and run the traditional regression, ordinal regression or multi-class classification to obtain the predicted rank. Assume we have k ordered categories $1,2,3\cdots k$ to indicate the strength of document relevance, and we predict the most probable category to which each document belongs. We briefly describe an ordinal regression algorithm, Prank, and a multi-class classification algorithm, McRank, to review this type of ranking approach.

PRank (Crammer and Singer (2001)) is a famous ordinal linear regression algorithm. It defines k thresholds $b_1 \leq b_2 \cdots \leq b_k = +\infty$ for each category. For an observation x_i, if $b_m \leq wx_i < b_{m+1}$, then the fitted value is $\widehat{y_i} = m + 1$. PRank is an online algorithm, which means it updates its parameters w and b whenever a new observation x_n becomes available. The algorithm modifies parameters only when it predicts the category of x_n mistakenly. First, it defines auxiliary variables:

$$y_r = \begin{cases} 1 & if\ wx_n \geq b_r \\ -1 & if\ wx_n < b_r \end{cases} \quad r = 1,2,\cdots k-1$$

If it predicts the category of x_n correctly, then $y_r(wx_n - b_r) \geq 0$ for all r. Otherwise, for those y_r that $y_r(wx_n - b_r) < 0$, the following update is done to correct the prediction mistake:

$$b_r \rightarrow b_r - y_r$$

$$w_r \rightarrow w + (\Sigma y_r)\, x_n$$

McRank Li, Burges et al. (2007) adopts a multi-class classification technique to tackle the ranking problem and claims that classification-based ranking algorithms outperform regression-based ranking algorithms. They use a boosting tree algorithm to learn the class probability $p_{i,k} = \Pr(y_i = k)$, and then the scoring function

$$s_i = \sum_{t=1}^{k} p_{i,t} T(t)$$

is used to output the ranked list, where $T(t)$ is a monotonic function of category t.

The relative orders between the objects are not explicitly modeled in the pointwise learning processes. Since the ranking problem is more dedicated to predicting the relative orders between objects rather than their absolute relevance, pairwise and listwise approaches are more suited in such applications.

2.1.2 Pairwise approaches

Pairwise approaches do not emphasize the absolute ranking scores for individual documents. Instead, they focus on the relative preference between any pair of documents. Pairwise ranking algorithms deal with preference data and seek the ranking scores which minimize the occurrences of contradicting pairs. Preference data consists of entries indicating the preference between two objects. For instance, in information retrieval, a document d_i is preferred to d_j if document i receives higher click-through rates. In case of donation tendency, one donor d_i is preferred to d_j if donor i is more likely to donate. In other occasions, the preference relationship may be determined by opinions from domain experts. A contradicting pair contains two objects whose scores contradict their ground truth preferences. For instance, two documents form a

contradicting pair if the document with higher ranking score is actually less relevant according to the click-through rates from a query return.

The study of ranking on paired preference can be dated back to late last century. Cohen, Schapire et al. (1998) proposed a two-stage approach. At the first stage, a preference function y_{x_u,x_v} is built on a pair of objects x_u, x_v. Preference functions can be learned by some conventional classification learning algorithm. The value of y_{x_u,x_v} is between 0 and 1, with 1 indicating x_u as more preferable to x_v while 0 indicates x_v is more preferable to x_u. However, converting the preference function to a ranked list is NP-hard. The researchers proposed a greedy algorithm at the second stage which tries to agree with the pairwise preference as much as possible and proved that this greedy algorithm has at least half the agreement of the optimal ranked order. More recent pairwise ranking approaches incorporating machine learning techniques are reviewed in subsequent sections.

2.1.2.1 Neural network based approaches

a) SoftNet

SoftNet Rigutini, Papini et al. (2011) uses a two-layer neural network to learn the preference function. The main idea behind SoftNet is the symmetry of the preference function—i.e., the probability $P(x_u \succ x_v)$ that u is preferable to v equals to the probability $P(x_v \prec x_u)$ that v is inferior to u. The input nodes are feature variables $x_{u,1}, \cdots, x_{u,t}, \cdots, x_{v,1}, \cdots, x_{v,t}, \cdots$, of two objects x_u, x_v. For each neuron h_i in the middle layer, a dual neuron h_i' exists with some weight-sharing schema to ensure the symmetry of the preference function. Specifically,

$$h_i(X_u, x_v) = \frac{\exp(\sum_t \theta_{u,t,i} x_{u,t} + \theta_{v,t,i} x_{v,t+b_i})}{1+\exp(\sum_t \theta_{u,t,i} x_{u,t} + \theta_{v,t,i} x_{v,t+b_i})} = \frac{\exp(\sum_t \theta_{u,t,i'} x_{u,t} + \theta_{v,t,i'} x_{v,t+b_{i'}})}{1+\exp(\sum_t \theta_{u,t,i'} x_{u,t} + \theta_{v,t,i'} x_{v,t+b_{i'}})} = h_{i'}(X_u, x_v)$$

with constraints

$$\theta_{u,t,i} = \theta_{v,t,i'};\ \theta_{u,t,i'} = \theta_{y,t,i}; b_i = b_{i'}$$

imposed to ensure that a neuron and its dual have the same value.

Similarly, at the top level,

$$P(x_u \succ x_v) = \frac{\exp(\sum_{i,i'} w_{i,\succ} h_i(X_u, x_v) + w_{i',\succ} h_{i'}(X_u, x_v) + b_\succ)}{1+\exp(\sum_{i,i'} w_{i,\succ} h_i(X_u, x_v) + w_{i',\succ} h_{i'}(X_u, x_v) + b_\succ)}$$

$$= \frac{\exp(\sum_{i,i'} w_{i,\prec} h_i(X_u, x_v) + w_{i',\prec} h_{i'}(X_u, x_v) + b_\prec)}{1+\exp(\sum_{i,i'} w_{i,\prec} h_i(X_u, x_v) + w_{i',\prec} h_{i'}(X_u, x_v) + b_\prec)} = P(x_v \prec x_u)$$

with constraints $w_{i,\prec} = w_{i',\succ};\ w_{i,\succ} = w_{i',\prec}; b_\succ = b_\prec$.

The learning process uses gradient descent to optimize the mean square loss

$$L = \left(y_{u,v} - P(x_u \succ x_v)\right)^2 + \left(y_{v,u} - P(x_v \prec x_u)\right)^2$$

After learning the pairwise preference, a common sorting algorithm can be deployed to output the ordered rank list. A typical modern ranking data set contains hundreds of thousands of observations. If we learn the preference function on all possible pairs, then the data size is in the magnitude of millions or even larger. To make the learning process more practical, the authors propose an incremental learning process. The algorithm starts with a random generated neuron network, and, at each iteration, the neuron network model will learn on the pairs misclassified in the previous iteration until there is no significant improvement. The improvement is defined by some evaluation measures which will be discussed in later sections.

b) RankNet

Another neural network based ranking algorithm is RankNet Burges, Shaked et al. (2005), which is used by the commercial search engine, Bing. RankNet differs from SoftNet for the loss

function and how to construct probability $P(x_u \succ x_v)$. RankNet uses cross-entropy as the loss function

$$L = -\overline{P_{uv}} log P_{uv} - (1 - \overline{P_{uv}})\log(1 - P_{uv})$$

where P_{uv} is short for $P(x_u \succ x_v)$ and P_{uv} is the target probability constructed according to the ground truth labels. Instead of a weight-shared dual-neuron structure in the middle layer, RankNet builds preference probability on a ranking score function $f(x_u)$:

$$P_{uv} = \frac{\exp(f(x_u) - f(x_v))}{1 + exp(f(x_u) - f(x_v))}$$

This structure guarantees the consistency when pairwise preference probabilities are converted to the ranking score. Thus, RankNet outputs the ordered rank list naturally based on $f(x_u)$.

There are two major problems associated with the cross-entropy loss function. First, it has non-zero minimum, which means under certain circumstances we will receive non-zero loss even after we have learned the pairwise preferences perfectly. Secondly, the loss function is not bounded. Some abnormal pairs may dominate the learning process and lead to poor performance. A fidelity loss function is then proposed in Frank's algorithm to overcome these problems (Tsai, Liu et al. (2007)).

$$L = 1 - \sqrt{\overline{P_{uv}} P_{uv}} - \sqrt{(1 - \overline{P_{uv}})(1 - P_{uv})}$$

The fidelity loss is between 0 and 1. However, it is not convex and therefore is difficult to optimize. An iterative procedure similar to the boosting technique helps to estimate parameters in Frank's algorithm.

2.1.2.2 Boosting-based approaches

a) RankBoost

Boosting technique has emerged in the last decade as a very successful approach to build predictive models. In a variety of applications, boosting algorithms have shown better prediction accuracy. Boosting belongs to the category of ensemble methods. There is growing agreement in the machine-learning community that data sets in the real world are more or less heterogeneous. Some part of the data space may behave quite differently than other parts. Thus, it is very difficult to use only one synthetic model to describe the behavior of the entire data set, no matter how comprehensive the model may be. In contrast, researchers have found it performs well if a relative simple model is built for each sub-data space that behaves more or less homogeneously, and all these simple models are combined into one final model. Adaboost (Freund and Schapire (1997)) is one of the earliest boosting algorithms to implement the ensemble idea. Initially, the algorithm assigns equal weights to all the data points $D_0(i) = \frac{1}{n}$. At each iteration, a simple weak model $h_t(x)$ is learned on the weighted data, and the weights are updated after the learning. Training data that are being correctly classified will be assigned lower weights and misclassified data will be assigned higher weights. Specifically,

$$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

where Z_t is a normalization factor to ensure the sum of $D_t(i)$ is 1. α_t is the weight that is used to combine weak learners learned at each iteration to build the final model $H(x) = \sum_t \alpha_t h_t(x)$.

Rankboost (Freund, Iyer et al. (2003)) extends the idea of Adaboost to the ranking problems. The only difference is that the weights in Rankboost are defined on pairs of observations. It assumes

that there exists a ranking score function $f(x_i)$ and

$$D_{t+1}(x_i, x_j) = \frac{D_t(x_i, x_j)\exp(-\alpha_t(f(x_i) - f(x_j)))}{Z_t}$$

Many beautiful properties of Adaboost still hold in Rankboost. In addition, an efficient algorithm is available for a bipartite group problem. In a bipartite group problem, the task is to divide the data set into two disjoint groups with ranks in one group consistently higher than those in the other group, and does not impose conditions on the rank orders within each group. Our problem—to identify a limited number of potential donors—is exactly a bipartite group problem. Whereas the positions of the returned documents matter in the document ranking problems, we are not concerned about the order of donors on our top list since the American Red Cross will contact about 5,000 donors every month anyway.

b) GBRank

Friedman (2001) proposed a functional gradient descent framework to search an optimal function mapping $h: x \to y$ with respect to certain loss function $L(h(x), y)$in a function space $h \in \mathcal{H}$. The process is similar to the gradient descent method to find the minimum or maximum value of a function defined on real value. The difference here is that the space we are searching in is not real value but a function space. This algorithm is also related to the idea of boosting, because at every iteration the function h_k will be corrected by its gradient, which is largely decided by the portion of data that h_k does not predict well. Zheng, Chen et al. (2007) applied the gradient descent framework to the pairwise ranking problem. The loss function is defined on pairs of objects $< x_i, y_i >$, with a high ranking object followed by a low ranking object.

$$L = \frac{1}{2}\sum_{i=1}^{N}(max\{0, h(y_i) - h(x_i)\})^2 \quad (1)$$

The corresponding gradients with respect to $h(x)$ are

$$\begin{aligned} &\max\{0, h(y_i) - h(x_i)\} && \text{for } h(x_i) \\ &-\max\{0, h(x_i) - h(y_i)\} && \text{for } h(y_i) \end{aligned} \tag{2}$$

If the predicted order between a pair of objects $< x_i, y_i >$ according to the current ranking scores contradict their ground truth order, the score function value $h(x_i)$ will be modified to $h(x_i) + [h(y_i) - h(x_i)] = h(y_i)$ and the learning function value $h(y_i)$ will be modified to $h(y_i) + [-\big(h(y_i) - h(x_i)\big)] = h(x_i)$. GBRank adds all the updated scores, $< x_i, h'(x_i) >$, into one training data set and runs regression to fit between x_i and $h'(x_i)$ to approximate the functional gradient. Suppose the regression output is $g(x)$, and then the ranking function will be updated as

$$h_k(x) = \frac{k h_{k-1}(x) + \eta g(x)}{k+1}$$

where η is the learning rate.

Most boosting approaches are non-parametric and impose no assumption on the scoring function $h(x_i)$

2.1.2.3 SVM based approach

Another successful extension from conventional machine learning algorithms to ranking algorithms is RankSVM (Herbrich, Graepel et al. (2000)). RankSVM assumes a linear ranking function $f(x) = w^T x$. Both RankSVM and SVM use hinge loss function and L_2 -norm regularization. Thus, RankSVM has the same objective function as SVM

$$\min \frac{1}{2}\big||w|\big| + \lambda \sum\nolimits_{u,v:y_{u,v}=1} \xi_{u,v}$$

except that the term $\xi_{u,v}$ is defined on pairs. Correspondingly, the constraints are slightly different from those in SVM.

$$w^T(x_u - x_v) \geq 1 - \xi_{u,v} \; if \; y_{u,v} = 1$$

Since RankSVM has such a strong connection to SVM, it inherits many merits from SVM such as not being prone to overfitting, easy-to-handle nonlinear function through kernel tricks, and performing well for high dimensional data.

