Multi-Label Learning with Millions of Labels for Query Recommendation

Rahul Agrawal
Microsoft AdCenter

Archit Gupta
IIT Delhi

Yashoteja Prabhu
Microsoft Research India

Manik Varma
Microsoft Research India
Recommending Advertiser Bid Phrases

Do you know how much you could save on car insurance by switching to GEICO?

Get an online quote: Enter ZIP

Get a quote by phone: 1-800-841-1588

Find a local agent: Enter ZIP

Need other insurance?
- ATV
- Boat
- Commercial Auto
- Condo/Co-op
- Flood
- Homeowners
- ID Theft Protection
- Life
- Mobile Home
- Motorcycle
- Overseas
- Renters
- RV
- Umbrella

Already a policyholder? Log in.

geico auto insurance
geico car insurance
geico insurance
www.geico.com
care.geicos
geico.com
needcheapautoinsurance
wisconsincheapcarinsurancequotes
cheapautoinsuranceflorida
allstatecarinsurancecouponcode
“Absolutely cheapest car insurance”
The GEICO Gecko
@TheGEICOGecko

The official Twitter home of the GEICO Gecko, helping people save hundreds on their car insurance.

Washington, DC · http://www.geico.com

1,542 TWEETS  1,237 FOLLOWING  9,113 FOLLOWERS

The GEICO Gecko @TheGEICOGecko
Feb 23
Catching up on the nominees before tomorrow. I've always been a fan of the "Short Film" category. pic.twitter.com/zjHzpYkXR7
Details

The GEICO Gecko @TheGEICOGecko
Feb 18
Anyone know where the halfpipe is around here? Gotta practice my tailgrab. pic.twitter.com/fvQTy8Fu
Details

geico auto insurance
geico car insurance
geico insurance
www.geico.com
care geicos
geico com
need cheap auto insurance
wisconsin cheap car insurance quotes
cheap auto insurance florida
geico twitter
Recommending Advertiser Bid Phrases

Do you know how much you could save on car insurance by switching to GEICO?

Get an online quote: Enter ZIP

Get a quote by phone: 1-800-841-1588

Find a local agent: Enter ZIP

Need other insurance?

- ATV
- Boat
- Commercial Auto
- Condo/Co-op
- Flood
- Homeowners
- ID Theft Protection
- Life
- Mobile Home
- Motorcycle
- Overseas
- Renters
- RV
- Umbrella

Already a policyholder? Log in.
Learning to Predict a Set of Queries

$f : X \rightarrow 2^Y$

X: Ads

- Ads
- Geico
- Online quote
- Need cheap auto insurance
- Car insurance
- Iphone

Y: Queries

- Italian restaurant
- Geico
- Online quote
- Car insurance
- Iphone
Learning to Predict a Set of Queries

\[ f(\,) \rightarrow \begin{align*} & \bullet \text{need cheap auto insurance} \\ & \bullet \text{geico} \\ & \bullet \text{car insurance} \end{align*} \]
Multi-Label Learning Challenges

- Infinite number of labels (queries)
- Training data acquisition
- Efficient training
- Cost of prediction

\[ f(\cdot) \]

- need cheap auto insurance
  - geico
  - car insurance
Binary Classification & Ranking

\[ h : (X, Y) \rightarrow \{\times, \checkmark\} \]

\[ h(\text{geico}) \rightarrow \checkmark \]
\[ h(\text{iphone}) \rightarrow \times \]

- Infinite number of labels (queries)
- Training data acquisition
- Efficient training
- Cost of prediction
Binary Classification

\[ h : (X, Y) \rightarrow \{\times, \checkmark\} \]

- Infinite number of labels (queries)
- Training data acquisition
- Efficient training
- Cost of prediction
Binary Classification – KEX

\[ h : (X, Y) \rightarrow \{\times, \checkmark\} \]

- Infinite number of labels (queries)
- Training data acquisition
- Efficient training
- Cost of prediction

Do you know how much you could save on car insurance by switching to GEICO?

- switching to geico
- geico online quote
- car insurance

GEICO

Get an online quote: Enter ZIP Go
Get a quote by phone: 1-800-GEICO-13000
Find a local agent: Enter ZIP Go

Need other insurance?
- ATV
- Boat
- Commercial Auto
- Condo/Rental
- Flood
- Homeowners
- ID Theft Protection
- Life
- Motorcycle
- Motor Home

Overseas
- Recreational Vehicles
- RV
- Umbrella

Already a policyholder? Log In.
Do you know how much you could save on car insurance by switching to GEICO?

Get an online quote: Enter ZIP

Get a quote by phone: 1-800-841-1588

Find a local agent: Enter ZIP

Need other insurance?

