L4: Outline: CoAd Lectures

- Introduction

L1 - Online advertising background
- Business models, Campaigns

L2 - Technology and Economics
- Forward Markets
  - Gradient Descent, Operations research, LP, QP
- Auction Theory and Game Theory

L3 - Spot Markets
- ML, Ad quality, Ranking, Budgeting

L4 - New Directions
- Challenges in online advertising
- Summary

<table>
<thead>
<tr>
<th>CoAd Lectures</th>
<th>Date</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friday</td>
<td>9/11/2009</td>
<td>10:30-12:00</td>
</tr>
<tr>
<td>Saturday</td>
<td>9/12/2009</td>
<td>8:30-10:00</td>
</tr>
<tr>
<td>Sunday</td>
<td>9/13/2009</td>
<td>8:30-10:00</td>
</tr>
<tr>
<td>Monday</td>
<td>9/14/2009</td>
<td>8:30-10:00</td>
</tr>
</tbody>
</table>

RuSSIR 2009, Petrozavodsk, Russia. Online Advertising © 2009 James G. Shanahan (San Francisco)

James.Shanahan_AT_gmail_DOT_com
Course philosophy

• Socratic Method (more inspiration than information)
  – participation strongly encouraged (please state your name and affiliation)

• Highly interactive and adaptable
  – Questions welcome!!

• Lectures emphasize intuition, less rigor and detail
  – Build on lectures from other faculty
  – Background reading will provide more rigor & detail

• Action Items
  – Read suggested books first (and then papers), read/write Wikipedia, watch/make YouTube videos, take courses, participate in competitions, do internships, network
  – Prototype, simulate, publish, participate
  – Classic (core) versus trendy (applications)
Lecture 2: Homework

- **Email solutions to**
  James.Shanahan__AT__gmail.com

- **Exercises**
  - Find a local minimum of the function
    \[ f(x) = 6x^5 - 8x^2 + 6 \]
  - Implement gradient descent version of Perceptron
  - Implement gradient descent version of OLS; show evolution of weight vector during training
Lecture 3: Homework

• Email solutions to James.Shanahan_AT_gmail.com

• Exercises
  – What is the mixed strategy equilibrium for Rocks Scissors Paper game?
  – What are the three equilibria for the following game?

<table>
<thead>
<tr>
<th>Player 1</th>
<th>Player 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Up</td>
</tr>
<tr>
<td>Up</td>
<td>(2, 2)</td>
</tr>
<tr>
<td>Down</td>
<td>(0, 0)</td>
</tr>
</tbody>
</table>
Outline

• Introduction
• Online advertising background
• Business models
• Creating an online ad campaign
• Technology and Economics
  – Advertisers (optimizing ROI thru ads and ad placement)
  – Publishers (optimizing revenue and consumer satisfaction)
    • Forward Markets
    • Auction Systems, Game Theory
    • Spot Markets
      – Background
      – Ad Quality
      – Budgeting
• New Directions
• Challenges in online advertising
• Summary
Ad Quality Outline

- Motivation
- Machine Learning Overview
- Metrics
- IR as a means of Ranking
- Learning to Rank (LETOR)
- Online Learning
- Open issues
Ad Network Architecture: Spot Market

Featurize Ads and LPs

- Ads, LPs
- Crawler
- Index
- TF-IDF
- (Webgraph)
- Anchor Text
- Classes
- Page Quality
- Historical
- Site-level
- Ratings

Rank Ads

- For a target page and user
- (ECPM)
- Compute price
- (CPX)

Featurize Target Page And User

- Query Price
- Analytics
- Logs
- WebPage Index
- Crawl+

Yield Management Ad Network

- Ad upload/SelfServe
- Creatives Constraints
- Generate AdCode
- ADashBoard
- PDashBoard
- Advertisers
- Publishers
- Users
- WebPage
- SERP
- WWW
- Analytics
- Ads, LPs
- Logs
- Events
- Historical Site-level
- Page Quality
- Classes
- Anchor Text
- TF-IDF
- (Webgraph)
- Index

Historical Site-level WebPage

RuSSIR 2009, Petrozavodsk, Russia. Online Advertising © 2009 James G. Shanahan (San Francisco) James.Shanahan_AT_gmail_DOT_com
Rank by ECPM; CPC Calculation

Payoff = Value – Price
Payoff = ValuePerClick – CPC
ECPM = Bid \times \text{QualityScore}

\[
ECPM_{Ad} = \text{Quality}_{Ad} \times Bid_{Ad} \times 1000
\]

For \( ad_i \) to maintain its current rank then \( Bid_i \) needs to be at least:

\[
Bid_1 \geq \frac{Bid_2 \times DQ_2}{DQ_1}
\]

**CPC for click at rank 1**

<table>
<thead>
<tr>
<th>Ad Id</th>
<th>Bid</th>
<th>Quality</th>
<th>Rank</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>123</td>
<td>$5.80</td>
<td>10</td>
<td>$58.00</td>
<td>$1.71</td>
</tr>
<tr>
<td>ABC</td>
<td>$4.25</td>
<td>4</td>
<td>$17.00</td>
<td>$3.01</td>
</tr>
<tr>
<td>NOP</td>
<td>$2.00</td>
<td>6</td>
<td>$12.00</td>
<td>$0.51</td>
</tr>
<tr>
<td>TUV</td>
<td>$3.00</td>
<td>1</td>
<td>$3.00</td>
<td>$1.66</td>
</tr>
<tr>
<td>XYZ</td>
<td>$0.55</td>
<td>3</td>
<td>$1.65</td>
<td>Reserve Bid</td>
</tr>
</tbody>
</table>
Online Ad Targeting Systems (OATS)

- **OAT systems has many similarities with “Direct Mail” systems**
  - Can borrow a lot from these experiences and ideas
  - Many studies, e.g.,
    - Data Mining for CRM
      - See [http://www.dmreview.com/article_sub.cfm?articleId=1046025](http://www.dmreview.com/article_sub.cfm?articleId=1046025)
  - Recently many OATS.....
Accurate CTR Estimates are Crucial

\[ ECPM_{Ad} = CTR_{Ad} \times Bid_{Ad} \times 1000 \]

• Very important to have accurate estimates of CTR<sub>Ad</sub> for a keyword or publisher page
  – for ranking and for revenue purposes

• E.g., A true CTR for an Ad is 2.6% must be shown 1,000 times before we are 95% confident that this estimate is within 1% of the true CTR

• Curiously, average CTR and CPC
  – 2.6% CTR for ads in sponsored search advertising
  – Average CPC (cost-per-click) on Google was $1.60
  – [MarketingSherpa, 9/2005]
Estimating CTR (and later AR)

Estimate using Binomial MLE Estimates
I.e., #Clicks/#Impression

$40/1,000 @CPC of $1.60
$400/10,000

95% Confident

1.6% 2.6% 3.6% CTR (after 1,000 impressions)

2.3% 2.9% (after 10,000 impressions)
Estimating CTR (and later AR)

For a network of
~10^9 target pages,
~10^6 ads
~10^7 users

- Cannot afford this evaluation/auditioning
- Borrow strength, marginalize
- CoD (curse of dimality)

1.6% 2.6% 3.6% CTR (after 1,000 impressions)
2.3% 2.9% (after 10,000 impressions)
Accurate CTR Estimates are Crucial

\[ ECPM_{Ad} = CTR_{Ad} \times Bid_{Ad} \times 1000 \]

- Very important to have accurate estimates of \( CTR_{Ad} \) for a keyword or publisher page
  - for ranking and for revenue purposes
  - CTR drop exponentially with position [enquiro.com]; NDCG Metric

- E.g., A true CTR for an Ad is 2.6% must be shown 1,000 times before we are 95% confident that this estimate is within 1% of the true CTR, i.e., \([1.6, 3.6]\)
  - Very noisy!!
• Motivation

• Machine Learning Overview
  – Background
  – Gradient Descent
  – From Perceptrons to SVMs
  – Probabilistic Models
  – Decision Trees
  – Bagging/Boosting

• Metrics

• IR as a means of Ranking

• Learning to Rank (LETOR)

• Online Learning

• Open issues
Machine Learning is about…

- **Models**
- **Loss functions**
  - Maximum likelihood, mean squared errors, accuracy
- **Optimization**
  - gradient descent, Newton, linear/quadratic programming, etc.
- **Data**
  - Training data, Validation data, Test data, crossfold validation
- **Metrics**
  - Precision, recall, DCG, Revenue etc.
- **And more!**
Machine Learning Background

Machine Learning (ML): ”a computer program that improves its performance at some task through experience” [Mitchell 1997]

GIVEN: Input data is a table of attribute values and associated class values (in the case of supervised learning)
GOAL: Approximate \( f(x_1, \ldots, x_n) \rightarrow y \)

<table>
<thead>
<tr>
<th>Instance</th>
<th>Attr</th>
<th>( x_1 )</th>
<th>( \ldots )</th>
<th>( x_n )</th>
<th>( y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>1</td>
<td>\ldots</td>
<td>-1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0</td>
<td>\ldots</td>
<td>+1</td>
<td>4</td>
</tr>
<tr>
<td>\ldots</td>
<td></td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>L</td>
<td></td>
<td>0</td>
<td>4</td>
<td>8</td>
<td>-1</td>
</tr>
</tbody>
</table>

What is Machine Learning?
• Automatically programming computers to perform tasks that humans perform well but find difficult to specify algorithmically
• Principled way of building high performance information processing systems
  – search engines, information retrieval
  – Natural language processing
  – Finance
  – Engineering
Data Modeling Process

1. Understand the domain and Define problems
2. Collect and Preprocess Data
3. Extract Patterns/Models
4. Interpret and Evaluate discovered knowledge
5. Deploy System in the wild (and AB test)

Data Modeling is inherently interactive and iterative.
Linear Regression

Predict systolic blood pressure (SBP) from Age using 33 adult women examples

\[ SBP = 81.54 + 1.222 \cdot \text{Age} \]

Types Machine Learning

• **Design of a learning algorithm is affected by**
  – Task to be learned?
  – What feedback is available from?
  – What representation is used for the learnt model?

• **Type of feedback:**
  – **Supervised learning**: correct answers for each example
  – **Unsupervised learning**: correct answers not given
  – **Reinforcement learning**: occasional rewards
  – **Co-learning, Transductive Learning, etc.**
Families of Supervised Learning

• **Generative Classifier (Bottom-up learning)**
  – Build model of each class
  – Assume the underlying form of the classes and estimate their parameters (e.g., a Gaussian)

• **Discriminative Classifier (Top down)**
  – Build model of boundary between classes
  – Assume the underlying form of the discriminant and estimate its parameters (e.g., a hyperplane)
Generative vs. Discriminative

• Discriminative learning (e.g., ANN, SVM, logistic regression, Conditional Random Fields (CRF)) typically more accurate
  – Better with small datasets
  – Faster to train

• Generative learning (e.g., Bayesian Networks, HMM, Naïve Bayes, Mixtures of Gaussians) typically more flexible
  – More complex problems
  – More flexible predictions
Ad Quality Outline

• Motivation
• Machine Learning Overview
  – Background
  – Gradient Descent (covered in Lecture 2)
  – From Perceptrons to SVMs (covered partially in lecture and in notes)
  – Probabilistic Models (in notes and see other lectures)
  – Decision Trees
  – Bagging/Boosting
• IR as a means of Ranking
• Learning to Rank (LETOR)
• Online Learning
• Open issues
Ad Quality Outline

- **Motivation**
- **Machine Learning Overview**
  - Background
  - Gradient Descent
  - From Perceptrons to SVMs
  - Probabilistic Models (in notes and see other lectures)
  - Decision Trees
  - Bagging/Boosting
- **IR as a means of Ranking**
- **Learning to Rank (LETOR)**
- **Online Learning**
- **Open issues**
Maximum Margin Classifier: take 3

• The classifier that produces the maximum margin (over the training data)
• I.e., the hyperplane that is furthest from the data

Better!
But the objective is non-convex

\[
\begin{align*}
\text{Maximize} & \quad \frac{\gamma_{\text{fun}}}{\|W\|} \\
\text{subject to} & \quad y_i (w \cdot x_i + b) \geq \gamma_{\text{fun}} \quad \forall \quad i = 1, \ldots, L
\end{align*}
\]