2.1.3 Listwise Approach

Listwise approaches directly optimize the permutation of the recommendation list. There are two strategies to incorporate position information in the training process: one adapts the loss function while the other directly compares the ranking result and the ground truth permutation of the list. In the former case, the problem of using direct measurement as the loss function is that neither Average Precision nor NDCG evaluation measure (discussed in the next chapter) is a smooth function and therefore both are very hard to optimize. Several attempts have been made to tackle the challenge. One such attempt can use a continuous and differentiable function to approximate the directly measured loss function. SoftRank (Taylor, Guiver et al. (2008)), Approximate Rank Taylor, Guiver et al. (2008)), and SmoothRank (Chapelle and Wu (2010)) are some exemplary algorithms which seek mathematical approximation of NDCG so that traditional optimization approaches can be applied. Another option is to find an appropriate upper bound of the directly measured loss function and optimize the upper bound. For example, SVM^{map} Yue, Finley et al. (2007) adapts the constraints in the traditional SVM formulation and shows that the sum of slack variables in the adapted constraints is the upper bound of Average Precision. As in the conventional SVM, the sum of slack variables is part of objective function to optimize.

Alternatively, one can also choose to utilize algorithms, such as boosting or genetic algorithms (GA), to directly optimize non-smooth objectives. In the GA case, the difficulty is the complexity of working directly on the permutation, as it is impossible to evaluate all possible permutations for a large data set.

2.1.4 The gap between the training loss function and evaluation measures

Average precision (AP) and discounted cumulative gain (DCG) are two common measures to evaluate the performance of ranking results. Assume that we have a ranked list output π from some ranking algorithm and a ground truth list l. $\pi^{(k)}$ and $l^{(k)}$ are the set of top k objects from each list. The precision for the top k ranked objects is the percentage of predicted objects that indeed exist in the ground truth top k list.

$$\boldsymbol{Precision@k} = \frac{\sum_{i \in \pi^{(k)}} \boldsymbol{I}(\boldsymbol{\pi_i} \in \boldsymbol{l^{(k)}})}{\boldsymbol{k}} \tag{3}$$

$$\boldsymbol{AP@n} = \frac{\sum_k \boldsymbol{Precision@k}}{\boldsymbol{n}}$$

In a typical information retrieval problem, the accuracy on the top positions is more important than that at the bottom. DCG explicitly discounts the accuracy requirement for low position predictions. Let G_i be the relevancy score of a returned document on position i, then

$$\boldsymbol{DCG@n} = \sum_i^n \frac{\boldsymbol{G_i}}{\boldsymbol{log_2(i+1)}} \tag{4}$$

IDCG, or ideal DCG, is the DCG when all the objects are ordered correctly, namely conforming to their ground truth scores. DCG can be normalized by being divided by IDCG, and the value of NDCG is between 0 and 1.

$$\boldsymbol{NDCG} = \frac{\boldsymbol{DCG}}{\boldsymbol{IDCG}} \tag{5}$$

As we can see from the pairwise ranking approaches discussed above, the loss function optimized in the training model is different from the measures used to evaluate ranking models. The major issue is that the loss functions try to find a model to predict as many pairs with correct relative preference as possible. In other words, they assume it is more important to correctly predict whether x_u is preferable over x_v (or vice versa) than to ensure that an object that should be on the top list actually does so. Thus, many improved algorithms tend to incorporate the position information of rank list into the loss the function. For example, Burges, Ragno et al. (2006) bring the *lambda* function into the RankNet loss function. The idea is that the absolute values of gradients of the cost function at the top position will be greater than the gradients at the lower position so that a document with higher rank is harder to be assigned a high relevance score in the next iteration. A sufficient condition for the cost function that holds this property is given in the paper. Many gradient descent-based approaches can adapt their cost function using the *lambda* function to include position information. At each iteration, incorrectly predicted objects at high positions of the rank list will have larger corrections (gradient) than those objects at lower positions. We implement this idea algorithmically instead of mathematically on top of a pairwise ranking algorithm to achieve better prediction performance. Detailed discussions follow in the next section.

2.2 MODEL SELECTION AND ADAPTION

Given the superior performance that learning-to-rank techniques hold over traditional statistical methods, we now introduce it to the marketing science community so as to help tackle the big challenges in business analytics discussed in Section 1.1. In this section, we first discuss the reasons why we choose a particular ranking method (i.e., GBRank) to implement. Then we describe how we tried to modify the prototype algorithm to achieve higher prediction accuracy near the top range and to better serve our needs in the specific business analytics application.

Among pointwise, pairwise and listwise ranking approaches, we prefer pairwise approaches for the following reasons. Pointwise approaches perform poorly in face of data with a lot of noise caused by a skewed distribution on the positive and negative cases. In contrast, pairwise and listwise approaches focus on relative orders between objects and therefore overcome the problems caused by skewed distributions. For the American Red Cross (ARC) data to be discussed in the next chapter, we prefer pairwise to listwise approaches due to the nature of the data itself; the listwise approach requires the training data to include the ground truth orders of donation tendencies for the entire list of candidate customers. However, this information is not available in our application. In contrast, pairwise approaches only require pairwise preference data that record the relative preference between pairs of objects. Such pairwise preference data can be handily constructed from the original ARC data by comparing customers' response to donation campaign: Customers who donated are preferable to those who did not. Moreover, even if we had the ground truth orders, we might still prefer pairwise to listwise approaches due to another advantage of pairwise approaches in handling large data. Generally speaking, listwise algorithms are more computationally expensive since they need to work on permutations of the

rank list. Our data set contains nearly one million donors, and the size of the data makes listwise algorithms computationally impractical. For these reasons, we choose the pairwise approach as the building block and tailor the algorithm to our specific application.

Among the many pairwise approaches reviewed in the previous section, we choose GBRank to others because of easy implementation and extendibility.

Recall that GBRank, as a boosting-based ranking algorithm, emerges as an iterative method to minimize the loss function $L(h(x), y)$ (see Eq. 1). The functional gradient of L with respect to h(x) has very simple forms (see Eq. 2), and the action of updating the scores according to the gradients can be simply implemented in several intuitive steps: create a list all incorrectly ordered pairs at the current iteration, exchange their current scores, do regression on the correction list, and update the ranking scores using the regression predictions.

The generality of the GBRank framework and its intuitive implementation leaves room for various modifications. For instance, one can choose to use a different form of the loss function, and, accordingly, the derived functional gradient used for score updating will also be different. One can also start by directly modifying the rules for updating scores at each iteration. The specified updating rules may or may not correspond to a loss function depending on whether the rules satisfy certain conditions.

We tried several ways to adapt the original GBRank algorithm by directly modifying the score-updating rules. One way is to incorporate positional information in the updating rules. The

original GBRank algorithm (Table 2-1) applies the same rule for correction, namely by adding or subtracting a fraction of the gradient $\pm[h_k(y_i) - h_k(x_i)]$, regardless of whether the score being updated takes a high or a low position in the current ranking. We modify this rule by letting the correction to be dependent of the customer's current ranking (see Table 2-2 for more details). Note that the adapted GBRank algorithm only modifies Step 3 of the original GBRank algorithm. Another modification that we have tried is to divide all customers into top-tier and lower-tier buckets based on their current ranking scores, and omit any contradicting pairs within the same bucket. In this manner, the relative orders within each bucket do not need to be adjusted due to "within-bucket errors." The size of the top-tier bucket is assigned as the preferred number of high-value customers. This modification is particularly designed due to the fact that in our ARC application (to be discussed with details in the next chapter), the relative orders of customers in the returned list do not play an essential role. We also tried other modifications which perform similarly as the above-mentioned modification.

Table 2-1 The original GBRank algorithm

Step 1.	Assign random scores $h_0(x)$ to all candidate customers.
Step 2.	Scan through all pairs in the preference data and find out their current scores $h_k(x)$. If any pairs have scores that contradict their ground truth orders, record these pairs and their current scores in a separate table called "contradicting pairs."
Step 3.	Exchange the two scores in each row of the "contradicting pairs" table to form a new table, titled "correction scores."
Step 4.	Run regression for the correction scores with all customers in the "contradicting pairs" table and output the fitted values of correction scores, named $g_k(x)$.
Step 5.	Update the current score $h_k(x)$ $$h_{k+1}(x) = \frac{kh_k(x) + \eta g_k(x)}{k+2}$$ and go to Step 2.

Table 2-2 The adapted GBRank algorithm

Step 3*.a.	Order all the customers according to their current scores $h_k(x)$, and assign their current ranks to a function $I_k(x)$. For instance, the customer with the highest current score has ranks I_k=1; the customer with the second-highest current score has rank I_k=2,…, etc.
Step 3*.b.	For each pair $<x_i, y_i>$ in the "contradicting pairs" table (suppose their current scores satisfy $h_k(y_i) > h_k(x_i)$), The customer with a currently lower score $h(x_i)$ will be reassigned a higher score $h_k(x_i) + [h_k(y_i) - h_k(x_i)] * log(I_k(y_i))$ The customer with a currently higher score $h(y_i)$ will be reassigned a lower score $h_k(y_i) - [h_k(y_i) - h_k(x_i)] \cdot \frac{1}{I_k(y_i)}$

3.0 EMPIRICAL ANALYSIS OF AMERICAN RED CROSS DATA

3.1 OVERVIEW OF THE AMERICAN RED CROSS (ARC) DATA SET

We use a data set from the American Red Cross to demonstrate how to leverage the learning-to-rank algorithms to identify the most valuable customers. In this section, we will take an overview of various aspects of this data set, and discuss the challenges faced by the American Red Cross in customer management.

The ARC database consists of over one million accounts who have made donations to or been contacted by American Red Cross. It records the date and amount of each donation. In addition, it keeps track of the contact history American Red Cross has made with its customers. The initiative of this database is to covert people who made donations in response to a rare disaster to regular donors. Russ Reid took over the project after 2009 and recorded the data more carefully and thoroughly than before. Thus, we use the data after 2009 for better quality.

The American Red Cross contacts its customers mostly through direct mailing to solicit donations. From 2009 to 2011, the American Red Cross conducted a total of 64 campaigns to its target customers, subsequently recording any payment received with the correlate campaign. Since the American Red Cross very rarely contacted a particular customer more than once in

each campaign, we can identify whether people responded positively to donation solicitation by matching the campaign and user identifier. As 92% of customers who decided to donate made their payments within 90 days after being contacted, we label a donation as a positive response to a contact if the customer makes the donation within 90 days.

Each donation campaign sends out solicitation mail to a specific type of targeted population. The American Red Cross has classified its customers into several categories, including "supporters" who have donated two or more times, "lapsed customers" who have donated before but have not done so in the last 18 months, and "pre-qualified lead" who donate in response to a disaster in an unsolicited program and may become a regular donor.

The typical size of the targeted customers in each campaign is around 5,000. The American Red Cross generally cannot contact more customers than this size due to cost concerns and budget control. However, there exist more than one million accounts in the database, among which 366,469 accounts made donations after 2009. How to choose 5,000 accounts from so many healthy accounts is a significant challenge faced by the American Red Cross. The goal of this study is to apply state-of-the-art machine-learning techniques and learning–to-rank algorithms to predict the small portion of customers who are most likely to provide a positive response to a donation solicitation.

3.2 DATA PREPROCESSING

There are different types of donations in the data set; one may write a big check to the American Red Cross while others may regularly spare $50 from a paycheck. The American Red Cross received $31 million donations after 2009, 99.86% of which are no more than $1,000. Even so, these donations amount to $28.8 million: 93% of total donations received. In addition, out of 269,277 total donors, only 503 accounts donated more than $1,000. These high-value customers are indeed important to the organization, but they can be treated as special donors and are handled manually by the American Red Cross employees. Our focus is to study how to apply statistical methods to aid managing a large corpus of customers. Thus, we remove all donation transactions amounting to $1,000 or more. This study treats the donation as a regular purchase behavior rather than a one-time event.

The data set consists of complete contact and donation history for each particular customer, from which we extract the following variables to rank customers: total number of donations; donation frequency (donations made over a given period); total number of contacts since last donation; time since last donation; time since last contact; and contact frequency. In addition, the average donation amount and demographic variables—race, education, income, gender, age, and church activity—are also taken into consideration. Note that the demographic information is not at individual level but the average at zip code level or county level.

We divide the data into two subsets. Data from 2009 to 2010 serve as the training data to tune the predictive model. Data recorded in 2011 are used as the test data set to verify our prediction. The descriptive statistics are summarized in Table 3-1.

Table 3-1 Descriptive statistics for the training (2009-2010) and test (2011) data sets

	Year 2009-2010	Year 2011
Number of people who made donations	230,488	94,458
Number of donations	380,239	128,228
Number of campaigns	38	26
Number of people been contacted	17,761	16,110
Number of interactions	236,777	133,857 (96,536 until 10/01/2011) [1]
Number of donations after interactions	16,004	5,317
Success rate	6.76%	5.51%*

[1] We consider only the interactions the American Red Cross has made until October 1, 2011 since the response to the interaction will be censored at the end of year 2011. We cannot accurately observe the response to the interaction after October 1, 2011.

Furthermore, due to their different donation tendencies, we find it necessary to build predictive models separately for different American Red Cross defined groups, such as "current supporters," "lapsed customers," etc., to more accurately capture this difference. See Table 3-2 for more details.

Table 3-2 Donation tendencies for various groups of customers as classified by the American Red Cross

	Contacts	Donations	Success Rate (%)
All	141,148	8,667	6.14
Current Supporter	116,008	4,935	4.25
Lapsed	30,086	808	2.69
Prequalified Lead	13,851	2,500	18.05
Others	14,354	918	6.40

3.3 METHODS

This section discusses the various ranking algorithms, including logistic regression, gradient boosting tree and pairwise boosting ranking algorithm to predict and recommend a list of people who are most likely to donate after being contacted by the American Red Cross.

Logistic Regression. This method uses the log odd from the logistic regression output as the ranking score to prioritize customers. One major drawback associated with logistic regression is that it does not handle missing value well. Logistic regression excludes the incomplete observations from the training data set or use attribute mean to impute the missing values. Both approaches perform poorly on our ranking task since only 16% of data has complete information on all independent variables. The majority of the missing values occur in demographic attributes.

To mitigate the missing value problem, we only include total counts, frequency, recency of donation, and recency of contact in the logistic regression model.