Versus

Best!
Convex objective

\[
\begin{align*}
\text{Maximize} & \quad \frac{1}{\|W\|^2} \\
\text{subject to} & \quad y_i (w \cdot x_i + b) \geq 1 \quad \forall \quad i = 1, \ldots, L
\end{align*}
\]

\[\gamma_{\text{Fun}} = 1\]

Scaling constraint: \(\gamma_{\text{Fun}}\) of \((W, b)\) WRT \(S\) is set to 1
SVMs find the center of the largest radius hypersphere whose center can be placed in version space and whose surface does not intersect with the hyperplanes corresponding to the labeled instances.
SVM

\[ \langle \hat{W}, \hat{X} \rangle + b = 0 \]

\[ \min \frac{\|w\|^2}{2} \]

\[ \max \frac{2}{\|w\|^2} \]

\[ \max L(\alpha) = \max \alpha^T \alpha - \frac{1}{2} \alpha^T Q \alpha \]

Find widest highway through the data

RuSSIR 2009, Petrozavodsk, Russia. Online Advertising © 2009 James G. Shanahan (San Francisco)
James.Shanahan_AT_gmail_DOT_com
Topic = Product Watch Filter for K2

\[ H_1: \langle \vec{W}, \vec{X} \rangle + b = 1 \]
\[ \langle \vec{W}, \vec{X} \rangle + b = 0 \]
\[ H_2: \langle \vec{W}, \vec{X} \rangle + b = -1 \]

Where \( \vec{W} \) denote the Support Vectors and have a non-zero alpha weight

\( x_1 = \text{Dublin} \)

\( x_2 = \text{Bloomsdale} \)
SVM Learning Algorithms

Primal (W,b) + Constraints

Encode as a Langrangian

Primal Langrangian(W,b,α)

Substitute α for W,B

Dual Langrangian(α) + (simpler) Constraints

Solve for α using either approach

Gradient Ascent (Dual Langrangian(α)) + KKT

Primal (Dual Langrangian(α, λ))

Use α to recover W,b (in linear case only!!)

SVM
SVM Learning Algorithms

• Gradient Descent
  – Perceptron Ensemble
  – Perceptron
  – Kernel-Adatron
  – PAUM

• Naïve Gradient Descent in dual mode

• Population-based search
  – E.g., Evolutionary algorithms

• Quadratic/linear programming based (QP/LP) Algorithms
  – Interior point algs; Fixed or variable sized Chunking

• Non-QP-based, with closed form
  – SMO, SMOK1, SMOK2
Gradient Ascent Method

\[
\max_{\alpha} W(\alpha) = \max \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{j=1}^{l} y_i y_j \langle X_i, X_j \rangle \alpha_i \alpha_j ,
\]

Given Training set S and learning rate, and \( \alpha = 0 \)

Repeat

for all train set \( i = 1 \) to \( l \)

update

\[
\alpha_i \leftarrow \alpha_i + \frac{\omega}{K(x_i, x_i)} \left( 1 - y_i \sum_{j=1}^{l} \alpha_j y_j K(x_i, x_j) \right)
\]

where \( \omega \in (0,2) \)

Ensure \( \alpha \) are within range \([0, C]\)

if \( \alpha < 0 \), then \( \alpha \leftarrow 0 \)

if \( \alpha > C \), then \( \alpha \leftarrow C \)

End for

Until stop criterion satisfied (KKT conditions)

Return \( \alpha \)
Learning Algorithms

SVM Learning Strategy

- **Non-QP-based are more desirable**
  - More efficient in terms of CPU and memory
  - Easier to program
  - Less intellectual property

- **Implemented non-QP-based algorithms**
  - Kernel-Adatron
  - SMO (patented by Microsoft)
  - Two modifications SMOK1, SMOK2 (IP free)
SVMs Basic Ideas: Learning

Trainig Set

SVM Parameter

Kernel Computation (Doc Similarity matrix)

Find Hyperplane in dual space

Convert Hyperplane to primal space

Hyperplane (W, b)

SVM Filter

<table>
<thead>
<tr>
<th>Term</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>T₁</td>
<td>Wgt₁</td>
</tr>
<tr>
<td>T₂</td>
<td>Wgt₂</td>
</tr>
<tr>
<td>....</td>
<td>....</td>
</tr>
</tbody>
</table>

Convert Hyperplane to primal space

Find Hyperplane in dual space

Hyperplane (W, b)

SVM Bayes Perceptron

Mac

Bayes Po
Conclusion

• Support vector machines (SVM) are a general purpose suite of machine learning algorithms for classification and regression. Generic support vector machines (SVMs) provide excellent performance on a variety of learning problems ranging from text to genomic data.
• There has been an explosion of new learning algorithms, kernels and applications
• Tuning SVMs remains a black art: selecting a specific kernel and parameters is usually done in a try-and-see manner.
• Many public domain packages available on the WWW
• Building kernels is a very hot of research area

• Anyone involved in text processing should consider using support vector machines but astutely
library(e1071)

iris.svm <- svm(Species ~ ., data = iris, probability = TRUE)

plot(iris.svm, iris, Petal.Width ~ Petal.Length, slice = list(Sepal.Width = 3, Sepal.Length = 4))
pred <- predict(iris.svm, iris, probability = TRUE)

attr(pred, "prob") # to get the probabilities
Software

- A list of SVM implementations can be found at http://www.kernel-machines.org/software.html

- LIBSVM (Java, and C++) can handle multi-class classification
  - http://www.csie.ntu.edu.tw/~cjlin/libsvm/

- Weka is open source machine learning toolkit (Java-based) available from
  - http://www.cs.waikato.ac.nz/~ml/weka/

- SVMLight is among one of the earliest implementations of SVMs (written in C)

- Several high level toolboxes for SVMs are also available
  - Matlab, R
Background Material

- **Background information about SVMs can be found in:**
  - SVMs in R
    - [http://www.jstatsoft.org/v15/i09/paper](http://www.jstatsoft.org/v15/i09/paper)
  - [Burges' tutorial](http://svm.research.bell-labs.com/SVMdoc.html)
  - Publication list at [kernel-machines.org](http://kernel-machines.org)
  - References in *An Introduction to Support Vector Machines*.
Potential Applications of SVMs

- Page Quality, Page Category, Webspam, LETOR
Ad Quality Outline

• Motivation
• Machine Learning Overview
  – Background
  – Gradient Descent
  – From Perceptrons to SVMs
  – Probabilistic Models
    – Decision Trees
    – Bagging/Boosting
• IR as a means of Ranking
• Learning to Rank (LETOR)
• Online Learning
• Open issues
Fruit Classifier: CART DT

FIGURE 8.2. A tree with arbitrary branching factor at different nodes can always be represented by a functionally equivalent binary tree—that is, one having branching factor $B = 2$ throughout, as shown here. By convention the “yes” branch is on the left, the “no” branch on the right. This binary tree contains the same information and implements the same classification as that in Fig. 8.1. From: Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification. Copyright © 2001 by John Wiley & Sons, Inc.
FIGURE 8.3. Monothetic decision trees create decision boundaries with portions perpendicular to the feature axes. The decision regions are marked $\mathcal{R}_1$ and $\mathcal{R}_2$ in these two-dimensional and three-dimensional two-category examples. With a sufficiently large tree, any decision boundary can be approximated arbitrarily well in this way. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.
Variable split? And node prediction?

• **For Classification**
  – Select variable split using Impurity measures and reduction impurity
    • Entropy, GINI, Misclassification Error
  – Node prediction: dominant class, or conditional probability of dominant class

• **For Regression**
  – Select variable split using squared prediction error
  – Node prediction: mean or median value of the response values of the training examples that make it to that node
Using information theory

- To implement Choose-Attribute in the DTL algorithm
- Information Content (Entropy):
  \[ I(P(v_1), \ldots, P(v_n)) = \sum_{i=1}^{n} -P(v_i) \log_2 P(v_i) \]
- For a training set containing \( p \) positive examples and \( n \) negative examples:
  \[
  I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n}
  \]

[Adapted from Russell and Norvig 1995]
Information gain

• A chosen attribute $A$ divides the training set $E$ into subsets $E_1, \ldots, E_v$ according to their values for $A$, where $A$ has $v$ distinct values.

\[
\text{remainder}(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I\left(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}\right)
\]

• Information Gain (IG) or reduction in entropy from the attribute test:

\[
IG(A) = I\left(\frac{p}{p + n}, \frac{n}{p + n}\right) - \text{remainder}(A)
\]

• Choose the attribute with the largest IG
Node Impurity for Binary Class

Impurity measures for 2-classes

- Entropy
- Misclassification
- GINI

$P(x=c_i|Node)$
Node Impurity for Binary Class

1. #
2. #plot binary impurity measures
3. # plotImpurity()
4. plotImpurity = function () {
   1. x=seq(0.00001, 0.99999, by=0.01)
   2. y =numeric(length(x))
   3. for (i in 1:length(x)) y[i] = -x[i]*log(x[i],2)-(1-x[i])*log(1-x[i],2)
   4. plot(x,y, type="l", lty=1, main="Impurity measures for 2-classes", xlab="P(x=c_i|Node)", ylab="Impurity(P)")
5. #plot GINI impurity measure
6. lines(x, x*(1-x), col="red", lty=5, type="l")
7. #plot misclassification error
8. x1=c(0, 0.5, 1)
9. y1=c(0, 0.5, 0)
10. lines(x1, y1, col="green", lwd=3, type="l", lty="dashed")
11. #some labels
12. text(0.3,1.0,"Entropy", col="black")
13. text(0.48,0.2,"GINI", col="red")
14. text(0.5,0.55,"Misclassification", col="green")
15. grid() #make it easy to read!
}
Decision tree learning

- **Aim:** find a small tree consistent with the training examples
- **Idea:** (recursively) choose "most significant" attribute as root of (sub)tree
- **Algorithms**
  - CART, ID3, C4.5, C5.0

```plaintext
function DTL(examples, attributes, default) returns a decision tree

    if examples is empty then return default
    else if all examples have the same classification then return the classification
    else if attributes is empty then return MODE(examples)
    else
        best ← CHOOSE-ATTRIBUTE(attributes, examples)
        tree ← a new decision tree with root test best
        for each value vi of best do
            examplesi ← {elements of examples with best = vi}
            subtree ← DTL(examplesi, attributes − best, MODE(examples))
            add a branch to tree with label vi and subtree subtree
        return tree

[Adapted from Russell and Norvig 1995]
```
Decision Trees: Strengths

• Decision trees are able to generate understandable rules (sometimes).
• Can be cheap to compute and in memory footprint.
• Decision trees are able to handle both continuous and categorical variables.
• Missing data is handled naturally
• Variable importance
  – Decision trees provide a clear indication of which features are most important for prediction or classification.

• Software: in R, use rpart, tree, gbm packages
Decision Trees: Weaknesses

• Decision tree can be computationally expensive to train.
  – Need to compare all possible splits
  – Pruning is also expensive

• Most decision-tree algorithms only examine a single field at a time.
  – This leads to rectangular classification boxes that may not correspond well with the actual distribution of records in the decision space.
  – Each node could be a linear regression of all variables
Ad Quality Outline

• Motivation
• Machine Learning Overview
  – Background
  – Gradient Descent
  – From Perceptrons to SVMs
  – Probabilistic Models
  – Decision Trees
  – Bagging/Boosting
• IR as a means of Ranking
• Learning to Rank (LETOR)
• Online Learning
• Open issues
Ensemble Models

An ensemble model is composed of multiple individual (weak) models that are combined in some way (e.g., by voting) to classify new examples.

Popular approaches include bagging [Breiman 1996] and boosting [Schapire 1990].
Bagging (bootstrap aggregation)

TRAINING

• Let m be the number of instances in the training data
• For each of t iterations:
  – Sample m instances with replacement from original training dataset
  – Apply the learning algorithm to the sample
  – Store the resulting model in Models[t]

CLASSIFICATION

• For each of the t models:
  – Predict class of the test instance using Models[t]
• Return class that has been predicted most often
Here, Doc has been classified as + by Filter1, and as − by Filter2, and Filter3. Since the positive class has the majority, Doc is classified as + using this ensemble.

NOTE: Here weight Wgt_+ and Wgt_- are uniform.
Boosting

TRAINING \( (m = \text{number of instances in the training data}) \)

- Assign equal weight to each training instance
- For each of \( t \) iterations:
  - Sample \( m \) instances with replacement from weighted training dataset
  - Apply the learning algorithm to the sample
  - Compute error \( e \) of model on training dataset
  - If \( e = 0 \) or \( e > 0.5 \) Terminate
  - Foreach instance in training dataset:
    - If instance classified correctly by model then
      - Multiply weight of instance by \( e/(1-e) \)
    - Normalise weights of all instances
  - Set weight of model to \(-\log(e/(1-e))\)
  - Store the resulting model in Models\([t]\) and weight
Boosting

CLASSIFICATION

• Assign weight of zero to all classes
• For each of the t models:
  – Predict class of the test instance using Models[t]
  – Add Weight[t] to weight of class predicted by model[t]
• Return class with highest weight
A Boosting based Filter

NOTE: Here weight $Wgt_+$ and $Wgt_-$ are **not** uniform
Gradient Boosted Decision Trees

**Algorithm 2: Stochastic Gradient TreeBoost**

1. $F_0(x) = \arg \min_\gamma \sum_{i=1}^{N} \Psi (y_i, \gamma)$
2. For $m = 1$ to $M$ do:
   3. $\{\pi(i)\}_1^N = \text{rand\_perm} \{i\}_1^N$
   4. $\tilde{y}_{\pi(i)m} = - \left[ \frac{\partial \Psi(y_{\pi(i)}, F(x_{\pi(i)}))}{\partial F(x_{\pi(i)})} \right] F(x) = F_{m-1}(x), \ i = 1, \tilde{N}$
   5. $\{R_{lm}\}_1^L = L - \text{terminal node tree} \left( \{\tilde{y}_{\pi(i)m}, x_{\pi(i)}\}_1^{\tilde{N}} \right)$
   6. $\gamma_{lm} = \arg \min_\gamma \sum_{x_{\pi(i)} \in R_{lm}} \Psi \left( y_{\pi(i)}, F_{m-1}(x_{\pi(i)}) + \gamma \right)$
   7. $F_m(x) = F_{m-1}(x) + \nu \cdot \gamma_{lm} 1(x \in R_{lm})$
8. endFor

[Friedman 1999]
Gradient Boosted Decision Trees are an additive classification or regression model consisting of an ensemble of trees, fitted to current residuals, gradients of the loss function, in a forward step-wise manner. In the traditional boosting framework, the weak learners are generally shallow decision trees consisting of a few leaf nodes. GBDT ensembles are found to work well when there are hundreds of such decision trees. Gradient Boosted Decision Trees was introduced by Jerome Friedman in 1999.
R code

```r
> load(" wine . RData ")
> ds <- wine
> ds$ Type <- as. numeric (ds$ Type )
> ds$ Type [ds$ Type >1] <- 0
> ds$ Type
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 [149] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

> ds.gbm <- gbm( Type ~ Alcohol + Malic + Ash + Alcalinity + Magnesium Phenols + Flavanoids + Nonflavanoids + Proanthocyanins Color + Hue + Dilution + Proline ,
  data =ds , distribution =" adaboost ", n. trees =100)
```
Relative Importance of Variables

Figure 10.14: Relative importance of the predictors for the California housing data.

\[
I_i^2(T) = \sum \text{over all node using } x_i \text{ for partition} \left( f(x_i) - y_i \right)^2
\]

\[
I_i^2 = \frac{1}{M} \sum_{m=1}^{M} I_i^2(T_m)
\]

\[
\sum \text{over all node using } x_i \text{ for partition} \left( \mu(x_i) - y_i \right)^2
\]
Gradient Boosted Trees

> gbm.show.rules(ds.gbm)
Number of models: 100

Tree 1: Weight XXXX
  Proline < 867.50 : 0 (XXXX/XXXX)
  Proline >= 867.50 : 1 (XXXX/XXXX)
  Proline missing : 0 (XXXX/XXXX)

[...]

Tree 100: Weight XXXX
  Proline < 755.00 : 0 (XXXX/XXXX)
  Proline >= 755.00 : 1 (XXXX/XXXX)
  Proline missing : 0 (XXXX/XXXX)
Relative Importance of Variables

– For a single tree, define the importance of $x_i$ as

$$I_i^2(T) = \sum \text{improve in square error risk over for a constant fit over the region over all node using } x_i \text{ for partition}$$

– For additive tree, define the importance of $x_i$ as

$$I_i^2 = \frac{1}{M} \sum_{m=1}^{M} I_i^2(T_m)$$

Figure 10.14: Relative importance of the predictors for the California housing data.
Ad Quality Outline

- Motivation
- Machine Learning Overview
  - Background
  - Gradient Descent
  - From Perceptrons to SVMs
  - Probabilistic Models
  - Decision Trees
  - Bagging/Boosting
- Metrics
  - IR as a means of Ranking
  - Learning to Rank (LETOR)
  - Online Learning
  - Open issues
Metrics (some examples)

• **ML Metrics**
  – Mean squared error, Accuracy, Maximum Likelihood, Deviance, Cross entropy, KL Divergence, etc.

• **IR metrics**
  – Precision, Recall, MRR, DCG, pRank etc.
  – CIKM 2009 Paper on a new flavor of DCG!!

• **Advertising Metrics**
  – CTRs, Revenue, ECPM
## Possible Document Classifications

<table>
<thead>
<tr>
<th>Document Type</th>
<th>Retrieved</th>
<th>Not Retrieved</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevant</td>
<td>$R^+$</td>
<td>$R^-$</td>
</tr>
<tr>
<td></td>
<td>Type II Error (False Negative)</td>
<td></td>
</tr>
<tr>
<td>Not Relevant</td>
<td>$N^+$</td>
<td>$N^-$</td>
</tr>
<tr>
<td></td>
<td>Type I Error (False Positive)</td>
<td></td>
</tr>
</tbody>
</table>

\[
\text{accuracy} = \frac{R^+ + N^-}{R^+ + R^- + N^+ + N^-} \quad \text{recall} = \frac{R^+}{R^+ + R^-} \quad \text{precision} = \frac{R^+}{R^+ + N^+}
\]

\[
F_{0.5} = \frac{1.25 \times R^+}{0.25 \times R^- + N^+ + 1.25 \times R^+}
\]
User Models

- User satisfaction/expectation can be expressed as a utility measure
- Linear utility provides an empirically proven evaluation metric
  - TREC (Text Retrieval Conference)
  - Linear Utility = 2R⁺ - N⁺
- Incorporate user models into SVM threshold relaxation operation

<table>
<thead>
<tr>
<th>Document Type</th>
<th>Retrieved</th>
<th>Not Retrieved</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevant</td>
<td>R⁺</td>
<td>R⁻</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type II (FN)</td>
</tr>
<tr>
<td>Not Relevant</td>
<td>N⁺</td>
<td>N⁻</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type I (FP)</td>
</tr>
</tbody>
</table>
Metrics for Web Search

• **Existing metrics limited**
  – Not always clear-cut binary decision: relevant vs. not relevant
  – Not position sensitive:
    d: *definitely relevant*, p: *partially relevant*, n: *not relevant*
    ranking 1: p d p n n n n
    ranking 2: d p n p n n n

• **How do you measure recall over the whole web?**
  – How many of the potentially billions results will get looked at? Which ones actually need to be good?

• **Normalized Discounted Cumulated Gain (NDCG)**
  – Gain: relevance of a document is no longer binary
  – Sensitive to the position of highest rated documents
    • Log-discounting of gains according to the positions
  – Normalize the DCG with the “ideal set” DCG (NDCG).
Normalized Discounted Cumulative Gain (NDCG)

- **Judgments**
  - E.g., Very relevant > Somewhat relevant > Not relevant
- **Want most relevant documents to be ranked first**

\[
DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}
\]

- **NDCG = DCG / ideal ordering DCG**
  - NDCG Ranges from 0 to 1
NDCG Example

- Proposed document ordering:
  4 2 0 1

- DCG = 4 + 2/log(2) + 0/log(3) + 1/log(4) = 6.5

- IDCG = 4 + 2/log(2) + 1/log(3) + 0/log(4) = 6.63

- NDCG = 6.5/6.63 = .98
Normalized Discounted Cumulative Gain (NDCG) evaluation measure

- Query: $q_i$
- DCG at position $m$: $N_i = Z_i \sum_{j=1}^{m} (2^{r(j)} - 1)/\log(1 + j)$
- NDCG at position $m$: average over queries
- Example
  - $(3, 3, 2, 2, 1, 1, 1)$ rank $r$
  - $(7, 7, 3, 3, 1, 1, 1)$ gain $2^{r(j)} - 1$
  - $(1, 0.63, 0.5, 0.43, 0.39, 0.36, 0.33)$ discount $1/\log(1 + j)$
  - $(7, 11.41, 12.91, 14.2, 14.59, 14.95, 15.28)$ $\sum_{j=1}^{m} (2^{r(j)} - 1)/\log(1 + j)$
  - $Z_i$ normalizes against best possible result for query, the above, versus lower scores for other rankings
    - Necessarily: High ranking number is good (more relevant)

[IIR secs 6.1.2–3 and 15.4]
Ad Quality Outline

• Motivation
• Machine Learning Overview
• Metrics
  • IR as a means of Ranking Ads
  • Learning to Rank (LETOR)
  • Online Learning
  • Open issues
Ad Quality Outline

• Motivation
• Machine Learning Overview
• Metrics
  • IR as a means of Ranking Ads
• Learning to Rank (LETOR)
• Online Learning
• Open issues
Ranking Ads using IR

User → Target Page → Targeting Engine → Ads Creatives → Ads → Landing Pages

Features: View TP as query

IR Engine:

View Ad as document with different sections
Keywords: Title + Description + URL

Landing Page

RuSSIR 2009, Petrozavodsk, Russia. Online Advertising © 2009 James G. Shanahan. All Rights Reserved.

James.Shanahan_AT_gmail_DOT_com
Ranking ads using IR

- Contextual advertising study
- View target page as a query and use IR techniques
- Features
  - Rank ads for target page
    - by matching target page words with ad keywords and ad creative words
  - Variations
    - Query Expansion; expand keywords of target page
    - Use words from landing page

\[
sim(q, d_j) = \frac{\vec{q} \cdot \vec{d}_j}{|\vec{q}| \times |\vec{d}_j|} = \frac{\sum_{i=1}^{n} w_{iq} \cdot w_{ij}}{\sqrt{\sum_{i=1}^{n} w_{iq}^2} \sqrt{\sum_{i=1}^{n} w_{ij}^2}}
\]

See [Berthier A. Ribeiro-Neto, Marco Cristo, Paulo Braz Golgher, Edleno Silva de Moura: Impedance coupling in content-targeted advertising. SIGIR 2005: 496-503]
Content-Targeted Advertising
Matching Strategies

- Match the ad and its keywords to the Web page

[Ribeiro-Neto et al., SIGIR 2005]
Content-Targeted Advertising
Experimental Evaluation

- **Ad Sample Collection**
  - 1,744 announcers (Advertiser)
  - 93,972 ads in 2,029 campaigns
  - 68,238 keywords

- **Test Collection** (100 Queries)
  - 100 pages of a Brazilian newspaper
  - Topics include economy, sports, culture, and politics

- **Matching Function**
  - Vector based with tf-idf weights

[Ribeiro-Neto et al., SIGIR 2005]
Test Set Creation

15 Users manually labeled up to 30 ads per target page (on average, 5 from 15 were judged relevant)

For each of our 100 target pages, we selected ranked ads provided by each of our 15 systems. Thus, for each target page, we selected 15 ads. These target pages were each pooled to a manual set of those ads that were submitted to a manual set of users. The average number of ads per page pool was 5.15. The same pooling method used to evaluate our previous work [6].
Content-Targeted Advertising
Comparison among All Methods

AAK_EXP_H = TP Expansion, Keywords, Landing Page

[Ribeiro-Neto et al., SIGIR 2005]
Ranking Ads using IR

• Ribeiro et al. study was limited
  – Evaluated ranking of ads but not CTR estimation; this is a much easier problem;
  – System would not yield accurate CTR estimates in practice
  – Study highlights gap between academia and industry and a way of bridging it

• Jones, R., Rey, B., Madani, O., and Greiner, W. 2006. Generating query substitutions.