Gradient boosting tree. In contrast to logistic regression, the gradient boosting tree (GBT) has the advantage of using surrogate splitting attributes to deal with the missing values. In the application to rank-lapsed customers, we build 2,500 successive trees in the boosting process and calculate the training error, cross validation error, and test error at each iteration. (See Figure 3-1 for an illustration.) To achieve highest prediction power, the optimal number of trees to be included in the final GBT model should correspond to minimal cross validation error. As we can see in Figure 3-1, the minimum of the cross validation error is achieved at iteration 442. After iteration 442, the training error keeps decreasing but the out-of-sample error has started to increase, which indicates overfitting. Therefore, we use the first 442 boosting trees to predict the donation outcome for the test data set.

Gradient Boosting Ranking: The newly introduced learning-to-rank techniques—GBRank pairwise boosting ranking algorithm and its adapted version—are also be applied and their performances will be compared with pointwise methods (logistic regression and gradient boosting tree). Results shown in Sections 3.4–3.7 are obtained using the original pairwise boosting ranking algorithm. Section 3.8 compares the results from the original and the adapted algorithms.

The descriptions of covariates used in the algorithms can be found in Appendix A.

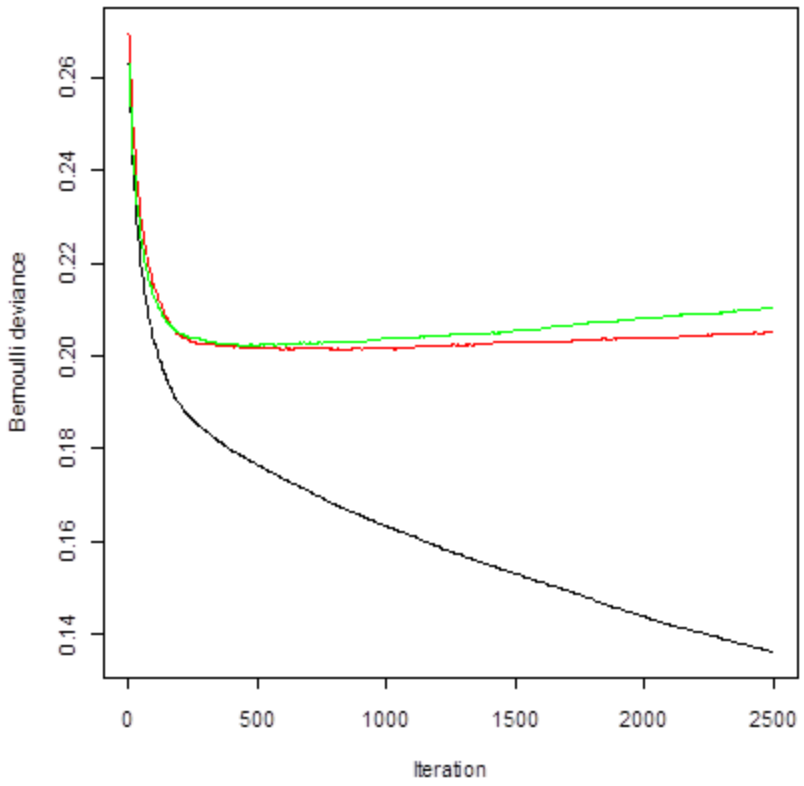


Figure 3-1 The training error (black), cross validation error (green) and test error (red) for boosting regression trees as functions of iteration numbers

As stated in Chapter 2, pairwise algorithms deal with preference data and thus we need to build our own preference data to meet this requirement. The preference data consists of about 144,000 entries, with each entry indicating the relative preference between two customers in a pair. Each pair consists of one customer d_i with positive response to a donation campaign and the other $d_j$with negative response. In such a case, the entry is recorded as $d_i > d_j$. No such pair contains two customers with both positive or both negative responses because there is no clear order between them. Here, we consider all campaigns in the training datasets (2009-2010)

homogeneous in nature and customer responses in one campaign comparable to those in another. Namely, we build the preference data by comparing all customer responses within campaigns and across campaigns.

To ensure the convergence of pairwise boosting ranking algorithms, we plot the number of contradicting pairs. A contradicting pair occurs when the preferred customer is assigned a lower ranking score. From Figure 3-2, we see that, in the training process of ranking "lapsed customers," the number of contradicting pairs in both training and test data sets appear to have stabilized after 100 iterations.

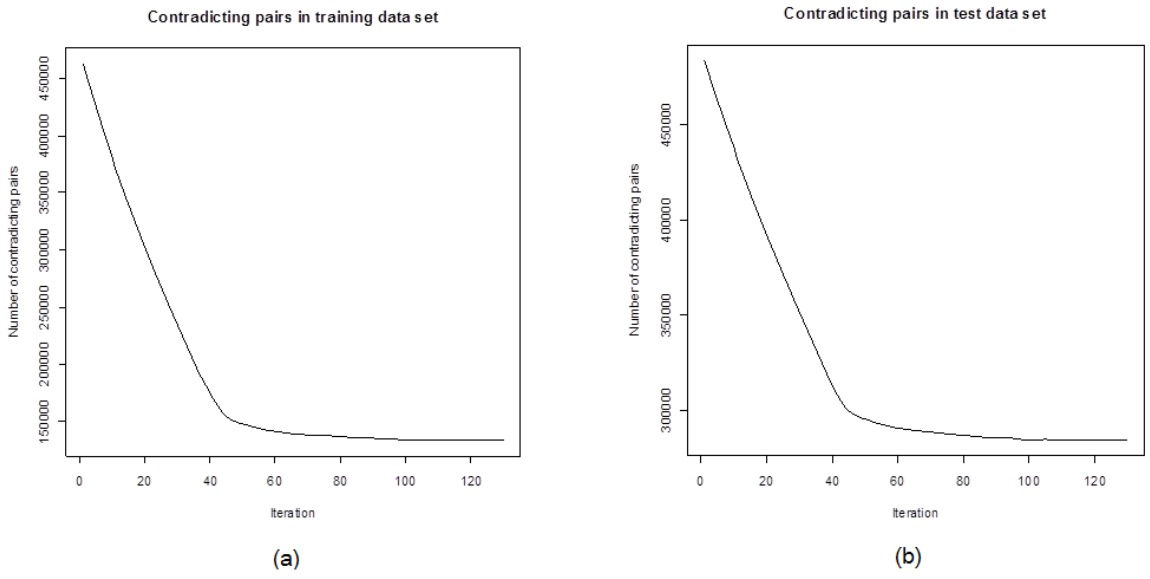


Figure 3-2 Numbers of contradicting pairs in the (a) training and (b) test data set as functions of iteration numbers

3.4 PREDICTION PERFORMANCE

In this section, we will first define a proper measure which can be used to evaluate model performance on the ARC data. We will then use this measure to compare different customer ranking algorithms discussed in the previous section. We find that the newly introduced pairwise ranking algorithm performs better on the ARC data than the traditional methods in terms of its prediction accuracy.

Conventional evaluation metrics such as *Precision@K, Discounted cumulative Gain (DCG)* are not applicable in the ARC data because we do not know the ground truth order of people's donation tendencies. However, we do know the outcome of an engagement, namely whether the American Red Cross successfully persuaded a particular customer to make a donation or not. In light of this, we need to improvise a new measure tailored to this particular scenario.

Suppose that we are now evaluating a predictive ranking algorithm which has generated a ranked list π_M of length M by comparing with the test data set. From the test data set, we know that there are a total of N customers who made the donations. If the American Red Cross contacted the first X customers ($X \leq M$) according to the ranked list π_M, then $K(X)$ of them would actually make donations. $K(X)$ must be a fraction of N, and we define

$$\boldsymbol{p(X) = \frac{K(X)}{N}} \tag{6}$$

as the proportion of True Valuable Customers covered by the X contacts. We name p(X) as the Valuable Customer Coverage Curve (VCCC). $p(X)$ has the following properties:

1. $p(X)$ monotonically increases with X
2. $p(X)$ bounded between 0 and min ($\frac{X}{N}$, 1).
3. $p(X)$ approaches 1 as X increases toward M where M is large enough
4. $p(X) \equiv \frac{X}{N}$ indicates perfect ranking. In practice, the closer $p(X)$ to $\frac{X}{N}$ the better ranking.

The reasons for the properties are as follows: $K(X)$ monotonically increases with X and is bounded between 0 and $min(X, N)$. This is because, as the American Red Cross contacts more customers down the ranking list, more people will make the donations until it reaches all N donors. This naturally leads to the above-listed properties 1 and 2. In practice, we make M large enough to cover all N donors, and thus min($\frac{X}{N}$, 1) ~ 1 as $X \uparrow M$, leading to property 3. Property 4 is true because $p(X) \equiv \frac{K(X)}{N}$ means that all contacted customers make donations and it marks a perfect ranking. Similarly to AUC for ROC curve, we calculate the area under the VCCC curve to come up with a numeric measure to evaluate the ranking algorithm.

From Figure 3-3 and Figure 3-4, we find that the pairwise boosting ranking algorithm significantly outperforms the other two approaches since its VCCC curve for stays well above the other two curves. Its advantage is especially obvious in ranking lapsed donors. Table 3-3 shows the area under the VCCC for all three algorithms, where the pairwise boosting ranking algorithm has the highest score in both scenarios. In the scenario to manage lapsed customers, the American Red Cross only needs to distribute 20% of printed campaign mails based on the boosting ranking score to reach 80% coverage of the customers who eventually will make a donation. In contrast, with predictions using boosting regression tree and logistic regression, the American Red Cross will need to distribute 60% and 80%, respectively, to cover the same number of valuable customers.

Both scenarios show that the pairwise boosting ranking algorithm improves prediction accuracy more obviously in the middle range, i.e., when the American Red Cross intends to contact 20-40% of its potential donors. Note that, on some occasions, the boosting ranking algorithm may not perform as well as the gradient boosting tree method, such as in the very top segment for lapsed donors. However, it is an acceptable compromise that we have chosen to make to achieve our purpose. Our purpose is to make better predictions for the middle ranking range as customers in this range are crucial in customer relationship management. These customers can be deemed as customers walking around the borderline. The extremely loyal customers are hard to lose and the extremely low-intentioned customers are hard to retain; the customers in the middle are worth great effort to maintain. This is the reason why we do not want to overemphasize the very top customers to cannibalize the customers at the borderline. Therefore, we intentionally designed our algorithm such that its loss function does not penalize mistakenly ranked instances at top positions more severely than those at the bottom. This is very different from any typical ranking algorithm in information retrieval in which incorrect orders on the top list are penalized more severely to ensure greater accuracy for only the first few entries.

The logistic regression almost follows the diagonal line with the area under the curve close to 0.5, which implies that the customers at the top rank list have nearly the same probability to make a donation as those at the bottom of the list. The ranking based on logistic regression prediction is not effective.

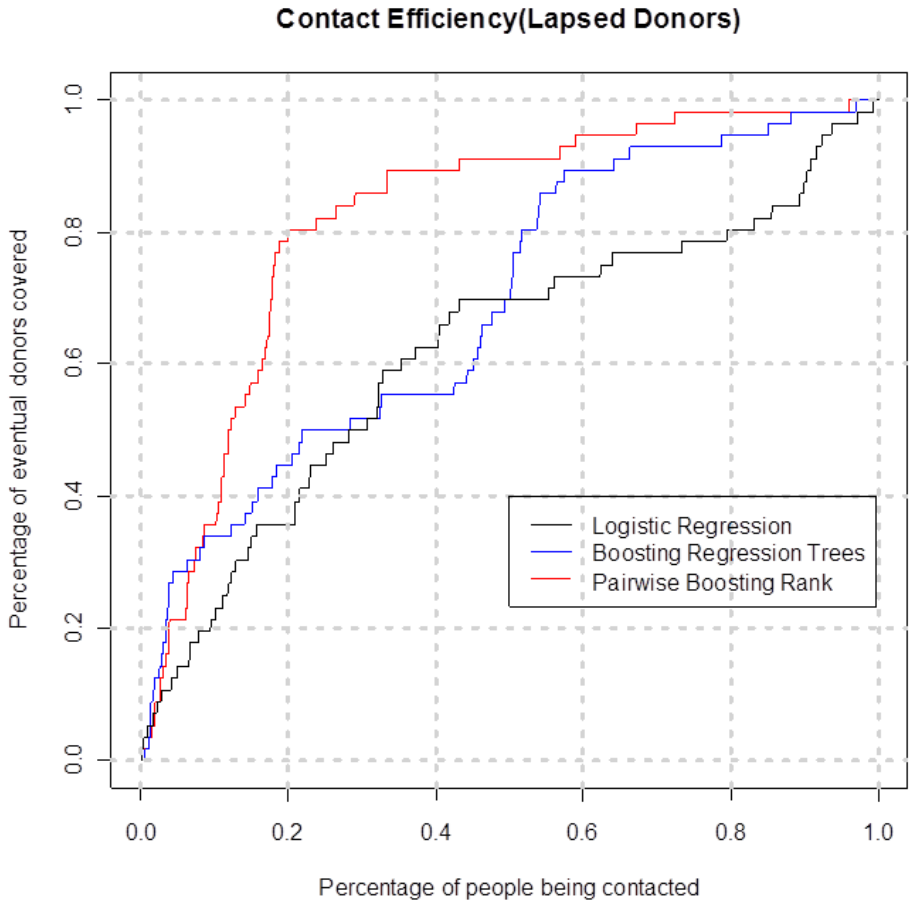


Figure 3-3 Ranking performance on the "Lapsed Donors"