• This could be seen as the first step in ad serving
  – i.e., select a subset of ads and then do ECPM-based ranking of the this subset
Estimating CTRs using KW Clustering

- Estimate the CTR of new ads by using the CTRs of existing ads with the same bid terms or topic clusters
  - See [Regelson and Fain, 2006]
  - However, keywords have highly variable CTRs (e.g., max of “Lawyers” is 300% that of the average CTR)

[Richardson et al, 2007]
Other IR approaches to OAT

• **Leverage a taxonomy distance and traditional text match**
  – Classify query and ad into a taxonomy (node represents a set of queries)
  – Weighted sum of taxonomy/topic similarity score and traditional IR score.

• **Keyword harvesting from webpages**
  – Treat target page as a query but can be long. Suggest keywords. Which terms to extract? Train a LogReg classifier on hand-picked keywords and their features. Predict if a word is keyword or not. Study performed at Microsoft Research
  – [Yih et al., WWW 2006, “Finding advertising keywords on web pages”]

• **Ad Matching for Rare Queries (enhancing broad match)**
  – Use classes of top ranked pages (for a query) to suggest ads
  – [Robust Classification of Rare Queries…, Broder, et al., SIGIR 2007]

• **All the problems in traditional IR + more** (see WWW, SIGIR CIKM)
  – Synonym detection; entity extraction from queries; query disambiguation
Ad Quality Outline

• Motivation
• Machine Learning Overview
• Metrics
• IR as a means of Ranking
• Learning to Rank (LETOR)
• Online Learning
• Open issues
Learning to Rank

• Learning to rank documents (main approaches)
  – Point-based
  – Pair-wise-based
  – List-based

• Very active area of research
  – LETOR is a package of benchmark data sets for LEarning TO Rank, released by Microsoft Research Asia.
  – Extracted features for each query-document pair in the OHSUMED and TREC collections
  – Benchmarked several state-of-the-arts ranking models with these features and provide baseline results for future studies

• Similar principles apply to ranking of ads
Table 3. All the features for the TREC datasets

<table>
<thead>
<tr>
<th>Feature</th>
<th>Descriptions</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BM25</td>
<td>[27]</td>
</tr>
<tr>
<td>2</td>
<td>document length (dl) of body</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>dl of anchor</td>
<td>[1]</td>
</tr>
<tr>
<td>4</td>
<td>dl of title</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>dl of URL</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>HITS authority</td>
<td>[21]</td>
</tr>
<tr>
<td>7</td>
<td>HITS hub</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>FostRank (SIGIR feature)</td>
<td>[34]</td>
</tr>
<tr>
<td>9</td>
<td>Inverse document frequency (idf) of body</td>
<td>[39]</td>
</tr>
<tr>
<td>10</td>
<td>idf of anchor</td>
<td>[1]</td>
</tr>
<tr>
<td>11</td>
<td>idf of title</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>idf of URL</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Sitemap based score propagation (SIGIR feature)</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>PageRank</td>
<td>[25]</td>
</tr>
<tr>
<td>15</td>
<td>LMIR.ABS of anchor</td>
<td>[35]</td>
</tr>
<tr>
<td>16</td>
<td>BM25 of anchor</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>LMIR.DIR of anchor</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>LMIR.JM of anchor</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>LMIR.ABS of extracted title (SIGIR feature)</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>BM25 of extracted title (SIGIR feature)</td>
<td>[29]</td>
</tr>
<tr>
<td>21</td>
<td>LMIR.DIR of extracted title (SIGIR feature)</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>LMIR.JM of extracted title (SIGIR feature)</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>LMIR.ABS of title</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>BM25 of title</td>
<td>[29]</td>
</tr>
<tr>
<td>25</td>
<td>LMIR.DIR of title</td>
<td>[35]</td>
</tr>
<tr>
<td>26</td>
<td>LMIR.JM of title</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Sitemap based feature propagation (SIGIR feature)</td>
<td>[28]</td>
</tr>
<tr>
<td>28</td>
<td>tf of body</td>
<td>[1]</td>
</tr>
<tr>
<td>29</td>
<td>tf of anchor</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>tf of title</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>tf of URL</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>tf*idf of body</td>
<td>[1]</td>
</tr>
<tr>
<td>33</td>
<td>tf*idf of anchor</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>tf*idf of title</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>tf*idf of URL</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>Topical PageRank (SIGIR feature)</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>Topical HITS authority (SIGIR feature)</td>
<td>[24]</td>
</tr>
<tr>
<td>38</td>
<td>Topical HITS hub (SIGIR feature)</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>Hyperlink base score propagation: weighted in-link (SIGIR feature)</td>
<td>[28] [30]</td>
</tr>
<tr>
<td>40</td>
<td>Hyperlink base score propagation: weighted out-link (SIGIR feature)</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>Hyperlink base score propagation: uniform out-link (SIGIR feature)</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>Hyperlink base feature propagation: weighted in-link (SIGIR feature)</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>Hyperlink base feature propagation: weighted out-link (SIGIR feature)</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>Hyperlink base feature propagation: uniform out-link (SIGIR feature)</td>
<td></td>
</tr>
</tbody>
</table>

Text, Anchor Text, Web graph, Log features, User features, and others.

Great start but be aware....

Missing Features
Current LETOR baselines

- **LETOR Algorithms**
  - Ranking SVM
  - RankBoost
  - AdaRank
  - Multiple hyperline ranker
  - FRank
  - ListNet

- **LETOR Datasets**

- **LETOR metrics such as DCG, NGCG**

(b) Mean average precision

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>RankBoost</td>
<td>0.383514</td>
</tr>
<tr>
<td>Ranking SVM</td>
<td>0.350459</td>
</tr>
</tbody>
</table>
LETOR test collection

- From Microsoft Research Asia
- An openly available standard test collection with pregenerated features, baselines, and research results for learning to rank
- It’s availability has really driven research in this area
- OHSUMED, MEDLINE subcollection for IR
  - 350,000 articles
  - 106 queries
  - 16,140 query-document pairs
  - 3 class judgments: Definitely relevant (DR), Partially Relevant (PR), Non-Relevant (NR)
- TREC GOV collection (predecessor of GOV2, cf. IIR p. 142)
  - 1 million web pages
  - 125 queries
Dataset Partitioning

- **five-fold cross validation**
  - The *training set* is used to learn the ranking model.
  - The *validation set* is used to tune the parameters of the ranking model.
  - The *test set* is used to report the ranking performance of the model.

### Table 4. Data Partitioning for 5-fold Cross Validation

<table>
<thead>
<tr>
<th>Folds</th>
<th>Training set</th>
<th>Validation set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold1</td>
<td>{S1, S2, S3}</td>
<td>S4</td>
<td>S5</td>
</tr>
<tr>
<td>Fold2</td>
<td>{S2, S3, S4}</td>
<td>S5</td>
<td>S1</td>
</tr>
<tr>
<td>Fold3</td>
<td>{S3, S4, S5}</td>
<td>S1</td>
<td>S2</td>
</tr>
<tr>
<td>Fold4</td>
<td>{S4, S5, S1}</td>
<td>S2</td>
<td>S3</td>
</tr>
<tr>
<td>Fold5</td>
<td>{S5, S1, S2}</td>
<td>S3</td>
<td>S4</td>
</tr>
</tbody>
</table>
Homework

- Download LETOR OHSUMED
- Model ranking function using linear regression (in R) and compare with a standard SVM, ranking SVM
- Compute DCG, NDCG scores and discuss your findings. Base results on five-fold cross validation
Experimental Results (OHSUMED)
Resources

- *IIR* secs 6.1.2–3 and 15.4
- **LETOR benchmark datasets**
  - Website with data, links to papers, benchmarks, etc.
  - [http://research.microsoft.com/users/LETOR/](http://research.microsoft.com/users/LETOR/)
  - Everything you need to start research in this area!
- **SIGIR 2007-9 workshops**
Learning to Rank Ads from Clicks

• LETOR Ads in the context of sponsored search
• More relevant versus absolutely relevant (preference judgements)
  – Infer pairwise preference of ads for a query using clicks based on the block approach proposed by [Joachims, KDD 2002]
  – Dropped clicked ads at rank 1; queries of 3 terms or more
  – 123,798 blocks [each query could be made up multiple blocks]
• Evaluated linear and nonlinear learning algorithms
  – Linear perceptron in classification mode {-1, +1}
  – Linear preceptron in pairwise learning mode [Ranking SVM]
  – Multilayer perceptron (Neural network) in classification mode {-1, +1}

[Ciaramita M., Murdock V., Plachouras, Online Learning from Click Data for Sponsored Search, WWW 2008]
Getting labeled data from clicks

<table>
<thead>
<tr>
<th>Ad Id</th>
<th>Clicked or not</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>Click</td>
</tr>
<tr>
<td>2</td>
<td>----</td>
</tr>
<tr>
<td>3</td>
<td>Clicked by user1</td>
</tr>
<tr>
<td>1</td>
<td>----</td>
</tr>
<tr>
<td>6</td>
<td>----</td>
</tr>
<tr>
<td>7</td>
<td>Clicked by user1</td>
</tr>
</tbody>
</table>

Delete Clicks at rank 1

Block 1: ad3 >> ad2

Block 2: ad7 >> ad6
    ad7 >> ad1

singles  ad3 labeled as +1 and ad2 as -1

pairwise  ad2 << ad3 is labeled -1
Ranking SVM: learning pairwise


- Input space: $X$
- Ranking function $f : X \rightarrow R$
- Ranking: $x_i \succ x_j \iff f(x_i; w) > f(x_j; w)$
- Linear ranking function: $f(x; w) = \langle w, x \rangle$
- $\langle w, x^{(1)} - x^{(2)} \rangle > 0 \iff f(x^{(1)}; w) > f(x^{(2)}; w)$
- Transforming to binary classification:

$$(\tilde{x}^{(1)} - \tilde{x}^{(2)}, z), \ z = \begin{cases} +1 & x^{(1)} \succ x^{(2)} \\ -1 & x^{(2)} \succ x^{(1)} \end{cases}$$

Shen and Joshi adapted this approach to the perceptron; [Shen and Joshi 2005]
## Features

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Abbrev.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td><strong>Word Overlap Features</strong></td>
</tr>
<tr>
<td>NoKey</td>
<td>N</td>
<td>1 if no query term is present in the ad materials; 0 otherwise</td>
</tr>
<tr>
<td>SomeKey</td>
<td>S</td>
<td>1 if at least one query term is present in the ad materials; 0 otherwise</td>
</tr>
<tr>
<td>AllKey</td>
<td>A</td>
<td>1 if every query term is present in the ad materials; 0 otherwise</td>
</tr>
<tr>
<td>PercentKey</td>
<td>P</td>
<td>The number of query terms present in the ad materials divided by the number of query terms</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Cosine Similarity Features</strong></td>
</tr>
<tr>
<td>Ad</td>
<td>B</td>
<td>The cosine similarity between the query and the ad materials (baseline)</td>
</tr>
<tr>
<td>Title</td>
<td>T</td>
<td>The cosine similarity between the query and the ad title</td>
</tr>
<tr>
<td>Description</td>
<td>D</td>
<td>The cosine similarity between the query and the ad description</td>
</tr>
<tr>
<td>Bidterm</td>
<td>D</td>
<td>The cosine similarity between the query and the bidded terms</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Correlation Features</strong></td>
</tr>
<tr>
<td>AvePMI</td>
<td>A</td>
<td>The average pointwise mutual information between terms in the query and terms in the ad</td>
</tr>
<tr>
<td>MaxPMI</td>
<td>M</td>
<td>The maximum pointwise mutual information between terms in the query and terms in the ad</td>
</tr>
<tr>
<td>CSQ</td>
<td>C</td>
<td>Number of query-ad term pairs that have a $\chi^2$ statistic in the top 5% of computed $\chi^2$ values.</td>
</tr>
</tbody>
</table>

Table 1. Summary of features. The column “Abbrev.” provides an abbreviated name for one or more features, as they will be used in the experiments.
MLP seems to give best results

• …though not sure about the significance compared to other learnt models

• Limited to ranking of ads

\[ RR = \frac{1}{\text{RankOfFirstRel}} \]

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Classification</th>
<th>Ranking</th>
<th>Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec at 1</td>
<td>MRR</td>
<td>Prec at 1</td>
</tr>
<tr>
<td>B</td>
<td>0.322</td>
<td>0.582 ± 0.306</td>
<td>0.333</td>
</tr>
<tr>
<td>BO</td>
<td>0.319</td>
<td>0.578* ± 0.306</td>
<td>0.352</td>
</tr>
<tr>
<td>BF</td>
<td>0.341</td>
<td>0.593* ± 0.309</td>
<td>0.347</td>
</tr>
<tr>
<td>BFO</td>
<td>0.357</td>
<td>0.605* ± 0.311</td>
<td>0.357</td>
</tr>
<tr>
<td>BFOP</td>
<td>0.357</td>
<td>0.604* ± 0.311</td>
<td>0.359</td>
</tr>
<tr>
<td>BFOC</td>
<td>0.351</td>
<td>0.601*† ± 0.310</td>
<td>0.364</td>
</tr>
<tr>
<td>BFOCP</td>
<td>0.360</td>
<td>0.606* ± 0.311</td>
<td>0.363</td>
</tr>
</tbody>
</table>

Table 4. The results for classification, ranking and regression, computed over all trials. The best result is indicated in bold. Results that are statistically significant with respect to the baseline are indicated with a star. Results indicated with a dagger are statistically significant with respect to the features B + F + O. The results for precision at one were not tested for statistical significance.
Ranking Ads using IR

User

Target Page

Targeting Engine

Ad Creatives

Landing Pages

ML Model

Zip Code
Browser Type
Etc

Terms etc.

Keywords
Title + Description + URL

Landing Page Appearance
Estimating CTRs using ML

- Estimate CTR using $\Pr_{\text{Ad}}(\text{Click}|\text{Keyword})$
- Frame as machine learning problem
  - E.g., Matthew Richardson, Ewa Dominowska, Robert Ragno: Predicting clicks: estimating the click-through rate for new ads. WWW 2007 pages 521-530
  - Model using Logistic Regression and MART (Boosted decision trees using stochastic gradient descent) [Friedman 2000]

What features could be used?
ML Features 1/2

- **Historical data**
  - CTR of KW based on other ads with this KW
  - Related terms CTRs

- **Appearance**
  - #words in title/body; capitalization; punctuation; word length

- **Attention Capture**
  - Title/body contain action words, e.g., buy/join/etc

- **Reputation**
  - .com/.net/etc, length of URL, #segments in URL, numbers in URL

- **Landing page quality**
  - Contains flash? Fraction of page in images? W3C compliant

- **Text Relevance**
  - keyword match with ad title/body; fraction of match

\[ p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \ldots + \beta_k x_{k,i})}} \]
ML Features 2/2

- Historical data
- Related terms CTRs
- Appearance
- Attention Capture
- Reputation
- Landing page quality

- **Text Relevance**
  - keyword match with ad title/body; fraction of match

- **10K unigrams (appearing in Ad title and Ad body); bi/trigrams did not bring significant improvement;**
  - Binary feature; 1 if term occurs in ad 0 otherwise

- **Freq of term on web; in query logs**

- **Many others could be used!!!**  
  [Richardson et al.]
Learning Setup

- **Logistic Regression**
  - Used a cross entropy loss function
  - Standardized all features using training data
    - (mean and variance, of 0 and 1)
  - Thresholded data beyond 5 std deviations
  - Added derived features
    - (i.e., foreach feature $f$, $\log(f + 1)$ and $f^2$)

- **Baseline**
  - Predict the average CTR of the training dataset

- **MART (Boosted decision trees using stochastic gradient descent [Friedman 2000])**
  - Experiments did not show significant improvement over LR
  - LR is a more transparent model

For LR see:
2. [http://statgen.iop.kcl.ac.uk/bgim/mle/sslike_4.html](http://statgen.iop.kcl.ac.uk/bgim/mle/sslike_4.html)
Learning Setup

• Error measures
  – Mean Squared error between predicted CTR and true CTR
  – KL Divergence between the predicted CTR and true CTR (in both cases lower is better; 0 is best)

• Issues?
  – Weighted?
  – ??
10,000 Advertisers
1 Million examples of <Keyword, Ad> -> CTR
- (view <Keyword, Ad> as <TP, Ad>)
Keywords are both exact and broad match
100,000 unique ad texts
Required that each example had more than 100 views
70-10-20 data split (train, validation, test)

[Richardson et al.]
Table 7: Comparison of results for a model trained and tested on ads with over 100 views vs. over 1000 views.

<table>
<thead>
<tr>
<th>Features</th>
<th>%Imprv &gt;100 views</th>
<th>%Imprv &gt;1000 views</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (CTR)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>+Term CTR</td>
<td>13.28</td>
<td>25.22</td>
</tr>
<tr>
<td>+Related CTR</td>
<td>19.67</td>
<td>32.92</td>
</tr>
<tr>
<td>+Ad Quality</td>
<td>23.45</td>
<td>33.90</td>
</tr>
<tr>
<td>+Order Specificity</td>
<td>28.97</td>
<td>40.51</td>
</tr>
<tr>
<td>+Search Data</td>
<td>29.47</td>
<td>41.88</td>
</tr>
</tbody>
</table>
## Transparency of Results

### Table 5: Non-unigram features with highest (lowest) weight

<table>
<thead>
<tr>
<th>Top ten features</th>
<th>Bottom ten features</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(#chars in term)</td>
<td>log(# terms in order)</td>
</tr>
<tr>
<td>$v_{12}$</td>
<td>log($v_{0+}$)</td>
</tr>
<tr>
<td>$v_{22}$</td>
<td>sq$(p_{00})$</td>
</tr>
<tr>
<td>log(order category entropy)</td>
<td>sq$(p_{0+})$</td>
</tr>
<tr>
<td>log(#most common word)</td>
<td>log(#chars in landing page)</td>
</tr>
<tr>
<td>sq$(#segments in displayurl)$</td>
<td>log($a_{01}$)</td>
</tr>
<tr>
<td>sq$(#action words in body)$</td>
<td>$a_{13}$</td>
</tr>
<tr>
<td>$P_{10}$</td>
<td>sq$(p_{+0})$</td>
</tr>
<tr>
<td>$p_{**}$</td>
<td>log(#chars in body)</td>
</tr>
<tr>
<td>log($v_{00}$)</td>
<td>sq$(#chars in term)$</td>
</tr>
</tbody>
</table>

### Table 6: Unigrams with highest (and lowest) weight.

<table>
<thead>
<tr>
<th>Top ten unigrams</th>
<th>Bottom ten unigrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>official body</td>
<td>quotes title</td>
</tr>
<tr>
<td>download title</td>
<td>hotels title</td>
</tr>
<tr>
<td>photos body</td>
<td>trial body</td>
</tr>
<tr>
<td>maps body</td>
<td>deals body</td>
</tr>
<tr>
<td>official title</td>
<td>gift body</td>
</tr>
<tr>
<td>direct body</td>
<td>have text</td>
</tr>
<tr>
<td>costumes title</td>
<td>software title</td>
</tr>
<tr>
<td>latest body</td>
<td>engine body</td>
</tr>
<tr>
<td>version body</td>
<td>compare title</td>
</tr>
<tr>
<td>complete body</td>
<td>secure body</td>
</tr>
</tbody>
</table>
CTR Evolution

Model predicted CTR using a Beta distribution where $p_0$ is the predicted (and baseline) CTR

\[ \hat{p} = \frac{\alpha p_0 + \text{clicks}}{\alpha + \text{views}} \]

Figure 6: Expected mean absolute error in CTR as a function of the number of times an ad is viewed.
Estimating CTRs using ML

Intermediate Conclusions

• Richardson et al. report a very interesting approach and case study
  – Despite realistic problem setting results are preliminary

• Transparency of model

• Using many features helps insulate from adversarial attacks (can be useful in adversarial detection)

• Applied to new ads but could be extended to deal with existing ads, display/graphical ads
  – Homework!!

• But many issues remain!!
Ad Quality Outline

- Motivation
- Machine Learning Overview
- IR as a means of Ranking
- Learning to Rank (LETOR)
- Online Learning
- Open issues
Modeling CTR Challenges

- Extremely rare events (Typical CTRs < 1% for contextual)
- Biased dataset (the rich get richer; suboptimal locking)
- Very sparse (only a small percentage of <TP, Ad> get impressions; can impede generalization)
  - Missed opportunities
- **Accuracy of estimates**
  - ML approaches are hugely biased; bias correction [see Provost and Domingos; Platt]
- **Scale and Speed**
- **Non-Stationary, new ads, changes in network**
- **Marginalization versus segmentation (resolution vs. sufficient data)**
- ....
Other Challenges in the CPC world

- Define a portfolio of keywords and bid prices
  - Number of keywords per campaign per month is increasing
  - 9,100 keywords in 9/2004 to 14,700 in 3/2005
- Manage keyword portfolio
- Impression Fraud, Click Fraud
CPA versus CPC

\[
\begin{align*}
&\text{CPC } _E \text{ ECPM}_A = \text{CTR}_A \times \text{Bid}_A \\
&\text{CPA } _E \text{ ECPM}_A = \text{CTR}_A \times [\text{AR}_A] \times \text{Bid}_A
\end{align*}
\]

- Similar estimation challenges (but on even rarer data)
- Non-keyword –based networks (term extraction)
- www.Turn.com is a trail blazer in automatically targeted CPA marketplaces
- Google is currently offering CPA in beta-mode (targeting done by publisher or automatically)
CPC versus CPA (e.g., Turn Inc)

User → Target Page → Targeting Engine → Ads → Ad Creatives → Landing Pages
Outline

• Introduction
• Online advertising background
• Business models
• Creating an online ad campaign
• Technology and Economics
  – Advertisers (optimizing ROI thru ads and ad placement)
  – Publishers (optimizing revenue and consumer satisfaction)
    • Forward Markets
    • Spot Markets
      – Background
      – Auction Systems, Game Theory
      – Ad Quality
      – Budgeting

• New Directions
• Challenges in online advertising
• Summary
ECPM with Budget Constraints

- Advertisers impose budgets on money spent
- Allocate ads greedily (consume budget of highest bidder)
  - *Not optimal*
  - Worst case scenario: 50% of Optimal
- Introduce a throttle factor to slow down budget exhaustion

\[
\text{CPC} - ECPM_{Ad} = CTR_{Ad} \times Bid_{Ad} \times \text{ThrottleFactor}
\]


RuSSIR 2009, Petrozavodsk, Russia. Online Advertising © 2009 James G. Shanahan (San Francisco)
James.Shanahan_AT_gmail_DOT_com
ECPM with Budget Constraints

• Advertises impose budgets on money spent
• Allocate ads greedily (consume budget of highest bidder)
  – Not optimal
  – Worst case scenario: 50% of Optimal
• Introduce a throttle factor to slow down budget exhaustion
  – Throttle algorithm: \( 1 - \frac{1}{e} \approx 0.63 \) competitive (63% competitive); theoretical result [Mehta et al.]

\[
CPC - ECPM_{Ad} = CTR_{Ad} \times Bid_{Ad} \times \text{ThrottleFactor}
\]

\[
\text{ThrottleFactor} = 1 - e^{-(1 - \text{fractionOfBudgetSpent})}
\]

Technology + Economics Summary

• Forward Markets vs Spot Markets
• Allocations via operations research
• Advertising is not search!
  – Relevance == Topicality|Usefulness? (hint $$)
  – Ranking versus prediction
  – Intellectually diverse discipline (Economics, IR, ML, Stats, Social)
• Key directions
  – Accurately estimating CTR/AR; auction systems; new learning algorithms; targeting graphical ads
  – Auction mechanisms
  – Global optimization algorithms
Technology + Economics Summary

• **Forward Markets vs Spot Markets**
• **Allocations via operations research**
• **Advertising is not search!**
  – Relevance == Topicality|Usefulness? (hint $$)
  – Ranking versus prediction
  – Intellectually diverse discipline (Economics, IR, ML, Stats, Social)
• **Key directions**
  – Accurately estimating CTR/AR; auction systems; new learning algorithms; targeting graphical ads
  – Auction mechanisms
  – Global optimization algorithms
Outline

• Introduction
• Online advertising background
• Business models
• Creating an online ad campaign
• Technology and Economics

• New Directions
  – Behavioral Targeting
  – Mobile
  – Web 2.0
  – Social Advertising
  – Data Exchanges

• Challenges in online advertising
• Summary
New Directions

• **Behavioral Targeting**
  – CTRs >> E.g., 11% to 840% higher than average
  – $8 Billion US Display ad market

• **Mobile**
  – By 2012, the number of devices accessing the Internet will double to more than three billion, half of which will be mobile devices [IDC, 2008]

• **Web 2.0**
  – E.g., Video: Research firm eMarketer predicts online video advertising will grow 45% this year to $850 million.
  – Freeium models are more actively being explored

• **Social Advertising**
  – leverages historically "offline" dynamics, such as peer-pressure, recommendations, and other forms of social influence

• **Data Exchanges**
  – Mining and auctioning audiences from user transactions and browsing at partner sites (ecommerce or otherwise)
Behavioral Targeting: Modeling The User

- Target ads based on user’s browsing behavior
  - Behavioral targeting uses information collected on an individual's web-browsing behavior, such as the pages they have visited or the searches they have made, to select which advertisements to display to that individual.
  - Helps them deliver their online advertisements to the users who are most likely to be influenced by them.