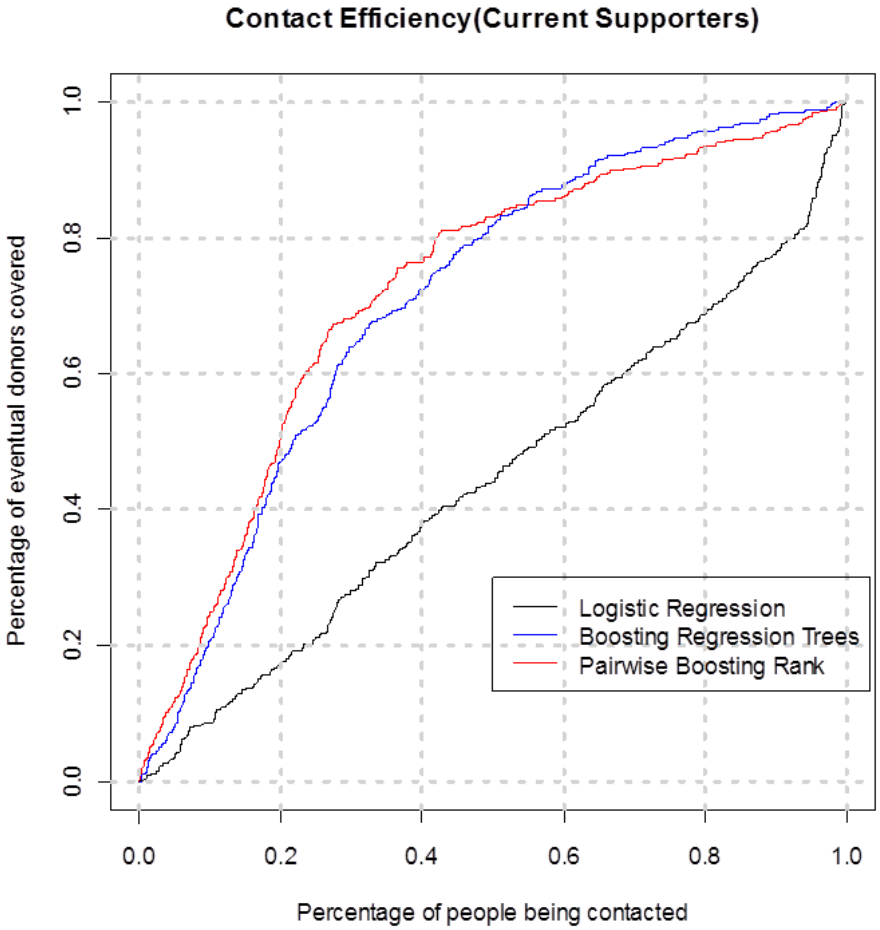


Figure 3-4 Ranking performance on the "Current Supporters"

Table 3-3 Area under the VCCC curves using three different machine-learning techiniques in predicting donation responses for lapsed donors and current supporters

	Lapsed Donors	Current Supporters
Pairwise Boosting Rank	0.822	0.722
Boosting Regression Trees	0.720	0.709
Logistic Regression	0.619	0.444

3.5 HIGH-VALUE AND LOW-VALUE CUSTOMERS

Ranking scores from the boosting ranking algorithm also facilitate the segmentation of customers into high- and low-donation tendencies. A closer look at these two groups separately can help reveal important factors in customers' decision making and correspondingly provide strategy in customer relationship management.

Figure 3-5 shows the score distributions at different iterations in lapsed donors ranking process. The scores range from 0 to 1000. After the boosting process converges, the distribution manifests a polarized shape. Nineteen percent of customers fall into the top tier, with scores above 900, while 64% of customers get squeezed into the bottom range, with the scores below 100. In other words, the ranking algorithm tends to separate customers into two disjoint sets: high-value customers and low-value customers. To gain insights of what factors attribute to the customer separation, we conduct a t-test to compare the sample mean of all attributes between these two groups with high and low donation tendencies. Table 3-4 presents the t-test results.

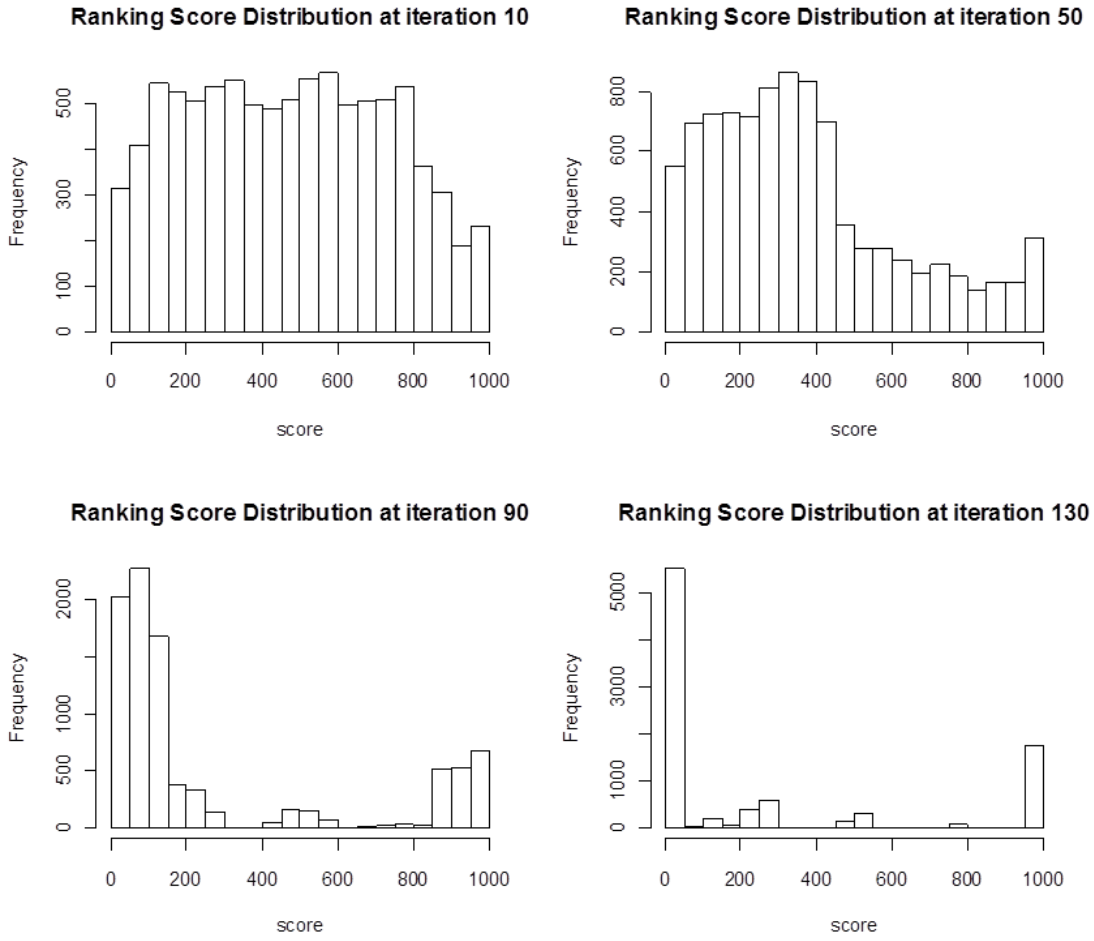


Figure 3-5 Score distributions at different iterations in the lapsed donors ranking, ranging from 0 to 1000

Table 3-4 Results of t-tests for sample means of all attributes between high- and low-donation tendency groups

	High Group	Low Group	p-value
Donation amount	52.99	52.1	0.81
Total number of donations	1.48	1.44	0.07*
Total number of contacts since last donation	18.53	18.76	0.2
Donation frequency	181.54	173.6	0.09*
Contact frequency	54.62	54.64	0.96
Time since last donation	1031.29	1042.28	0.14
Time since last contact	87.85	77.25	0**
Age	39.71	39.53	0.21
Gender	48.72	48.71	0.87
Race	78.56	77.9	0.15
Education	58.14	58.05	0.72
Income	10.99	10.98	0.08*
Church	524.05	524.62	0.86

* $0.05 < p$-value < 0.1 **$0.01 < p$-value < 0.05

The high-value group has a significantly higher total number of donations and donation frequency. In contrast, there is no significant difference in the total number of contacts or contact frequency between the high-value group and the low-value group. This suggests that, for the dormant customers who have not donated for a long time, simply increasing the contact intensity may not encourage them to donate again. Their donation history may serve as a good indicator for the likelihood of whether the lapsed customer will make another donation in the future. The most significant variable in the t-test is the time since last contact. Surprisingly, the high-value group has a much greater time span since the last contact. Demographic information has minor impact on the donation tendency ranking. People with higher incomes are likely to donate again. This implies that the best strategy to manage lapsed customers is to pick up customers with high incomes who have donated more times in the past while not contacting them too frequently. Sending them occasional reminders would be just enough.

3.6 WHAT'S THE DIFFERENCE BETWEEN LAPSED DONORS AND CURRENT SUPPORTERS?

In this section, we investigate the difference between lapsed donors and current supporters. In particular, we will address what attributes make the highest impact on the donation tendency for each group. This is done using boosting regression tree, which yields a score of relative importance for each attribute by summarizing their involvement in the ensemble tree splittings. Table 3-5 and Table 3-6 list the relative importance of attributes for the lapsed donors and

current supporters, respectively. When comparing these two groups, we find that their top influential attributes are very different: it appears that intensity of contact plays important a role in persuading lapsed donors to make donations again, while the donation recency is a dominant factor in current supporters' decision making.

A noteworthy inconsistency exists between results using different machine-learning techniques, one that may lead to contradictory strategies for handling lapsed customers. In the last section, we have shown that the boosting ranking algorithm advises against contacting lapsed customers too frequently because the intensity of contact does not appear to play an important role in stimulating donation tendencies for these customers. In contrast, according to the boosting regression tree, it is recommended to contact lapsed customers more frequently and more times to spur future donation. The reason for the inconsistency is not well understood yet, and it is worth exploring later.

Table 3-5 Relative importance of attributes in lapsed donors' donation tendency based on results from the boosting regresion tree meethod

	Variables	Relative Importance (%)
1	Contact_Frequency	40.65
2	Number_Contacts	13.31
3	Timesincelastdonation	7.63
4	Number_Donations	6.36
5	Church	4.41
6	Donation_Frequency	4.36
7	Timesincelastcontact	3.97
8	Race	3.94
9	Education	3.94
10	Income	3.54
11	Gender	3.12
12	Age	2.72
13	Average_Donations	2.05

Table 3-6 Relative importance of attributes in current supporters' donation tendency based on results from the boosting regression tree meethod

	Variables	Relative Importance (%)
1	Timesincelastdonation	66.88
2	Average_Donations	6.65
3	Number_Contacts	6.00
4	Donation_Frequency	4.94
5	Timesincelastcontact	3.06
6	Race	2.90
7	Gender	2.08
8	Contact_Frequency	2.00
9	Number_Donations	1.71
10	Education	1.19
11	Church	1.17
12	Income	0.81
13	Age	0.61

3.7 LONGITUDINAL ANALYSIS ON RANKING SCORE

One interesting problem in customer relationship management is the task of monitoring customers' loyalty over time. How does a customer's attitude change after certain encounters

with the company? For instance, does a customer's interest in a company drop after he or she completes a transaction? In this section, we investigate the sentimental development after people make donations or after being contacted by the American Red Cross several times.

We monitor the changes in the ranking scores before and after a customer is contacted, and before and after a donation. From Table 3-7, we observe that the ranking score will increase significantly after a customer makes a donation. However, the magnitude of the change has great variance and depends on other characteristics of the customer. No significant changes on ranking scores are observed after a customer gets contacted by the American Red Cross. This indicates that contacting customers frequently does not have adverse effects on customers' willingness to donate.

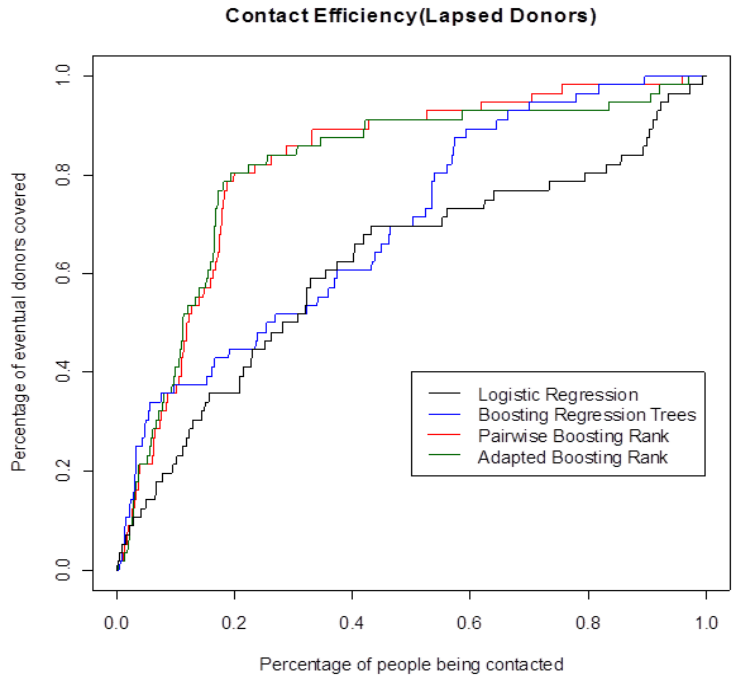
Table 3-7 Results of t-tests for sample means of ranking scores before and after customers receive contacts from the American Red Cross, and before and after they make donations

	Before	After	95% confident interval	p-value
Contact	277.7611	269.2674	[-3.75 20.74]	0.174
Donation	263.1055	326.1806	[6.42,119.72]	0.029*

* $0.05 < p$-value < 0.1 **$0.01 < p$-value < 0.05

3.8 IMPROVEMENT OF THE ADAPTED ALGORITHM OVER THE ORIGINAL PAIRWISE BOOSTING RANKING ALGORITHM

In chapter 2, we highlighted the adaptions made to the original pairwise boosting ranking algorithm in order to tailor it to our direct marketing problem. We now compare its performance with the original algorithm. From Figure 3-6a, we see that the adapted and the original boosting ranking algorithms perform similarly. In fact, the original ranking algorithm has a higher area under VCCC. However, a closer look at the VCCC for the top 5% to 20% of customers reveals that our adapted boosting ranking algorithm outperforms the original one in this segment (see Figure 3-6b). This is because the adaptive algorithm considers position information. Recall that, in the original boosting ranking algorithm, the scores are adjusted according to their functional gradients at each iteration. The adapted algorithm manipulates the adjustment such that it not only depends on the gradients, but also the relative position of the candidate at the current iteration. This position dependent gradient adjustment makes it harder for a candidate to climb up the ranking list to reach the top range. In addition, it also makes it easier for a candidate to slide down the ranking list if it was mistakenly positioned in the top range. This adjustment favors the ranking accuracy on the top rank list.