- Commonly used by:
  - Website owners (e.g., e-commerce websites)
  - Ad networks

- Key players include:
  - E-commerce websites such as Amazon
  - Blue Lithium (acq by Yahoo!, $300M), Tacoda (acq by AOL, $275M), Burst, Phorm and Revenue Science, Turn.com, and others...

[For more background see: http://en.wikipedia.org/wiki/Behavioral_targeting]
Ad Network Behavioral Targeting

• Selling Audiences (and not sites)
  – Traditionally did this based on panels (user surveys or using Comscore/NetRatings); very broad and not very accurate
  – Through a combination of cookies and log analysis BT enables very specific segmentation

• BT based on:
  – Website/web page visited
    • E.g., Users who frequent Orbitz.com and Expedia.com or the travel section of USAToday.com would become part of the “Travel Shoppers” segment. Then, these users are re-targeted when they are found on other more general content type sites
  – Keyword profile (using recent searches or content that was read)
  – Retarget past visitors to your website
    • The goal is to use BT to let you send the appropriate message to each user based on where they are in your product sales cycle.
BT Extends Targeting Variables

- **Behavioral variables**, e.g.,:
  - Client site visits/conversions
  - Email/newsletter list membership
  - User interest category

- **Demographic/psychographic variables:**
  - Geography
  - Prizm™ Cluster
  - Day part/day of week
  - Gender
  - At work/at home

- **Other variables:**
  - Connection speed
  - Browser/OS
  - ISP/domain”
  - *And many others (that we have looked at already)*
• **Segmenting users?**
  – Clustering, data mining, classification
  – Rule-based system, hybrid systems
  – Segmenting publisher real estate into categories

• **Collaborative filtering**
  – People who bought this also bought X…

• **Collaborative filtering algorithms**
  – Yehuda Koren, Factorization meets the neighborhood: a multifaceted collaborative filtering model. KDD 2008 426-434
  – Remember NetFlix problem??
  – Gradient descent etc…

• **Very active area!!!**
BT is a happening place

• $11B M&A of real estate and behavioral targeting
  – 2007 Yahoo acquires RightMedia and BlueLithium (1$B)
  – Google acquires DoubleClick ($3B)
  – Microsoft Acquires Aquantive ($6B)
  – AOL acquires Quigo ($340M), Tacoda (~$300M)

• Google’s Interest-based Advertising announced [3/2009]

• Yahoo! SmartAds (dynamically assemble ads)
  – Behavioral, demographic and geographic targeting
  – For example, if a user is browsing for hybrid cars in Yahoo! Autos and has selected San Francisco as their default location in Yahoo! Weather, Yahoo!'s SmartAds platform can assemble and deliver a display ad in real time that showcases a hybrid vehicle from a major auto brand, as well as local dealer information and current lease rates.
Behavioral Targeting (BT) attempts to deliver the most relevant advertisements to the most interested audiences, and is playing an increasingly important role in online advertising market. However, there have been not any public works investigating on how much the BT can truly help online advertising in commercial search engines? To answer this question, in this paper we provide an empirical study on the ads click-through log collected from a commercial search engine. From the comprehensively experimental results on the sponsored search log of a commercial search engine over a period of seven days, we can draw three

Conclusions:
- Users who clicked the same ad will truly have similar behaviors on the Web;
- The Click-Through Rate (CTR) of an ad can be averagely improved as high as 670% by properly segmenting users for behavioral targeted advertising;
- Using the short term user behaviors to represent users is more effective than using the long term user behaviors for BT.
- Represented users in terms of a search keyword profile and URL-profile over one day or one week.

[How much the Behavioral Targeting can Help Online Advertising? Jun Yan, et al., WWW 2009]
Large-Scale Behavioral Targeting - Ye Chen* Yahoo! Labs; Dmitry Pavlov Yahoo! Labs; John Canny Computer Science Division University of California Berkeley

- Best Application Paper Award Winner
- Describe a highly scalable and efficient solution to BT using Hadoop MapReduce framework.
- Parallel algorithm that builds 450 BT-category models from the entire Yahoo's user base
- Describe approach has yielded 20% CTR lift over the existing production system by leveraging a well-grounded probabilistic model fitted from a much larger training dataset.
NebuAd suffers a public outcry

• NebuAd (Silicon Valley) was one of several companies developing behavioral targeting advertising systems, seeking deals with ISPs to enable them to analyse customer's websurfing habits in order to provide them with more relevant, micro-targeted advertising.

• Due to fallout following (U.S.) public and Congressional concern, NebuAd's largest customers have all pulled out.
Privacy

• Behavioral targeting agencies claim that their actions are legal, due to the privacy policy information located on websites. Most of them contain links that allow consumers to opt out of tracking.

• BlueKai lets consumers opt out of having their data traded online, as well as letting you “see exactly what data is known about you by marketers” through its BlueKai Registry program. Other networks allow similar capabilities.

• A common complaint is the lack of or not-so-easy-to-access information.

[See http://digg.com/d1o46k]
In the US

- [http://digg.com/d1mVck](http://digg.com/d1mVck) (Techrepublic.com)
- Self-regulation
- The [FTC’s](http://www.ftc.gov/privacy/) (governing body) response (somewhat aligned with Eric Schmidt’s) is to let the market regulate itself, but with new and stronger guidelines. I penned an article titled, “[Behavioral targeting: FTC still prefers self-regulation](http://www.ftc.gov/privacy/)” that explains the new guidelines.
Ad Matching Opt-Out

When possible, we try to match the ads that we show you to your interests. We call this "ad matching," (sometimes also called ad customization) and we do this in order to make the ads you see more relevant and useful for you. Ad matching uses data about your visits to both Yahoo! and our partner sites and about the ads you view and click.

We offer you the choice to opt-out of ad matching. If you opt-out the ads we show you will be more general and less relevant to you. Opting-out will not turn off ads.

We also use web beacons for a number of purposes including research and reporting for some of our clients. If you opt-out out of ad matching, Yahoo! also will not use the information gathered through our web beacons for research and reporting purposes.

You are currently not opted out

Opt-Out

If your status has not changed after selecting Opt-Out, your browser may be set to block Yahoo! Cookies. Please unblock Yahoo! Cookies and try again.

Note: This opt-out applies to a specific browser rather than a specific user. Therefore you will have to opt-out separately from each computer or browser that you use.

Yahoo! is a participating member of the Network Advertising Initiative. You can exercise the same Yahoo! opt-out on the NAI site and also manage your other ad network opt outs.

To see which ad networks serve ads on Yahoo!, you can visit our page about Third Party and Affiliate Cookies on Yahoo!
The Network Advertising Initiative (NAI)

A consortium of approximately 30 companies that use BT technology. Opt-out easily!!

Opt Out of Behavioral Advertising

The NAI Opt-out Tool was developed in conjunction with our members for the express purpose of allowing consumers to "opt out" of the behavioral advertising delivered by our member companies.

Using the Tool below, you can examine your computer to identify those member companies that have placed an advertising cookie file on your computer.

To opt out of an NAI member's behavioral advertising program, simply check the box that corresponds to the company from which you wish to opt out. Alternatively, you can check the box labeled "Select All" and each member's opt-out box will be checked for you. Next click the "Submit" button. The Tool will automatically replace the specified advertising cookie(s) and verify your opt-out status.

Opting out of a network does not mean you will no longer receive online advertising. It does mean that the network from which you opted out will no longer deliver ads tailored to your Web preferences and usage patterns.

If you have any questions, please visit our FAQ section.

Opt-Out Status

<table>
<thead>
<tr>
<th>Member Company</th>
<th>Status</th>
<th>Opt-Out</th>
</tr>
</thead>
<tbody>
<tr>
<td>aCerno</td>
<td>No Cookie</td>
<td>Opt-Out □</td>
</tr>
<tr>
<td>Advertising.com</td>
<td>Active Cookie</td>
<td>Opt-Out □</td>
</tr>
<tr>
<td>Akamai</td>
<td>Active Cookie</td>
<td>Opt-Out □</td>
</tr>
</tbody>
</table>

Google
- More Information

interCLICK
- More Information

Media6degrees
- More Information

Mindset Media
- More Information

NextAction
- More Information

Safecount
- More Information

SpecificMEDIA
- More Information

Traffic Marketplace
- More Information

Turn
- More Information

24/7 Real Media
- More Information

Undertone Networks
- More Information

[x+1] (formerly Poindie
- More Information

Yahoo! Ad Network
Consumer Notice

Making advertising relevant, while respecting your choice.

TACODA works with hundreds of existing websites to ensure that the online advertising you see is relevant and useful.

We anonymously categorize web surfing interests using a small text file in the browser called a cookie to deliver targeted advertising. You may opt-out of these targeted ads at any time.

For example:

With TACODA-enabled ads you will be more likely to receive advertising that is relevant to you. For example, if you are researching a new car you will see:

- Make adventure thrilling in the adventurer.
- Do more

If you opt-out of TACODA, you are more likely to see advertising that is not relevant to you.

Sheep Herding

Herb 20 sheep, get a free lawn mower

TACODA's Use of Cookies Statement

- How Does TACODA Improve Online Advertising?
  For many, the term "online advertising" conjures up images of web pages cluttered with irrelevant ads and annoying offers. TACODA believes your online experience can be better. We deliver ads that are likely to offer greater relevance by working with hundreds of name brand websites and advertisers to deliver ads based on your previous web viewing interests.

- What is a Cookie?
Behavioural Targeting

- Holds huge promise to monetize longtail inventory (such as Web2.0)
- Privacy concerns need to be addressed before consumers, advertisers and publishers embrace this direction in a big way
Outline

• Introduction
• Online advertising background
• Business models
• Creating an online ad campaign
• Technology and Economics
• New Directions
  – Behavioral Targeting
  – Mobile
  – Web 2.0
  – Social Advertising
  – Data Exchanges
• Challenges in online advertising
• Summary
Mobile Web vs. Internet

Users

WWW

Advertisers

Publishers

Online

Users

Carriers

WWW

Advertisers

Publishers

Mobile
Tipping Points

• The true turning point for the industry was the introduction of the smartphone, heightened by Apple’s iPhone launch in mid-2007.
• Tearing down the carrier wall
• The development of third-generation (3G) mobile phones led to better connection speeds
• Wi-Fi connectivity and the rise of mobile Internet browsing.
• Richer Content: Open Internet
• In developing markets, mobile is the primary gateway to the Internet not PC!
Mobile Advertising

[John Gantz, IDC 2008]

• While the PC is currently the dominant means of gaining access to the Internet, IDC expects the number of mobile devices accessing the Internet will surpass the number of online PCs by 2012.

• Roughly 40 per cent of all Internet users worldwide currently have mobile Internet access. The number of mobile Internet users will reach 546 million in 2008, almost twice as many as in 2006, and is forecast to surpass 1.5 billion worldwide in 2012.

• The latest wave of post-iPhone smart phones have put mobile social networking, search, location services and even gaming and video into the fast lanes of user adoption.

• Data usage is up 39% in a year.
Branded Landing Page & Click-to-Video

CPM Banner Ad

Branded Landing Page

Gallery/Image

Movie Trailer

VERSUS

2005-2007 AdMob, Inc. Confidential & Proprietary

RuSSIR 2009, Petrozavodsk, Russia. Online Advertising © 2009 James G. Shanahan (San Francisco)
James.Shanahan_AT_gmail_DOT_com
Mobile Monetization

- All the internet monetization models are applicable to mobile.
- New monetization models that leverage unique mobile features e.g. pay per call.

<table>
<thead>
<tr>
<th></th>
<th>Search</th>
<th>Contextual</th>
<th>Display</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPM</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>CPC</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>CPA</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>PPC</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>
Select Best Ads to Display

• **Requires:**
  – Ads to select from (i.e., ads with budget remaining)
  – Ability to determine best ad to show (see below);
  – Serve ad or not (social responsibility)

• **Selecting Best Ad(s)**
  – Estimate the CTR of each ad using available features.
  – A number of techniques for estimating CTR:
    • Regularized Regression models
    • Multi-armed bandit models
    • Non parametric nearest neighbor methods.
  – Features for estimating CTR:
    • Site characteristics
    • Ad Characteristics
    • User Characteristics
Targeting for Mobile versus Online

How to predict CTR and allocate based on the product of CTR and CPC (ECPM)?

Do not have the features available to search engines for ranking sponsored search.

Contextual information is often limited because of lack of content.

Large Percentage of the traffic is user generated content which is difficult to monetize. Ads are often targeted down to the carrier and device level. e.g., Phones with video capability on Sprint!  
-----But have location based info..... Treat in a e.g., collaborative way
Sizing the US Mobile Web Display Advertising Opportunity

Source: Yankee Group, 2008

Note: eCPM represents the average. Pricing levels vary considerably.
Ad spend will quadruple in 4 years

US Mobile Advertising Spending, 2008-2013 (millions and % change)

- 2008: $648 (35.0%)
- 2009: $760 (17.3%)
- 2010: $995 (30.9%)
- 2011: $1,410 (41.7%)
- 2012: $2,390 (69.5%)
- 2013: $3,330 (39.3%)

Note: Includes mobile message advertising, mobile display advertising and mobile search advertising
Source: eMarketer, February 2009

www.emarketer.com
Outline

• Introduction
• Online advertising background
• Business models
• Creating an online ad campaign
• Technology and Economics
• New Directions
  – Behavioral Targeting
  – Mobile
  – Web 2.0
  – Social Advertising
  – Data Exchanges
• Challenges in online advertising
• Summary
Web 2.0

- Online reviews: [Amazon.com](http://www.amazon.com), [Ebay](http://www.ebay.com), [Epinions.com](http://www.epinions.com)
- Social networking: [Facebook](http://www.facebook.com), [MySpace](http://www.myspace.com), [LinkedIn](http://www.linkedin.com)
- Social media sites: [YouTube](http://www.youtube.com), [Hulu](http://www.hulu.com)
- Tagging/folksonomies systems: [del.icio.us](http://www.del.icio.us), [Digg](http://www.digg.com)
- Online encyclopdedia: [Wikipedia](http://www.wikipedia.org), [Yahoo! Answers](http://www.yahoo.com)
- Online blogs: [Blogger](http://www.blogger.com), [Twitter](http://www.twitter.com)
- Online question-answer forums: [Taskcn](http://www.taskcn.com), [Yahoo! Answers](http://www.yahoo.com)
- CrowdSourcing forums: [Amazon Mechanical Turk](http://www.mturk.com)
- Cloud Computing
- Wikis
- Online Gaming
Online Video Advertising

- Online Video Advertising report examines this nascent but rapidly growing ad channel
  - YouTube signs deal with Hollywood for TV shows and Movies [4/2009]; Ads and possibly payment for premium content
- YouTube will lose $470M in 2009 [Credit Suisse 2009]
Social Network Ad Spend Cools Off

- With the economy taking a toll on even social-networking sites, eMarketer reduced its **global** forecast for advertising in the sector this year to $2.3 billion, according to a WSJ (3/2009).
  - Facebook and MySpace claim $900M revenue in US and over 50% of the world wide spend (WSJ)

![US Online Social Network Advertising Spending, 2008-2013 (millions and % change)](image)
Web 2.0-> Web 3.0

• Very low CPMs
• Taking Web 2.0 to Web 3.0?
  – Behavioral Targeting
  – Social search or a social search engines
    • a type of web search method that determines the relevance of results by considering the interactions or contributions of users
    • How to leverage community and collaboration?
  – Social behavioral targeting?
  – Are new business models required?
  – ??
Web 2.0 Conclusions

• **Social Networks**
  – 79.5 million people—41% of the US Internet user population—visited social network sites at least once a month, an 11% increase over the previous year. By 2013, an estimated 52% of Internet users will be regular social network visitors.
  – Growth in spending on social network advertising is slowing around the world, a result of poor economic conditions and a lack of proven ad models
  – Marketers continue to show interest in social networks but increasingly are looking outside buying ads!!
    • Building fan communities on social network sites.
    • Subscriber –based; pay for premium content
    • [http://www.shutterstock.com/](http://www.shutterstock.com/)

• **Online video -> profitability**

• **Microblogging is growing rapidly; Twitter; Mixi (Japan)**
Outline

• Introduction
• Online advertising background
• Business models
• Creating an online ad campaign
• Technology and Economics
• New Directions
  – Behavioral Targeting
  – Mobile
  – Web 2.0
  – Social Advertising
  – Data Exchanges
• Challenges in online advertising
• Summary
Social Advertising

• **Social advertising** was introduced in the context of community-oriented websites such as Facebook and Digg.

• This form of advertising systematically leverages historically "offline" dynamics, such as peer-pressure, recommendations, and other forms of social influence.

• In the case of Digg, this translates to enabling users to determine what ads appear on the website news streams by voting up or "burying" ads, much the way they can digg or "bury" organic news items.
Social Advertising (at Digg)

Social Online Advertising

Advertiser-Publisher-User-Community

User+Community+
Diggs

Select Ads

Advertiser

Digg Social Advertising © 2009 James G. Shanahan (San Francisco)
Lance Armstrong in hospital after race crash

cnn.com — American cyclist Lance Armstrong, the only man to win the Tour de France seven times, crashed on the first stage of a five-day race in Spain on Monday and was taken to a hospital by ambulance. More... (Other Sports)

29 Comments  Share  Bury  jdm2 made popular 39 min ago

I Think my Social Studies Teacher is Spying on Me

answers.yahoo.com — "his # keeps coming up on the caller id and i see his car outside my house a lot and sometimes he's even talking to my mom inside but when he sees me he gets nervous and says has there to drop off my homework even though i already took it home. i'm thinking about putting up secret cameras in my house cause my friend dave is a computer nerd." More... (Comedy)

72 Comments  Share  Bury  sixmonkey made popular 39 min ago

Mr. Obama: Populist Anger Is Hard to Contain

online.wsj.com — The president could have spoken more responsibly about AIG. A political leader simply stir up a little bit of populism, then turn it off when it gets inconvenient -- not even a leader as eloquent as President Barack Obama. More... (Political Opinion)

87 Comments  Share  Bury  Blinker1315 made popular 49 min ago

EA to launch ‘Sims 3’ on June 2nd

ee.com — Electronic Arts said on Tuesday that The Sims 3, the third full iteration of one of the most successful video game franchises of all time, will hit store shelves on June 2.

29 Comments  Share  Bury

Cookie Monster's Conundrum

newmoderntuxury.ning.com — oh that Cookie Monster, he loves his cookies, but is he willing to give them up to clean his computer? More... (Odd Stuff)

21 Comments  Share  Bury  estodesigns made popular 7 min ago

French women Europe’s thinnest and most worried about weight

The Economist — Sweden recently abolished commercial advertising in 2006. But for women in other European countries, the struggle to look and feel good continues. More... (Europe)

100 Comments  Share  Bury

[http://blog.digg.com/]
Google Connect

- Soon websites that use Friend Connect a new option - add Adsense-like advertising within the Friend Connect and Open Social widgets that they’ve added to their websites.
- Publishers will get a percentage of the revenue generated from the advertising.
- And that’s the big monetization scheme behind Open Social and Friend Connect for Google. And that’s why they call it Friendsense internally.
- And occasionally let it slip to outsiders.

• “Google Friend Connect instantly awakens and strengthens the community that visits your site by enriching it with social features”.

Love Da' Guacamole

Guac of the day

This recipe calls for preparing the guacamole in a molcajete, a mortar and pestle carved from volcanic rock. One of the world's oldest kitchen tools, it was used by the Aztecs in Mexico for grinding corn into flour.

Ingredients:
1 ripe tomato, finely chopped
2 Tbs. finely minced white onion
2 serrano chilies, finely chopped
1 Tbs. fresh lime juice
1/2 tsp. sea salt, plus more as needed
2 large avocados, preferably Haas
2 Tbs. finely minced fresh cilantro

For the garnish:
1 Tbs. small fresh cilantro leaves (optional)
1 Tbs. finely chopped white onion (optional)
1 Tbs. finely chopped ripe tomato (optional)

Directions:
Put the tomato, onion, chilies, lime juice and the 1/2 tsp. salt in a molcajete or small bowl, and smash with a pestle or fork to a coarse paste. Cut the avocados in half, remove the pits and scoop the flesh into the tomato mixture. Add
Outline

• Introduction
• Online advertising background
• Business models
• Creating an online ad campaign
• Technology and Economics
• New Directions
  – Behavioral Targeting
  – Mobile
  – Web 2.0
  – Social Advertising
  – Data Exchanges
• Challenges in online advertising
• Summary
Data Exchanges

- Publishers get paid for their audiences and content
- eCommerce sites get paid for their audiences and transactions
- Data Exchanges
  - Categorize users based on their behaviors
  - Auction off audiences to publishers/ad networks
  - E.g., BlueKai, eXelate and Datran (behavioral exchanges)
    - E.g., BlueKai: Intent mining from eCommerce sites
    - Lotame, SocialMedia and Media6Degrees working the social-media world.
Commercial Intent: Shop/Browse

[Adapted from Lucian Vlad Lita 2009, BlueKai]

RuSSIR 2009, Petrozavodsk, Russia. Online Advertising © 2009 James G. Shanahan (San Francisco)

James.Shanahan_AT_gmail_DOT_com
IN-MARKET AUTO:
- Autos by Make & Model
- Class
  - Compacts & Sub-Compacts
  - Convertibles
  - Family Sedans
  - Green Cars
  - Luxury Cars
  - Midsize Cars
  - Minivans
  - Pickup Trucks
  - Sport Utility Vehicles
  - Station Wagons

IN-MARKET TRAVEL:
- Air Travel
  - International Flyers
  - Domestic Flyers
  - By Destination City
  - By Departure City
  - Length of Stay
  - By Advanced Booking
- Hotels & Lodging
  - By City
  - By Star Rating
  - Length of Stay
  - By Advanced Booking
- Car Rentals
  - By City
  - By Rental Company
  - By Type of Car
  - By Advanced Booking
  - Saturday Rental
- Cruises
  - By Destination
  - By Cruise line
  - By Class
  - Duration
  - Departure Month
- Packages
  - By Domestic Destination
  - By International Destination

IN-MARKET SHOPPING:
- Auto Parts & Accessories
- Bags
- Books & Magazines
- Cameras
- Cell Phones & PDA’s
- Clothing & Accessories
  - Women’s > Jeans, Dresses, Shoes, Outerwear
  - Men’s > Accessories, Shoes, Outerwear, etc
  - Boys
  - Girls
  - Infants & Toddlers
  - Jewelry & Watches, etc.
- Computers
- Consumer Electronics
  - Apple iPod, MP3 players
  - DVD and Home Theater
  - MP3 Accessories
  - Satellite, Cable TV, etc.
- DVD & Movies
- Health
  - Fragrance
  - Hair Care
  - Makeup
- Home & Garden
  - Furniture
  - Appliances
  - Kitchen, etc
- Toys, Games, Video Games

[Lucian Vlad Lita 2009, BlueKai]

© 2009 Blue Kai. Confidential. All Rights Reserved.

RuSSIR 2009, Petrozavodsk, Russia. Online Advertising © 2009 James G. Shanahan (San Francisco)
James.Shanahan_AT_gmail_DOT_com
Mining Commercial Intent

**Commercial Intent**

1: Shops Searches

2: Sells

[Data Exchange Auction](#)

3: Bids

4: Select Winner

5: Winner Deposits Cookie

**Cookie Buyer**

- Publisher
- Media Agency
- Ad Network

---

*RuSSIR 2009, Petrozavodsk, Russia. Online Advertising © 2009 James G. Shanahan (San Francisco) James.Shanahan_AT_gmail_DOT_com*
Context + User Behavior

Context + Behavior Class

CPM = ~$1-$2


CPM = ~$4-15
How does the data exchange works?