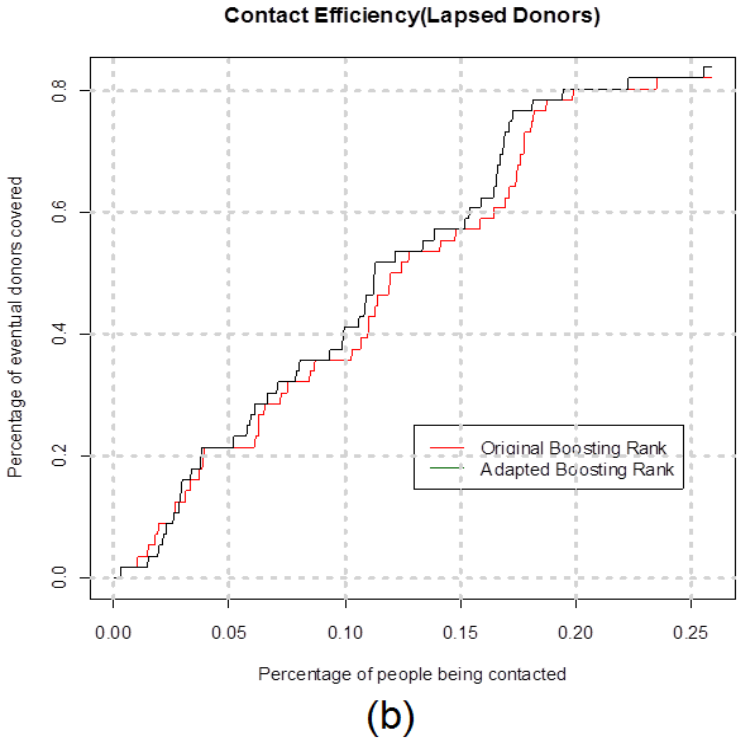


Figure 3-6 Ranking performances on the "lapsed donors" using original and adapted boosting ranking algorithms. (a) Overview of ranking performances. (b) Enlarged near the top list.

4.0 MODELING BROWSING BEHAVIOR AND AD CLICK INTENTION ON A HEDONIC WEB SITE

4.1 INTRODUCTION

There are two types of Internet business models: e-commerce and site traffic-based revenue generation models. The former model prevailed at the onset of Internet business. A natural thought is to imitate business activities from the real world in a new online environment. Business owners typically set up a web site to sell products or to provide services to site viewers. The advantages of e-commerce relative to traditional marketing channels include easy access to products, time saved travelling to stores and the ability to compare product prices and features quickly (Schaupp and Bélanger (2005)). Amazon.com is an example of a successful e-commerce company.

As new information technologies rapidly develop, many people's lives have significantly changed. Besides purchasing goods or services online, people can use the Internet to study and learn, collect information, communicate with others, stay entertained online, and more. Basically, people access the Internet to search, browse, and consume information. Due to the convenience and ubiquity of the Internet, good web site content can easily attract millions of viewers. Many industry management practitioners treat the site traffic as a type of asset that can

generate tremendous profits. Google Inc. valuated YouTube's site traffic at $1.65 billion when the acquisition occurred in 2006. Similarly, Adobe Systems Inc. bought Omniture for $1.8 billion (Clark and Vranica (Sept. 16, 2009)). Omniture produces software that monitors consumers' web-browsing habits and enables clients to tailor their advertising accordingly.

Although it is still unclear for many companies exactly how to produce steady cash flow through web sites, much like Facebook.com and Twitter.com, advertising is probably the simplest way to exploit the volume of site traffic (Vascellaro (Apr. 22 2010)). However, the key success factors for a web site that primarily focuses on web content are quite different than those for an e-commerce site. E-commerce sites are concerned with the conversion rate while web content sites are concerned with the *ad click-through rate* and *web traffic volume*, two major factors cited in negotiations with ads partners. In the marketing literature, many researchers have investigated users' behavior on an e-commerce web site and suggested how to better predict and improve browsers' intention to buy after they viewed certain web pages. To our best knowledge, no research has empirically examined users' actions on a hedonic-content web site.

Concerning research about online advertising, Chatterjee, Hoffman et al. (2003) examined the relationship between a user's browsing behavior and his/her propensity to click a banner ad. Recent advances in broadband networking technology make it possible to transmit large bulk data in a short time. For instance, people now watch online videos regularly, and, as a result, video clips have become a new media tool for advertising. However, compared to the traditional banner ad, a video ad is more disruptive and inescapable: viewers must finish the ad before they are allowed to watch their target video. Briggs and Hollis (1997) claimed that whether and how

long an advertisement can be remembered by consumers varies widely in different forms of media. Their research showed that a web banner ad is more similar to conventional print than a TV advertisement because print and banner ads emphasize viewer involvement and prominent placement while TV advertising requires more creative power to draw viewers' active attention. These distinctions imply that users' behaviors under video advertisements may differ from those under banner ads. For example, the banner ad *click-through rate* is generally below 0.3% (eMarketer (2002)), but the *click-through rate* on our video ad dataset is about 0.7%. Thus, it is important to reexamine how a user's actions on a web content site affect his/her proneness to click on an advertisement.

Besides ad click-through, another concern of site managers is that the mandatory video ad could hinder users' site activities. Some visitors close the video play page even before the video ad ends. The negative effects of video ads are so severe that many users leave before the target videos start. MacMillan (2009) reported that online viewers are more impatient than TV watchers and running 15- or 30-second ads may deter online viewers.

In this chapter, we will analyze a clickstream dataset collected from a video streaming site to empirically investigate the gravity of the negative effect of video ad clips and study influential factors that can affect ad click-through rates. Our contribution is two-fold. First, we aim to potentially help the site to improve the click-through rate, which is a core metric for a web content hosting business. Clicking a video ad clip will take site visitors to advertisers' sites and directly benefit partners' business. YouTube overlaid a corresponding "click to buy DVD" advertisement on a video clip "Monty Python" and DVD sales increased by 23,000%

(MacMillan (2009)). Therefore, click-through rate is one major determinant of how much profit a hedonic-content website can gain from its advertisement partner.

On the other hand, our research can help managers estimate to what extent the intrusive video ads will discourage site viewers to continue watching videos on this site. Every hedonic-content site understands that site traffic is a key for the site to survive. The managers definitely do not want to see potential customers leave the site due to annoying advertisements. We will provide several suggestions about how to choose the frequency and timing of video ads so that visitors can watch a reasonable amount of ad exposures without losing great interest in the site.

4.2 LITERATURE REVIEW

4.2.1 Research on clickstream data

Compared to traditional brick-and-mortar stores, it is much easier for e-commerce sites to track transactions and users' actions. Generally, a web site tracks the path of URLs a visitor takes (i.e., clickstream) and every click event a visitor triggers on each web page. By analyzing the rich information embedded in the clickstream data, management can understand visitors' decision choices, discern their purchase intention and predict their next decision better than with the traditional scanner panel data because clickstream data contain visitors' decision paths before they make a final decision. Internet choice behavior is different from choice behavior in a supermarket setting. Bucklin, Lattin et al. (2002) summarized clickstream research in a 2-by-2

matrix. One dimension depends on whether the research focuses on a single site or across several sites. Cross-site studies generally require user-centric data instead of site-centric data. To obtain user-centric data, a monitoring device is usually installed client-side to track the usage of each individual's web browser. An example of a cross-site study can be found in (Park and Fader (2004)), which noted that a person's browsing behavior at one site correlates with his/her browsing behavior at another site. Johnson, Moe et al. (2004) studied consumers' cross-site search behavior and found that consumers search no more than two sites before they decide to buy a certain product. The data used in this study come from the server log of a web site. Thus, it is a site-centric data set and we will limit our discussion to the research within one single site. The other dimension is the objective of the individual, i.e., browse versus purchase. The research in the former category studies the site navigation behavior and research in the latter category studies the relationship between the individual's browsing behavior and conversion rate. Our research focuses on how to predict site visitors' intentions to click video advertisement clips based upon their browsing behavior and how the video advertisement in turn influences visitors' browsing behavior.

4.2.2 Browsing behavior

Huberman, Pirolli et al. (1998) took the first step in studying web surfing patterns. In their model, each page has a certain utility and the utility of the current page U_t equals the utility carried over from the previous page plus a random Gaussian noise, i.e., $U_t = U_{t-1} + \epsilon_t$. When the page utility exceeds a certain threshold, the user continues to request an additional web page. Otherwise, the user stops web browsing. Based on the page utility model, they derived that the distribution of the total number of pages viewed by site visitors follows a power law distribution,

which is a long tail distribution. Hence, strong regularity exists in web browsing. While a small number of users in the population browse many pages, the majority views only a few pages. They also warned that care must be exercised when researchers use the average number of clicks to study the depth of web surfing as the average over-estimates the browse depth. One drawback of this study is that the authors do not take covariates into account and hence cannot explain why some users browse more pages than others. In addition, their analysis is based on aggregated data, while many researchers have pointed out that results from aggregate level analysis might be misleading and can possibly mask some unique patterns at the individual level. In addition, their analysis did not reflect the dynamic nature of web browsing.

Bucklin and Sismeiro (2003) continue to adopt the page utility assumption in a binary probit choice model. It is easy to show that the binary probit choice model is consistent with that of Huberman, Pirolli et al. (1998) except that instead of a random walk structure, Bucklin and Sismeiro (2003) tie page utility to variables that reflect browsing behavior. In addition to modeling a user's decision to stay or exit browsing, they also model the time duration that users spend on each page. They demonstrated several phenomena during a web-browsing process.

4.2.2.1 Learning effect and involvement

Learning effect is persistent in repetitive visits. Visitors spend less time when they return to the site because they become more familiar with the web interface and site structure and can navigate the site more efficiently. Johnson, Bellman et al. (2003) empirically showed that the cognitive cost of using a site decreases with experience and the decrease can be modeled by a power law of practice. Furthermore, the reduction in time spent when people return to a site is due to fewer page views but not the time spent on each individual page (Bucklin and Sismeiro

(2003)). This may indicate that the learning effects are due to better site navigation rather than faster processing of information.

As for within-site behavior, there are two counter forces governing browsing behavior: visitor involvement and time constraints (Bucklin and Sismeiro (2003)). As visitors go deeper into the site, they may become more involved and hence have larger propensity to request additional pages and stay longer at each page. However, after visitors have browsed many pages, they are running out of time. Thus, they may spend less time on each page and have a low probability to request another page. Their results show that the involvement effect dominates time constraint in terms of page-view duration, while time constraints play a more important role in terms of the probability of requesting additional pages.

4.2.2.2 Dynamics and evolvement

Visitor behavior at a site is dynamic. The initial intention why a visitor came to this site is not clear. A visitor could come to the site through a search engine with a predefined goal, or he/she might come casually without a particular objective at all. In addition, a visitor's state of mind may change from time to time. When visitors who browse web pages casually find a particular attractive topic, they may start to focus on this topic and spend more time on related pages. When visitors get bored of the topic, they may switch from the focused status to another topic randomly. Moreover, a visitor's attitude or behavior can change after repeat visits. Moe and Fader (2004) examined how visit frequencies can predict a visitor's purchase intention. They found evidence that, although higher visit frequency does lead to a greater propensity to buy, the increase in an individual's visit frequency is a stronger indicator in predicting purchase intention even though the frequency itself is low.

Many clickstream researchers consider the dynamics and evolution of browsing behavior in their models, either within-session or cross-session. The easiest way is to model some relevant parameters as a function of some measurement of time. For instance, Moe and Fader (2004) allow the purchase motive gained at each site visit and purchase threshold to increase or decrease exponentially over time. Chatterjee, Hoffman et al. (2003) consider several coefficients in their logit model as a linear function of the number of sessions that visitors have spent. Johnson, Moe et al. (2004) model users' linear increase in search propensity with respect to the logarithm of the months. Montgomery, Li et al. (2004) introduce a vector autoregressive component in their model to capture the dynamics of users' browsing behavior.

4.2.2.3 Heterogeneity

Many researchers have pointed out that web analytic measuring at the aggregate level might be misleading and can possibly disguise some unique patterns at the individual level. For example, Bucklin and Sismeiro (2003) showed that the page-view duration is positively correlated to the number of visits to the site at the aggregate level but uncorrelated at the individual level due to the compensation of learning effect and involvement. Moe and Fader (2004) studied conversion behavior at an online book store where the conversion rate for the whole population increased from 13.2% to 14.7%. However, for loyal customers who made at least one visit in the first two months and last two months, the conversion rate dropped from 26% to 20.8%. Because analysis at the aggregate level risks providing incorrect managerial opinions, we argue that browsing behavior should be modeled at the individual level.

4.2.2.4 Types of browsing behavior

The motivation for visiting a site varies by the user. Browsing behavior can be classified into several categories. Childers, Carr et al. (2001) found that, besides traditional utilitarian motivations, when users go to online retail stores, the immersive, hedonic aspects of web user interactions also occur during the browsing process and may become another motivation to visit the site. Browsing behaviors under these two motivations have different characteristics. Similarly, Janiszewski (1998) classifies offline shopping behaviors into exploratory and goal-directed search. Moe (2003) extended this dichomatic classification into four categories: directed buyers, search /deliberation visitors, hedonic browsers, and knowledge-building visitors. Visitors in each category have their signature browsing patterns. Directed buyers come to the site with preset products to buy. Deliberation visitors have a general category of products in mind and come to the site to gather more information. Hedonic browsers are casual visitors without particular products in mind though the site interface and stimuli can motivate impulsive purchase. Finally, knowledge-building visitors come to the site to do research on products available without any intention to buy. In our application, some visitors jump onto the site through a search engine with a clear target video to watch (goal-directed search) while others come to the site seeking any videos that can entertain them (hedonic browsing). Switching between different categories is obviously permissible. For instance, after a goal-directed visitor watches the desired video, he/she may not leave immediately but show interest in other videos promoted by the site, switching his/her status to that of hedonic browsing.

Working from concepts developed by Csíkszentmihályi (1975), Hoffman and Novak (1996) constructed a flow framework to model the types of browsing behavior. They emphasize user

experience during a browsing session rather than the intention to purchase. The researchers model the browsing process as a series of stimulus-response pairs between the site and the visitor. They define the flow experience as a site navigation status when visitors lose self-consciousness to some extent and self-reinforce that the page is intrinsically enjoyable. At this status, visitors may not even notice time passing. When a visitor's skill set fits the web site's challenge, and the web content is appropriate and interesting, he/she can get involved deeply and generate focused attention—a necessary condition for the visitor to reach flow experience. Novak, Hoffman et al. (2000)) claimed that the flow experience is higher for visitors with hedonic or experimental motivation than those who are task-oriented and utilitarian motivated.

4.2.3 Conversion at e-commerce sites

The key problem in studying the conversion rate at e-commerce sites is to predict conversion and identify the factors that influence a site visitor's purchase decision. Since the conversion rate is very low—typically below 3%—predicting conversion is not easy. Sismeiro and Bucklin (2004) claimed that a purchase decision process can be decomposed into several tasks. The completion of each task reveals the progress of a user's purchase intention. Browsing behavior varies from task to task. In the authors' application of online car selling, among users who have completed a car configuration task, the ones who spend more time but request few pages are more likely to eventually buy a car. In addition, they found that the same variable may have a positive effect on the completion on one task but a negative effect on another. A single-stage model cannot discover the dynamics and evolvement of a user's browsing behavior. Similarly, Moe (2006) divided the online nutritional product purchase process into two stages: choosing a product and purchasing a product. Consumers behave differently in the two stages.

At the early stage, they tend to use simple attributes, such as size or color, to choose from available products while making decisions deliberately at the later purchase stage. In contrast to considering user's decisions to occur on the task completion, Montgomery, Li et al. (2004) adopted a hidden Markov model on dynamic multinomial probit model to examine the browsing behavior in a more detailed level. After viewing each web page, a visitor chooses whether to stay or exit the site. If the visitor stays, the challenge is to determine which type of pages he/she would like to continue to browse. All web pages are categorized into seven categories. Based upon a user's browsing path, the model can predict after six page viewings whether this user will reach the category of "Order" (making a purchase) with 40% accuracy. The authors also found that the utility obtained from one page may spill over to succeeding page views.

A user's goals or state of mind may change abruptly during the web browsing process. The research of Montgomery et al. (2004) focused only on the modeling browsing path within a session without considering the repeat visits. To incorporate repeat visits, Scott and Hann (2007) add another session-level hidden Markov chain on the top of within-session page level hidden Markov chain as that in Montgomery et al. (2004). The hidden states at the session level can help to cluster browsing behavior into three types: decisive shoppers, deliberators and "never buyers." Decisive shoppers rarely place an item in the shopping cart but are very likely to complete the purchase if/when they do so. Deliberators collect a lot of information before they make the purchase. "Never buyers" come to the site to check price or product features and buy it somewhere else. Another finding related to repeat visit comes from Moe and Fader (2004), who

argued that subsequent visits have diminishing impact on purchasing behavior, and purchasing threshold increases as potential customers revisit the site.

4.2.4 Advertising

Research about the effectiveness of banner ads on web pages has drawn attention from multiple disciplines. In this paper, we will restrict our topic to how a web user's browsing behavior reflects his/her propensity to click a banner ad. Chatterjee, Hoffman et al. (2003) specified a random coefficient logit model with evolution to examine *wirein* and *wireout* effects. In the wirein stage, additional ad exposure increased a user's tendency to click an ad, while, in the wireout stage, satiation makes users feel that ads are annoying and additional exposure has significant negative effects. The results show that wireout dominates wirein on intra-session exposures and earlier ads have a higher probability of being clicked. Across sessions, wirein dominates wireout. Longer intersession times in prior sessions, more banner exposures in previous sessions, and a longer time span since the last click generally leads to high proneness of ad clicks. In addition, the researchers found that new visitors and less frequent visitors are more likely to click on ads than more regular visitors.

Moe (2006) studied another type of disruptive advertisement, pop-up windows. The results suggested that, instead of an ad window popping up as soon as a user hits the target web page, a delayed popup window may increase the total number of pages viewed. However, delayed popup windows have no effect on click-through rates.

In our application, a complete video must be divided into several parts due to the limitation of the size of each file. The ad video is currently delivered in the beginning of the first part and no more ad videos are exposed in the beginning of ensuing parts. This result suggests that it might be wise to delay the ad to the later part of video content.

4.3 OVERVIEW OF THE DATA SET

The data set we obtained is from a video streaming service web site in China, which is very similar to www.hulu.com in the U.S. The site purchases TV programs and high definition movies from its partners and makes them available free for all site visitors. However, visitors must watch an advertisement video before they are allowed to watch the intended program or video. The advertisement presented to visitors is randomly assigned by the server system. From the server log, we extract one month of clickstream data, from April 1, 2009 to April 30, 2009. The data contains the URL of every web page that users have browsed as well as the time when each page was loaded. If a visitor jumps to the current page through a search engine, the keyword used in the search is available. Additionally, the data contain records of users' actions triggered by videos, including play, stop, pause and cancel. Other actions, such as fast-forwarding and/or rewinding the video by manually dragging or otherwise altering the progress bar, are not recorded. In the database, an advertisement video is recorded the same as regular videos except that it has a special flag to indicate that it is an advertisement program. When a user clicks on a video ad, it pops up a new window linked to the sponsor's web page and also triggers a pause action. Thus, by looking at the pause action attached to the advertisement, we are able to tell whether a user clicked on the video ad or not. However, we cannot track or monitor users'

actions after the initial click since the user will be led to the third party web page beyond our control.

A snapshot of sample data can be found below. In the "videotype" column, the number 1 indicates an advertisement while the number 0 indicates a regular video. The "status" column indicates users' actions on the video, like so: 1- play; 2- stop; 3-pause; 4- cancel pause; 5- load; 6- loading complete. The "clength" column references the time the user spent on the video. The "servertime" column marks the timestamp when the action on this video is triggered.

Table 4-1 A snapshot of sample data

userid	userip	program	status	isrepeat	videotype	clength	servertime	url
GFZjxvnZX	992169271	hdmovie_200902:	1	1	0	1517	6:11:29 PM	http://hd.openv.com/m
GFZjxvnZX	992169271	hdmovie_200902:	6	1	0	1462	6:15:36 PM	http://hd.openv.com/m
GFZjxvnZX	992169271	hdmovie_200902:	2	1	0	1517	6:16:32 PM	http://hd.openv.com/m
GFZjxvnZX	992169271	hdmovie_200902:	5	0	0	1819	6:16:33 PM	http://hd.openv.com/m
GFZjxvnZX	992169271	hdmovie_200902:	1	0	0	1819	6:16:33 PM	http://hd.openv.com/m
GFZjxvnZX	992169271	hdmovie_200902:	6	0	0	1751	6:20:28 PM	http://hd.openv.com/m
GFZjxvnZX	992169271	hdmovie_200902:	2	0	0	1819	6:21:36 PM	http://hd.openv.com/m
GFZjxvnZX	992169271	hdmovie_200902:	5	0	0	2121	6:21:37 PM	http://hd.openv.com/m
GFZjxvnZX	992169271	hdmovie_200902:	1	0	0	2121	6:21:37 PM	http://hd.openv.com/m
GFZjxvnZX	2030788662	GuangXiTVprog_2	2	0	1	7	10:24:26 PM	http://t.openv.com/zj/z
GFZjxvnZX	2030788662	GuangXiTVprog_2	1	0	0	13	10:24:26 PM	http://t.openv.com/zj/z
GFZjxvnZX	2030788662	GuangXiTVprog_2	5	0	0	13	10:24:26 PM	http://t.openv.com/zj/z

4.4 MODEL THE CLICKTHROUGH RATE

4.4.1 Each individual's click choice

We model visitors' browsing behavior and ad click intention at the individual level so as to easily incorporate visitors' heterogeneity and make a prediction of each visitor's click-through

probability. Prediction at the individual level is highly valuable for management. The web content displayed to online visitors during their browsing process is dynamic and interactive. If the site knows what factors will influence each visitor's ad click intention, the site can customize the timing, frequency or content of video ads to maximize the likelihood of visitors clicking video ads.

As with many advertisement research papers that model users' click-through decisions, we use the logit choice model to link users' browsing covariates with the probability that users will click on a video advertisement. Although Chatterjee, Hoffman et al. (2003) noted that the positive effect of an advertisement can spill over into the ensuing sections, we cannot track how many times a user has watched a video advertisement in repeated visits. Therefore, we model the click-through rate within one session. Let p_{io} be the probability that user i clicks on the o^{th} video advertisement exposure. p_{io} is not observable and we can only observe the click outcome C_{i0}. $C_{io} = 1$ if an ad click occurs and 0 otherwise. Thus,

$$C_{io} = Bernolli(p_{io}) \tag{7}$$

We use logit link function to connect click propensity to covariates.

$$p_{io} = Logit(\alpha_i + F_{io} + \beta_i X_{io}) \tag{8}$$

α_i is the intrinsic click intension for visitor i. F_{io} is the click proneness determined by the users' flow experience θ_{io}, which will be specified in the next section. X_{io} are covariates that can directly influence the click probability.

Cross-section variables not only affect visitors' browsing behavior due to the learning effect as discussed in section 2.3, but may also affect visitors' intention to click video ads. Chatterjee,

Hoffman et al. (2003) reported that the time span between sessions has a significant impact on click behavior, and cumulative ad exposures in previous sessions has a small positive effect on click probability. In addition, the researchers observed that, as the number of visits increases, the increased familiarity with the advertisement (wirein effect) may raise the probability to click ads. The learning effect in repeat visits also can affect visitors' ad click tendency. On one hand, the better understanding of the site structure and operational process may encourage visitors to neglect the time frame allocated to advertisement; they know every ad video runs 15 seconds and can intentionally browse other pages when the ad video is running, only to come back after 15 seconds to watch the target video. On the other hand, the learning effect enables visitors to browse web pages more efficiently and reduce anxiety under the time constraints; visitors are in a more relaxed mood and are more likely to watch the ad due to curiosity. We included cross-session covariates such as inter-session time in our pilot study and found they have no significant impact on visitors' intentions to click video ads. To mitigate the tremendous heterogeneity in our data set, we focus on the repeated visitors and the learning effects carried over sessions are minimum. Thus, we do not consider inter-session variables in our model.

4.4.2 Intra-session Covariates

Chatterjee, Hoffman et al. (2003) claimed that addition ad exposure generates negative marginal effect on ad click probability. The time since the last click and time since the last ad exposure affect users' click probability considerably because the ad satiation effect is relieved to some extent. Although their results came from an empirical study on banner ad instead of video advertisement, we include all of these variables in the model and reexamine whether the same results will hold for video advertisement. Unfortunately, the clickthrough rate is only about 1%

and the variable "time since last click" is not applicable in the other 99% occasions. Therefore, we use another simplified binary variable as the surrogate, whether the visitor has already clicked a video ad during the current session. In addition, Bucklin (2003) argued that the browsing depth and the time constraints may affect users' browsing behavior. We use two measurements "the number of pages browsed" and "number of videos viewed" to represent the construct "browsing depth". We calculate how long the visitor has been at the web site to reflect his/her time constraint. In summary, the intra-session component consists of the following five variables. The detailed variable descriptions can be found in appendix B.

$$X_{io} = \begin{bmatrix} NumberOfExposures \\ AlreadyClick \\ NumberOfVideos \\ NumberOfPages \\ TimeElapsed \end{bmatrix}$$

4.4.3 The hidden flow status

Montgomery, Li et al. (2004) and Scott and Hann (2007) modeled users' browsing types as discrete states of Markov chains. Most of the contents that visitors consume at a hedonic web are recreational videos. Thus, users' browsing processes are intrinsically motivated, experimental and ritualistically oriented instead of extrinsically motivated, goal-directed and instrumentally oriented as in product purchase practice.

We adopted the flow experience concept from the research of Hoffman and Novak (1996) to model a visitor's state of mind. Since Hoffman and Novak argued that the flow should be measured in a continuous space, we hypothesize that flow status θ_{iso} can fall at anywhere on an axis that measures the extent of recreational experience throughout the web surfing session. At

one end of the spectrum, a site visitor is completely immersed in the web contents and filters out other irrelevant perceptions such as time pressure. At the other end of spectrum, visitors have clearly defined objectives when they come to the site and exit the site immediately after they complete the predefined task. Stimuli displayed at the site cannot change their browsing path to complete the task.