• When someone does a search, for example, on Kayak.com for first-class flights to Paris in September, that information can be captured by a cookie, and Kayak.com can sell that cookie using eXelate or BlueKai.

• More targeted
  – A buyer would want that cookie so the company could cut down on wasted ads. Sure, Hilton could blanket sites with its ads, but it would rather show an ad to someone who has searched for flights to Paris recently.
  – “When you see a cookie on that user, you can show them travel to Paris even though they may be on MySpace or The New York Times,” said Mark Zagorski, the chief revenue officer of eXelate.

• Who buys this data?:
  – the Hilton itself or its media agency may buy the information.
  – A publisher may buy data so it can sell more expensive ads to Parisian hoteliers and restaurants.
  – Or the buyer may be an ad network — a company that handles sales for big groups of sites at once.
Outline

• Introduction
• Online advertising background
• Business models
• Creating an online ad campaign
• Technology and Economics
• New Directions
  – Behavioral Targeting
  – Mobile
  – Web 2.0
• Challenges in online advertising
• Summary
Challenges in Online Advertising

- Privacy
- Modeling the user
- Ad placement
- Social responsibility – Long term user value
- Deception (e.g., phishing)
- Green advertising
  - 67% of printing magazines are returned [Sierra Club, AdTech 2008]; inflated subscription rates to fulfill advertiser commitments.
- Fraud
  - Impression Fraud
  - Click fraud
Privacy Concerns (US view)

- “the data game right now is a little bit of the Wild West.”
- All this tracking has raised privacy concerns. Some privacy advocates have asked Congress and the Federal Trade Commission to investigate the issue, seeking clear policies about sensitive data, more information on the way companies are tracking consumers and options for consumers to avoid online tracking.
- So far, the commission has recommended that the industry police itself.
- But Jon Leibowitz, one of the commissioners, warned in February that the industry needed to do a better job or face new laws and regulations.

[http://www.nytimes.com/2009/05/31/business/media/31ad.html?_r=2&em=&pagewanted=print]
Data Exchanges and Privacy

• Both BlueKai and eXelate
  – provide a page where consumers can refuse all targeting,
  – they are allowing consumers to see what information has been collected about them,
    • exelate.com/new/consumers-optoutpreferencemanager.html for eXelate,
    • and tags.bluekai.com/registry for BlueKai.
Fraud on Internet

• Fraud is prevalent on the internet today
  – Publisher side (e.g., domain-name parking, Internet pirates, domain name bandits, cybersquatters, typosquatters)
  – Advertiser side (both click and impression fraud)

• Difficult to quantify
  – Studies/organizations will claim [5, 40]% clicks are fraudulent

• CPA Advertising model directly address click fraud
A Bitter-Sweet Advertising Moment!
Deception

McAfee has detected a potential phishing Web site.

Phishing sites appear to be legitimate, but they request you to provide sensitive information, which can be used to commit fraud.

McAfee recommends that you block this Web site.

Block this Web site.
Allow this Web site.

OK
Click Fraud versus SPAM

• $21.4 billion in U.S. advertising spending, according to an Oct. 16 report by research firm eMarketer. (Google has 32% of the market)

• Average click fraud rate of PPC advertisements on search engine content networks, including Google AdSense and the Yahoo Publisher Network, was 28.1% for Q3/2007.

• The California legislature found that spam cost United States organizations alone more than $13 billion in 2007, including lost productivity and the additional equipment, software, and manpower needed to combat the problem.
Click Fraud

• Click fraud is a type of internet crime that defrauds advertisers
• The act of clicking on an ad (e.g., text or banner) with the intention of either:
  – To inflate a Publisher’s revenue
  – Or manipulating a competitor’s advertising behavior by artificially decreasing their ROI or depleting their budget
• Committed by a person, or by a computer program that imitates a legitimate user of a web browser
• Difficult to quantify
  – Studies/organizations will claim [10, 20]% of clicks are fraudulent
  – Dizzying collection of scams and deceptions that inflate advertising bills for thousands of organizations of all sizes
Conferences/Workshops

• **WWW 2008: WS6 - Targeting and Ranking for Online Advertising**
  – Ewa Dominowska and Vanja Josifovski

• **SIGIR 2008: Information Retrieval in Advertising**
  – Ewa Dominowska, Eugene Agichtein, James G. Shanahan, Evgeniy Gabrilovich

• **DIMACS Workshop on the Boundary between Economic Theory and Computer Science, 10/2007**

• **Click fraud workshop at Stanford (9/2007)**

• **KDD Workshop on Advertising**
  – Focused on the online and offline advertising worlds
  – Cross fertilization of both worlds (digital marketplaces, audience understanding)

• **Many workshops and conference session on online advertising**
  – See WWW, SIGIR, KDD, CIKM
Outline

• Introduction
• Online advertising background
• Business models
• Creating an online ad campaign
• Technology and Economics
• New Directions
• Challenges in online advertising
• Summary

Business, Gold rush
Tech
Hot Areas
Bridging the gap: DIY Software+Data

• OpenX
  – OpenX Ad Server is a powerful open source platform that manages advertising for more than 150,000 websites in over 100 countries. More than 300 billion ads run through OpenX Ad Servers each month. And since it's available as a software download or as a hosted service, you can decide to run it yourself or let us take care of managing the infrastructure for you.

• Solr (Solr is an open source enterprise search server based on the Lucene Java search library, with XML/HTTP and JSON APIs, hit highlighting, faceted search, ...)  

• Data:
  – Simulators; Mechanical Turk; Collaborations with advertising industry
R: open-source statistical software

- **R** is a free, open-source software package/programming language for statistical computing.
  - (commercial version is S-plus)

- **Resources on R** [adapted from http://www.stat.cmu.edu/~cshalizi/350/]
  - The official intro, "An Introduction to R", available online in HTML, PDF
  - John Verzani, "simpleR", in PDF
  - Quick-R. This is primarily aimed at those who already know a commercial statistics package like SAS, SPSS or Stata, but it's very clear and well-organized, and others may find it useful as well.
  - Patrick Burns, The R Inferno. "If you are using R and you think you're in hell, this is a map for you."
  - T. Lumley, "R Fundamentals and Programming Techniques" (large PDF)

- **Books:** [adapted from http://www.stat.cmu.edu/~cshalizi/350/]
  - Braun and Murdoch's *A First Course in Statistical Programming with R* (official site, Powell's), suitable for absolute beginners
  - Venables and Ripley's *Modern Applied Statistics with S* (official site, Powell's), useful but more advanced material
  - John M. Chambers, *Software for Data Analysis: Programming with R* (official site, Powell's) is the best book on writing programs in R.
• MapReduce is a software framework introduced by Google to support distributed computing on large data sets on clusters of computers.

• See [http://en.wikipedia.org/wiki/MapReduce](http://en.wikipedia.org/wiki/MapReduce)

• See Hadoop for open source implementation in Java

• Can be used for simple tasks such as counting, to inverted index creation to PageRank calculation for extremely large datasets (billions of rows).
Homegrown Evaluations

• **Use publicly available datasets**
  – LETOR, Yandex LETOR, Netflix, MSAd data, TREC, CLEF, INEX

• **Simulate**

• **Label data**
  – Use Mechanical Turk

• **Collaborate with advertising industry; intern**
Ad Network Architecture: Forward Market

- Ad upload/SelfServe
  - Creatives Constraints
- Ad CTRs
- Generate Display Schedule
- Segment Real Estate & Pricing
- Yield Management Ad Network
- Clicks/LPs Logs Ratings
- Advertisers
- Publishers
- ADashBoard
- PDashBoard
- Generate AdCode
- WebPage
  - SERP
  - WWW
- Behavioral
  - MLR Ranker
  - Query Proc
  - Analytics
- Ads, LPs
- Landing Pages
  - Features
    - Index
    - TF/IDF
    - (Webgraph)
    - Anchor Text
    - Classes
    - Page Quality
    - Historical
    - Site-level
- Ad Network
- MLAB Test DashBoard
- Users

Estimate Ad CTRs
Generate Display Schedule
Segment Real Estate & Pricing

RuSSIR 2009, Petrozavodsk, Russia. Online Advertising © 2009 James G. Shanahan (San Francisco) James.Shanahan_AT_gmail_DOT_com
Ad Network Architecture: Spot Market

Featurize Ads and LPs

Crawler

Features
- Index
- TF/IDF
- (Webgraph)
- Anchor Text
- Classes
- Page Quality
- ... Historical
- Site-level

Ratings

Ads, LPs

Rank Ads
For a target page and user (ECPM), Compute price (CPX)

Index

Ad

Analytics

Query Logs

Events

Crawl+

(Index, Feat.)

ML

AB Test

DashBoard

Users

Users

Ad upload/ SelfServe
Creatives Constraints

ADashBoard

PDashBoard

Generate AdCode

WebPage

SERP

WWW

Email

....

Advertisers

Publishers

Publishers

Publishers

Publishers

Publishers

Publishers
Executive Summary

- Advertisers deliver a message to consumers via publishers
- Online advertising revenue growing annually @ 10% (from 26%)
  - $23.4 billion in USA (2008), $65B worldwide (10% of overall ad spend)
  - Business Models: CPM, CPC, CPA (not popular yet but will be), dCPM
  - Online advertising is location, location and location ($11B M&A, 2007)
  - …..and data, features, objectives and optimization

- Key directions
  - Forward markets → **Spot Markets**
    - Advertising transformed from a low-tech, “Mad Men” process to highly optimized, mathematical, computer-centric (Wall Street-like) process
  - **Technology**: Economics, IR, ML, OR, Game Theory, Stats, Social Sciences (estimating CTR/AR; auction systems; learning algorithms; targeting display advertising; behavioral targeting)
  - **Click fraud** is the spam of online advertising ($Billion problem)
  - **New Directions**: Web 3.0; games; social advertising; data exchanges
  - **Gap**: perceived gap between academia and industry?
From Mad Men To Wall Street

• Set in New York City, Mad Men begins in 1960 at the fictional Sterling Cooper advertising agency on New York City's Madison Avenue.
Course philosophy

• **Socratic Method**
  – participation strongly encouraged (please state your name and affiliation)

• **Highly interactive and adaptable**
  – Questions welcome!!

• **Lectures emphasize intuition, not rigor or detail**
  – Build on lectures from other faculty
  – Background reading will have rigor & detail

• **Action Items**
  – Read suggested books first (and then papers), read/write Wikipedia, watch/make YouTube videos, take courses, participate in competitions, do internships, network
  – Prototype, publish, participate, simulate
  – Classic (core) versus trendy (applications)
Thanks

EMAIL:
James_DOT_Shahanhan_AT_gmail_DOT_com
Bibliography

• Shanahan James G., 2010, Online Advertising: From “Mad Men” to Wall Street.
• Large-Scale Behavioral Targeting - Ye Chen* Yahoo! Labs; Dmitry Pavlov Yahoo! Labs; John Canny Computer Science Division University of California Berkeley

• Optimal Auction Design in a Multi-unit Environment: The Case of Sponsored Search Auctions. With Michael Schwarz. (December 2006)
• Statistics for Experimenters: Design, Innovation, and Discovery, 2nd Edition by George Box, Stuart Hunter and William G. Hunter, 2005
• David Maxwell Chickering, David Heckerman, Targeted advertising on the Web with inventory management, Interfaces, v.33 n.5, p.71-77, September 2003
• http://www.cis.upenn.edu/~mkearns/teaching/SponsoredSearch/
• S. Pandey, C. Olston, 2006, Handling Advertisements of Unknown Quality in Search Advertising
• TNS Media Intelligence
• Global Entertainment and Media Outlook: 2006-2010
Bibliography

Machine Learning

- Friedman, J. H. "Stochastic Gradient Boosting ." (March 1999b) [software](http://statgen.iop.kcl.ac.uk/bgim/mle/sslike_4.html)

Online Advertising Background:


• OTHER Resources
  – See Proceedings WWW 2008, WWW 2009 for other papers on Ranking ads, economic models etc.
  – See OMMA
  – www.eMarketer.com
THANKS!

The End