Hoffman and Novak (1996) also noted that flow affects many browsing patterns, such as time duration, depth of search, repeat visit, navigation path, and so on. Visitors with high flow status are entertained by the web contents and generate positive recognition on site. They may watch the video ad earnestly and, as a result, have a higher probability to click it. Therefore, the term F_{io} in the logit model represents the amount of ad click intention attributed to visitors' flow status. It follows a normal distribution with the hidden flow status θ_{io} as the mean. Note that θ_{io} is not constant; rather, it varies within a session.

$$F_{io} = N(\theta_{io}, \sigma) \tag{9}$$

The hidden flow status θ_{is} is determined by a set of variables Z_{io}.

$$\theta_{io} = \pi_i + \eta_i Z_{io} \tag{10}$$

π_{is} represents flow status that cannot be reflected in browsing behavior specified in Z_{iso}. Holbrook and Gardner (1993) and Olney, Holbrook et al. (1991) argue that web surfing duration time and advertisement viewing time are good indicators of experimental versus goal-directed orientation. One prominent characteristic of information consumption on the web is interaction. Trevino and Webster (1992) use employees' control/actions on an email system as a measure of the flow construct. Pausing and resuming videos also to some extent reflect visitors' interest in the videos. Similarly, visitors will replay the video only if they enjoyed watching it and want to

repeat the pleasurable experience. Novak, Hoffman et al. (2000) conducted an empirical online survey and argued that flow is higher for users who use the web for experimental uses, such as online chatting or entertainment, than for task-oriented uses, such as work or searching for specific references. this justifies the usage of binary variable *"FromSearchengine"*. If a user jumps to the site from a search engine, he/she is more likely task-oriented. Finally, we add an additional variable to describe the extent to which visitors are immersed in the content consumption. Intuitively, whether a visitor watches the entire video clip without premature exit measures his/her interest in the video. Therefore,

$$Z_{io} = \begin{bmatrix} AverageViewTime \\ VideoInteraction \\ Replay \\ FromSearchEngine \\ WatchVideoInFull \end{bmatrix}$$

4.4.4 Heterogeneity

To model visitors' heterogeneity, we apply a random effect specification. Subscript k is the index of variable in the vectors X_{io} and Z_{io}.

$$\begin{aligned} \boldsymbol{\alpha_i} &\sim \boldsymbol{N}(\boldsymbol{\alpha^0}, \boldsymbol{\sigma_\alpha}) \\ \boldsymbol{\beta_{ik}} &\sim \boldsymbol{N}\left(\beta_k^0, \boldsymbol{\sigma_{\beta k}}\right) \\ \boldsymbol{\pi_i} &\sim \boldsymbol{N}(\boldsymbol{\pi^0}, \boldsymbol{\sigma_\pi}) \\ \boldsymbol{\eta_{ik}} &\sim \boldsymbol{N}\left(\eta_k^0, \boldsymbol{\sigma_{\eta k}}\right) \end{aligned} \tag{11}$$

4.5 ESTIMATION

4.5.1 Preprocess data

Clickstream data at hedonic wet sites usually contain many visitors who come to the site accidentally, with most of them exiting immediately. Also, many customers come to the site through reference links from third-party web sites. Most of these visitors will never come back to the site thereafter. Their browsing behaviors exhibit great heterogeneity and generate significant noise in the data. To alleviate this problem, we only choose visitors who have come to the site at least twice in our experiment data. We extract one month of clickstream data between April 1, 2009 and April 30, 2009. In total, there are 19,937 ads exposures during this period, with 212 click-throughs. The click-through rate is 1.06%. If a user either watched a single video clip for more than an hour or spent more than three hours in the current session, we consider it an outlier and remove it from out data set. We introduce four control variables—1) video popularity, 2) video length, 3) visitor gender, and 4) visitor membership length to reflect observed video and visitor heterogeneity. Table 4-2 presents the summary statistics for all dependent and independent variables used in our model. Some non-binary variables have significantly larger magnitude and make them dominate in the estimation computation. We normalize all non-binary variables. There are strong correlations between variables "NumberOfExposures", "NumberOfPages" and "NumberOfVideos". We compare the estimation results of model including all three variables and models including only one or two of the variables and conclude that the multicollinearity does not change the results significantly. We reports the results of the model using all three variables.

Table 4-2 Summary statistics for dependent and independent variables

Variables	Mean	Std Dev	Min	Med	Max
ClickOrNot	0.0106	0.0007	0	0	1
AlreadyClick	0.0161	0.0009	0	0	1
NumberOfExposures	3.5582	0.0487	0	1	101
NumberOfPages	15.5251	0.1613	0	8	329
NumberOfVideos	4.9525	0.0499	1	3	102
TimeElapsed (seconds)	5246.3022	54.5313	0	1992.5	66993
WatchTheVideoInFull	0.2999	0.0032	0	0	1
VideoInteraction	0.4145	0.0035	0	0	1
Replay	0.0061	0.0006	0	0	1
FromSearchEngine	0.0190	0.0010	0	0	1
AverageViewTime (seconds)	1598.2464	19.1425	0	478	49095
VideoLegnth	4.6764	0.0393	0	2	38
VideoPopularity	30.2720	0.3733	1	9	336
VisitorGender	0.6987	0.0036	0	1	2
VisitorMemberlength (days)	62.9165	0.3763	1	48	279

4.5.2 Estimation Methods

We use simple logistic regression as the null model. We also do logistic regression with random effects to capture visitor heterogeneity and use it as an alternative benchmark. In addition, we run our proposed hierarchical linear model with flow status as the latent variable and compare its results with the previous two benchmarks. From this point on, we will refer to these three models as “null,” “alternative,” and “proposed” models. Deviance information criterion (DIC) has been suggested as a criterion for model fit. Models with smaller DIC have better out-of-sample predictive power (Gelman, Carlin et al. (2004)).

Estimation for logistic regression can be done in many standard statistic packages. We use a Bayesian approach, specifically MCMC simulation to estimate *"proposed"* hierarchical model with latent variables and *"alternative"* logistic regression model with random effect. We conduct the MCMC simulation process through the software WINBUGS. Since we have no knowledge of parameters to estimate, we impose non-informative priors on all parameters. Therefore,

$$\alpha^0 \sim N(0,10^3),\ \beta_k^0 \sim N(0,10^3), \pi^0 \sim N(0,10^3), \eta_k^0 \sim N(0,10^3),$$

$$\sigma_\alpha \sim U(0,100),\ \sigma_{\beta k} \sim U(0,100),\ \sigma_\pi \sim U(0,100),\ \sigma_{\eta k} \sim U(0,100)$$

We build three Monte Carlo chains with different initial values. One chain assigns 0.1 to all parameters, one chain assigns -0.1 to all parameters and the other chain randomly assigns initial values between -1 to 1 . We monitor three chains to ensure the convergence of the simulation. We burn the first 50,000 draws and simulate 10,000 more draws to be used in posterior analysis. We only keep one in every ten draws from each chain to reduce memory usage. The estimation results derived from the resulting 3,000 simulations are summarized in Table 4-3.

We report the estimated mean for each simulated variable with its 2.5% and 97.5% quartiles. If the 2.5% and 97.5% quartiles have the same sign, then the corresponding variable is statistically significant with 95% confidence. The last row reports the DIC scores for the model fits.

In addition to information criteria score, hit rates of training sample and hold-out sample are reported to validate the proposed model. We keep the outcomes of visitors' click decisions during their last session as the hold-out sample and use the remaining data to obtain the parameter estimation. The posterior distributions from MCMC estimation are at the individual level. It is straightforward to generate the simulated distribution of predicted click

probability p_{io} for each person at each occasion of being forced to watch a video ad. If the distribution mean is greater than 0.5, we may predict the visitor will click the video ads. However, the click event is considerably rare and if we use probability 0.5 as the cut-off rule to predict click outcomes, all methods perform badly and it is hard to differentiate them. Another prediction rule as described in (Chatterjee, Hoffman et al. (2003)) is adopted. First we rank order the 2,019 observations in the hold-out sample in descending order of their predicted probabilities and predict the first 29 observations (the actual number of clicks) as clicks.

4.6 RESULTS

4.6.1 Model fit

Simple logistic regression has a huge DIC score and none of the variables are significant. Thus, logistic regression cannot discover any relationship between a visitor's browsing behavior and his/her intention to click an advertisement. One of the reasons for the poor performance is that visitors who come to the site bear significant heterogeneity. The extent to which certain explanatory variables affect the click-through inclination varies significantly across visitors; it is impossible for a single coefficient to pool the effects.

After we add random effects to capture visitor heterogeneity, the alternative model achieves a much lower DIC score of 275.24 and some significant variables appear. Moreover, our proposed hierarchical linear model with latent variables achieves a better DIC score at 55.09.

Table 4-4 presents the confusion matrix for click outcomes. All three methods predict 29 clicks. 65.5% predictions of the proposed model are correct, 27.6% predictions of the alternative model are correct and only 7% predictions of the null model are correct. The results demonstrate the superiority of the proposed model.

Table 4-3 Comparison of results from the three predictive models

	Variables	Simple Logistic Regression (Null model)	Logistic Regression With Random Effect Alternative model)	Hierarchical Linear Model With Latent Variable Proposed Model)
Intercept		-0.03 [-23.27, 16.42]	**-52.59** **[-66.19, -42.39]**	**-14.4** **[-20.09, -10.06]**
Within Session Variable				
	AlreadyClick	-0.16 [-21.11,21.06]	6.31 [-2.42, 13.68]	**2.2** **[0.44,3.88]**
	NumberOfExposures	-0.04 [-20.32,20.01]	1.63 [-5.24, 10.67]	0.08 [-2.30, 2.09]
	NumberOfPages	-0.27 [-21.35,19.90]	0.17 [-2.56, 2.47]	0.17 [-0.43,0.78]
	NumberOfVideos	-0.05 [-19.38, 19.48]	**-5.45** **[-13.49, -0.70]**	-1.48 [-3.77, 0.90]
	TimeElapsed	0.14 [-20.21,20.16]	-1.54 [-4.19, 0.49]	-0.32 [-1.07, 0.15]
FlowStatus				**-1.65** **[-2.93, -0.60]**
	WatchTheVideoInFull	0.06 [-20.68,21.16]	**-18.25** **[-32.64, -8.59]**	**10.38** **[2.14, 26.54]**
	VideoInteraction	0.12 [-19.51,19.10]	-1.41 [-6.91, 3.19]	-0.5 [-1.28, 0.21]
	Replay	-0.59 [-28.36,25.95]	0.85 [-9.02, 9.31]	6.56 [-0.73, 17.79]
	FromSearchEngine	-0.18 [-22.65,21.19]	-2.18 [-12.39, 5.62]	3.87 [-0.52, 10.94]
	AverageViewTime	0.37 [-19.93,20.37]	**7.1** **[4.09, 9.62]**	**-0.91** **[-1.73, -0.49]**
Control Variable				
	VideoLegnth	-0.26 [-19.72,19.12]	**-23.92** **[-30.64, -17.16]**	**-8.55** **[-13.18, -5.34]**
	VideoPopularity	-0.14 [-20.98,20.21]	0.41 [-0.76, 1.95]	-0.08 [-0.47,0.22]
	VisitorGender	-0.15 [-19.56,19.33]	-1.24 [-5.50, 3.07]	-0.36 [-1.69, 0.77]
	VisitorMemberlength	-0.1 [-19.28,18.54]	0.96 [-1.05, 3.15]	0.01 [-0.66,0.66]
DIC		>1000	275.24	55.09

Table 4-4 Confusion matrix for click outcomes

	Predicted Outcomes					
Observed Decision	Null		Alternative		Proposed	
	Click	No Click	Click	No Click	Click	No Click
Click(29)	2	27	8	21	19	10
No Click(1990)	27	1963	21	1969	10	1980

4.6.2 Model implication

In this section, we discuss specific findings in our proposed hierarchical linear model with a latent variable and compare its estimation results against the alternative model without any latent variable.

4.6.2.1 Intrasession effects

The negative intercept for click-through proneness confirms our intuition that people generally dislike commercials.

The number of advertisement exposures has a non-significant coefficient, which implies that visitors' click-through intentions do not decrease as they receive more advertisement exposures. However, Chatterjee, Hoffman et al. (2003) reported that repeated banner ads have a negative and nonlinear effect on click probability. The inconsistency may be due to the format of the advertisement. The advertisement in our program is a video clip, and the tedious effect is not as strong as in banner ads. Another explanation is that the visitors in our data set are frequent

visitors. Once they get familiar with the site advertisement, they tend to ignore it and become progressively insensitive to it (Benway (1998)).

The number of pages browsed, number of videos watched and how long the visitor has stayed in the current session are non-significant variables. The time constraint and within-site lock-in reportedly do not affect visitors' click intention (Bucklin and Sismeiro (2003)).

If a visitor has already clicked the ad in current session, he/she is more likely to click again after exposure to an advertising program when compared to those who never click in this session. This is consistent with the known conclusion related to banner ads (Chatterjee, Hoffman et al. (2003)). We may view this as an indicator of visitors' attitudes towards advertisements. Some people dislike advertisements and almost never click on any ads; others possess a more tolerant attitude and click ads occasionally.

4.6.2.2 Flow Status

The flow status has negative effect on click probability. The more interest visitors have in the video content, the less likely they will spare their attention on ads and click them. Casual visitors without predefined target content are more likely to be attracted by the advertisement. What factors can reflect visitors' flow status? Variables "WatchTheVideoInFull" (10.38), "Reply" (6.56) and "FromSearchEngine" (3.87) all have positive effect on flow status. (The latter two variables are significant at the 90% confidence level.) Watching the entire video clip, replaying the video, or accessing the video from a search engine are all signs that the visitor has considerable interest in the video and is likely to be immersed in the video content. Visitors' actions on videos are not indictors of flow status. The negative sign (-0.91) of

"AverageViewTime" is confusing. Intuitively, the flow status should be high if a visitor spends a long time watching the video. Our explanation is that, unlike other explanatory variables which are related only to the current video, “AverageViewTime” is an aggregate measure across all the current and previous videos that a visitor has watched. Since browsing behavior is very dynamic and a visitor’s flow status continually changes, this traditionally favored measure is no longer suitable as an indicator for a visitor’s instantaneous flow status.

Another noteworthy fact is the distinction between our proposed model and the alternative model on variables "WatchTheVideoInFull" and "AverageViewTime." In the alternative model, "WatchTheVideoInFull" has a direct negative effect on click probability. In the proposed model, "WatchTheVideoInFull" has a positive effect on flow status and flow status has a negative effect on click probability, and these two effects combined contribute to the overall negative effect of "WatchTheVideoInFull" on click probability. Thus, the proposed hierarchical model with latent variable reveals more fine-scale details of the click-through dynamics that simpler models may have overlooked.

4.6.2.3 Control variables

Among control variables, only "VideoLength" is significant. Long video clip has negative effect on click probability. After examining the contents of videos, we find that videos of long length are TV series or movies. We suspect that visitors are less likely to click the advertisements before TV series or movies because they are eager to watch the video as soon as possible.

4.7 MANAGERIAL SUGGESTIONS

Web site managers do not need to worry that too many video advertising programs will deter visitors from clicking ads because they are bored of video advertisement programs. Showing the advertising programs either early on, when visitors first enter the session, or later in the session does not affect the click probability. Casual visitors, or visitors in an aimless browsing mode, are more likely to click video ads. It is not effective to play an advertising program after a visitor replays a video. If a visitor comes to the site from a search engine for a particular video clip, do not bother to show him/her a video ad since the visitor is unlikely to click anyway. It is wiser to place advertising video before a gossip video clip than a movie program.

5.0 CONCLUSION AND FUTURE WORK

In this dissertation, we demonstrated how statistics-based business analytics can help identify valuable customers and understand customers' behavior. Specifically, in the application of managing customers of a non-profit organization, we adopted learning-to-rank techniques from the information retrieval discipline, and the experimental results show that we only need to contact 20% of customers to cover 80% of valuable customers. The second application modeled how likely a web site's visitors click advertising video clips. We employed a hierarchical linear model with latent variable to understand a visitor's intention to click ads. Several managerial suggestions have been proposed, including when, how often and to whom the site should deliver advertising programs.

There are several directions that we can pursue in the future. First, we can develop an online ranking system. In chapter 3, we discussed that a customer's rank changes after certain events occur. If we manage to find an approach to update a customer’s score incrementally based on features of the events, then we can keep track of a customer's scores without having to re-rank him/her frequently.

Another feasible research direction is to rank customers/visitors' value based on the posterior distribution of the random coefficients in hierarchical models. In the second application, we used Bayesian statistics to predict ad click probability. Values of coefficients with random effect are visitor-dependent, and the MCMC estimation provides us with posterior distributions at the individual level. To some extent, these individual-based coefficients reflect a user's tendency to click advertisements. For example, if the coefficient of variable "NumberOfExposure" for visitor i is bigger than visitor j, then visitor i is more sensitive to the number of ad exposures and his/her proneness to click ads decreases more drastically than visitor j. In other words, customer j is more prone to a larger number of ad exposures than customer i, and may be viewed by the web site as a more valuable visitor. Since the posterior distributions at individual level embed personalized attitude information, ranking on these posterior variables have the potential to perform better than ranking in the original feature space.

APPENDIX A

DESCRIPTIONS OF VARIABLES USED IN ADAPTED GBRANK ALGORITHM

Variables	Descriptions
Timesincelastdonation	Days since last donation
Average_Donations	Average donation amount
Number_Contacts	Number of contacts that ARC has made since last donation
Donation_Frequency	Average interval, in days, between two consecutive donations
Timesincelastcontact	Days since last donation
Race	Percentage of white people
Gender	Percentage of male
Contact_Frequency	Average interval, in days, between two consecutive contacts
Number_Donations	Total number of donations
Education	Percentage of population who have at least college degree
Church	An index measuring how active people go to church
Income	Average gross income of households with same zip code , in thousands
Age	Average age of the county population

APPENDIX B

DESCRIPTIONS OF VARIABLES USED IN HIERARCHICAL LINEAR MODEL WITH HIDDEN VARIABLES

Variables	Descriptions
AlreadyClick	Dummy variable set equal to 1 if the visitor already clicked video ads at least once during current session
NumberOfExposures	Number of video ads the visitor has watched during current session
NumberOfPages	Number of pages the visitor has browsed during current session
NumberOfVideos	Number of vidoes the visitor has watched during current session
TimeElapsed	Time, in seconds, since the visitor came the web site
WatchTheVideoInFull	Dummy variable set equal to 1 if the visitor finished watching the entire video clip without premature exit
VideoInteraction	Dummy variable set equal to 1 if the visitor has any interactive action with the video, including pause, resume, replay
Replay	Dummy variable set equal to 1 if the current video is a replay
FromSearchEngine	Dummy variable set equal to 1 if the visitor came to the site from a hyperlink provided in a search engine such as Google
AverageViewTime	Average duration, in seconds, of video view
VideoLegnth	Number of parts that the video contains, each part is about five minute video clip
VideoPopularity	Number of views of current video
VisitorGender	Dummy variable set equal to 1 if the visitor is male and 2 if the visitor choose not to report
VisitorMemberlength	Number of days since the visitor registered at the web site

BIBLIOGRAPHY

Benway, J. P. (1998). Banner Blindness: The Irony of Attention Grabbing on the World Wide Web. Proceedings of the Human Factors and Ergonomics Society Annual Meeting October

Briggs, R. and N. Hollis (1997). "Advertising on the Web: Is There Response before Click-Through?" Journal of Advertising Research **37**(2): 33-45.

Bucklin, R. E., J. M. Lattin, A. Ansari, S. Gupta, D. Bell, E. Coupey, J. D. C. Little, C. Mela, A. Montgomery and J. Steckel (2002). "Choice and the Internet: From Clickstream to Research Stream." Marketing Letters **13**(3): 245-258.

Bucklin, R. E. and C. Sismeiro (2003). "A Model of Web Site Browsing Behavior Estimated on Clickstream Data." Journal of Marketing Research **XL**: 249-267.

Bucklin, R. E. and C. Sismeiro (2003). "A model of web site browsing behavior estimated on clickstream data." Journal of Marketing Research **40**(3): 249-267.

Burges, C., T. Shaked, E. Renshaw, M. Deeds, N. Hamilton and G. Hullender (2005). Learning to rank using gradient descent. Proceedings of the 22nd International Conference on Machine Learning (ICML 2005).

Burges, C. J. C., R. Ragno and Q. V. Le (2006). Learning to rank with nonsmooth cost functions. Advances in Neural Information Processing Systems (NIPS 2006).

Chapelle, O. and M. Wu (2010). "Gradient descent optimization of smoothed information retrieval metrics " Information Retrieval **13**(3): 216-235.

Chatterjee, P., D. L. Hoffman and T. P. Novak (2003). "Modeling the Clickstream: Implications for Web-Based Advertising Efforts." Marketing Science **22**(4): 520-541.

Childers, T. L., C. L. Carr, J. Peck and S. Carson (2001). "Hedonic and Utilitarian Motivations for Online Retail Shopping Behavior." Journal of Retailing **77**: 511-535.

Clark, D. and S. Vranica (Sept. 16, 2009). Adobe to Acquire Omniture in $1.8 Billion Deal. Wall Street Journal.

Cohen, W. W., R. E. Schapire and Y. Singer (1998). "Learning to Order Things." Advances in Neural Information Processing Systems **10**: 243-270.

Crammer, K. and Y. Singer (2001). Pranking with Ranking. Advances in Neural Information Processing Systems

Csíkszentmihályi, M. (1975). Beyond Boredom and Anxiety: Experiencing Flow in Work and Play, San Francisco: Jossey-Bass.

Cui, D. P. and D. Curry (2005). "Prediction in marketing using the support vector machine." Marketing Science **24**(4): 595-615.

Cui, G., M. L. Wong and H. K. Lui (2006). "Machine learning for direct marketing response models: Bayesian networks with evolutionary programming." Management Science **52**(4): 597-612.

Duh, K. K. and K. Kirchhoff (2008). Learning to Rank with Partially-Labeled Data. SIGIR.

eMarketer (2002). Essential e-business Numbers for Marketers. New York.

Freund, Y., R. Iyer, R. E. Schapire and Y. Singer (2003). "An efficient boosting algorithm for combining preferences." The Journal of Machine Learning Research **4**(12): 933-969.

Freund, Y. and R. E. Schapire (1997). "A decision-theoretic generalization of on-line learning and an application to boosting." Journal of Computer and System Sciences **55**(1): 119-139.

Friedman, J. H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine." Annual Statistics **29**(5): 1189-1232.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin (2004). Bayesian Data Analysis.

Henneges, C., G. Hinselmann, S. Jung, J. Madlung, W. Schütz, A. Nordheim and A. Zell (2009). "Ranking Methods for the Prediction of Frequent Top Scoring Peptides from Proteomics Data." Journal of Proteomics & Bioinformatics **2**: 226-235.

Herbrich, R., T. Graepel and K. Obermayer (2000). Large Margin Rank Boundaries for Ordinal Regression. Advances in Large Margin Classifiers, MIT Press**:** 115-132.

Hoffman, D. L. and T. P. Novak (1996). "Marketing in Hypermedia Computer-Mediated Environments:Conceptual Foundations." Journal of Marketing **60**: 50-68.

Holbrook, M. B. and M. P. Gardner (1993). "An Approach to Investigating the Emotional Determinants of Consumption Durations: Why Do People Consumer What They Consume for as long as They Consumer It?" Journal of Consumer Psychology **2**(2): 123-142.

Huberman, B. A., P. L. T. Pirolli, J. E. Pitkow and R. M. Lukose (1998). "Strong Regularities in World Wide Web Surfing." Science **280**(3): 95-97.

Janiszewski, C. (1998). "The Influence of Display Characteristics on Visual Exploratory Search Behavior." Journal of Consumer Research **290-301**(25): 290-301.

Johnson, E. J., S. Bellman and G. L. Lohse (2003). "Cognitive Lock-In and the Power Law of Practice." Journal of Marketing **67**: 62-75.

Johnson, E. J., W. W. Moe, P. S. Fader, S. Bellman and G. L. Lohse (2004). "On the Depth and Dynamics of Online Search Behavior." Management Science **50**(3): 299-308.

Kim, Y., W. N. Street, G. J. Russell and F. Menczer (2005). "Customer Targeting: A Neural Network Approach Guided By Genetic Algorithms." Management Science **51**(2): 264-276.

King, E. N. and T. P. Ryan (2002). "A Preliminary Investigation Of Maximum Likelihood Logistic Regression Versus Exact Logistic Regression." American Statistician **56**(3): 163-170.

Li, P., C. Burges and Q. Wu (2007). McRank: Learning to Rank Using Multiple Classifications and Gradient Boosting,. Neural Information Processing Systems (NIPS).

Liu, T.-Y. (2011). Learning to Rank for Information Retrieval, Springer.

MacMillan, D. (2009). What Works in Online Video Advertising? BusinessWeek.

Moe, W. W. (2003). "Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstream." Journal of Consumer Psychology **13**(1): 29-39.

Moe, W. W. (2006). "An Empirical Two-Stage Choice Model with Varying Decision Rules Applied to Internet Clickstream Data." Journal of Marketing Research **XLIII**(680-692).

Moe, W. W. (2006). "A Field Experiment to Assess the Interruption Effect of Pop-Up Promotions." Journal of Interactive Marketing **20**(1): 34-44.

Moe, W. W. and P. S. Fader (2004). "Capturing Evolving Visit Behavior in Clickstream Data." Journal of Interactive Marketing **18**(1): 5-19.

Moe, W. W. and P. S. Fader (2004). "Dynamic Conversion Behavior at E-Commerce Site's." Management Science **50**(3): 326-335.

Montgomery, A. L., S. Li, K. Srinivasan and J. C. Liechty (2004). "Modeling Online Browsing and Path Analysis Using Clickstream Data." Marketing Science **23**(4): 579-595.

Novak, T. P., D. L. Hoffman and Y.-F. Yung (2000). "Measuring the Customer Experience in Online Environments: A Structural Modeling Approach " Marketing Science **19**(1): 22-42.

Novak, T. P., D. L. Hoffman and Y.-F. Yung (2000). "Measuring the Customer Experience in Online Environments: A Structural Modeling Approach." Management Science **19**(1): 22-42.

Olney, T. J., M. B. Holbrook and R. Batra (1991). "Consumer Responses to Advertising: The Effects of Ad Content, Emotions, and Attitude toward the Ad on Viewing Time." Journal of Consumer Research **17**(4): 440-453.

Park, Y.-H. and P. S. Fader (2004). "Modeling Browsing Behavior at Multiple Websites." Marketing Science **23**(3): 280-303.

Rigutini, L., T. Papini, M. Maggini and F. Scarselli (2011). "SortNet: Learning To Rank By a Neural-Based Sorting Algorithm." IEEE TRANSACTIONS ON NEURAL NETWORKS **22**(9).

Rossi, P. E. and G. M. Allenby (2003). "Bayesian Statistics and Marketing." Marketing Science **22**(3): 304-328.

Schaupp, L. C. and F. Bélanger (2005). "A Conjoint Analysis of Online Consumer Satisfaction." Journal of Electronic Commerce Research **6**(2).

Scott, S. L. and I.-H. Hann (2007). A Nested Hidden Markov Model for Internet Browsing Behavior.

Sculley, D. (2010). Combined Regression and Ranking. KDD'10, Washington, DC, USA.

Sismeiro, C. and R. E. Bucklin (2004). "Modeling purchase behavior at an E-commerce web site: A task-completion approach." Journal of Marketing Research **41**(3): 306-323.

Taylor, M., J. Guiver, S. Robertson and T. Minka (2008). Softrank: optimising non-smooth rank metrics. Proceedings of the 1st International Conference on Web Search and Web Data Mining**:** 77-86.

Trevino, L. K. and J. Webster (1992). "Flow in Computer-Mediated Communication " Communication Research **20**(2): 249-276.

Tsai, M.-f., T.-Y. Liu, T. Qin, H.-H. Chen and W.-Y. Ma (2007). FRank: A Ranking Method with Fidelity Loss. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007).

Vascellaro, J. E. (Apr. 22 2010). Facebook Wants to Know More Than Just Who Your Friends Are. Wall Street Journal.

Yue, Y., T. Finley, F. Radlinski and T. Joachims (2007). A support vector method for optimizing average precision. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007).

Zheng, Z., K. Chen, G. Sun and H. Zha (2007). A regression framework for learning ranking functions using relative relevance judgments. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007).