- ATV
- Boat
- Commercial Auto
- Condo/Co-op
- Flood
- Homeowners
- ID Theft Protection
- Life
- Mobile Home
- Motorcycle
- Overseas
- Renters
- RV
- Umbrella

Already a policyholder? Log in.
Query Recommendations by KEX

Do you know how much you could save on car insurance by switching to GEICO?

$h(\text{GEICO}, \text{car insurance}) \rightarrow ?$

$h(\text{GEICO}, \text{iphone}) \rightarrow ?$

Flood  Motorcycle

Already a policyholder? Log in.

© 1996-2012 GEICO
Query Recommendations by KEX

Simone & Sylvia

plastic ponies
simone
plastics
clothing and accessories
sylvia
pony clothing
couture
playground
plastic recycling
children's clothing
Multi-Label Learning Formulation

\[ f : X \rightarrow 2^Y \]

X: Ads

Y: Queries

- need cheap auto insurance
- geico
- online quote
- car insurance
- iphone
- italian restaurant
Learning with Millions of Labels

$f : X \rightarrow 2^Y$

X: Ads

Y: 10 Million Queries

- italian restaurant
- need cheap auto insurance
- geico online quote
- car insurance
- iphone
We develop Multi-Label Random Forests with logarithmic prediction costs that make predictions in a few milliseconds.

We train on 200 M points, 100 M categories and 10 M features in 28 hours on a grid with 1000 compute nodes.

We develop a tree growing criterion which learns from positive data alone.

We generate training data automatically from click logs.

We develop a sparse SSL formulation to infer beliefs about the state of missing and noisy labels.
No annotator can mark all the relevant labels for a data point.

We have missing labels during
- Training
- Validation
- Testing.

Even fundamental ML techniques such as validation can go awry.

One can’t design error metrics invariant to missing labels.
Training Data and Features

TF-IDF Bag of Words Features

iPhone color material

TF-IDF Bag of Words Features
## Training Labels

<table>
<thead>
<tr>
<th>Case for iPhone</th>
<th>Best iPhone Case</th>
<th>Apple iPhone 3G Metallic Slim Fit Case</th>
<th>Best iPhone NN4 Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>iPhone Cases</td>
<td>Best iPhone Cases</td>
<td>Apple iPhone 4G Cases</td>
<td>Black White Premium Bumper Case Apple iPhone NN4 ATT</td>
</tr>
<tr>
<td>Best iPhone NN4 Case</td>
<td>Case iPhone</td>
<td>Apple iPhone 4G Premium Soft Silicone Rubber Black Phone Protector Skin Cover Case</td>
<td>Bunny Rabbit Silicone Case Skin iPhone NN4 Stand Tail Holder</td>
</tr>
<tr>
<td>iPhone 3GS Cases</td>
<td>Otterbox Universal Defender Case iPhone NN4 Black Silicone Black Plastic</td>
<td>Apple iPhone NN4 Cases</td>
<td>iPhone Case</td>
</tr>
<tr>
<td>iPhone 4S Case</td>
<td>Sena iPhone Cases</td>
<td>Belkin Grip Vue Tint Case iPhone NN4 Clear</td>
<td>iPhone 4G Cases</td>
</tr>
<tr>
<td>iPhone Case</td>
<td>Speck iPhone Case</td>
<td>Best Case iPhone 4S</td>
<td>iPhone 4GS Cases</td>
</tr>
<tr>
<td>iPhone NN4 Case</td>
<td>Switcheasy Neo Case iPhone 3G Black</td>
<td>Best Case iPhone NN4</td>
<td>iPhone 4S Defender Series Case</td>
</tr>
<tr>
<td>3G iPhone Cases</td>
<td>Waterproof iPhone Case</td>
<td>Best iPhone 3G Cases</td>
<td>iPhone Case Design</td>
</tr>
<tr>
<td>Apple iPhone Cases</td>
<td>Waterproof iPhone Cases</td>
<td>Best iPhone 4S Case</td>
<td>iPhone Cases 3G</td>
</tr>
<tr>
<td>Best iPhone 3G Case</td>
<td>AmazonBasics Protective TPU Case Screen Protector ATT Verizon iPhone NN4 iPhone 4S Clear</td>
<td>Best iPhone 4S Cases</td>
<td>iPhone Cases 4G</td>
</tr>
</tbody>
</table>
Training Labels
## Missing and Noisy Labels

<table>
<thead>
<tr>
<th>best italian restaurants philadelphia</th>
<th>italian restaurant chains</th>
</tr>
</thead>
<tbody>
<tr>
<td>italian restaurants</td>
<td>italian restaurant connecticut</td>
</tr>
<tr>
<td>italian restaurant</td>
<td>italian restaurant district columbia</td>
</tr>
<tr>
<td>italian restaurants arkansas</td>
<td>thai restaurant</td>
</tr>
<tr>
<td>italian restaurants connecticut</td>
<td>thai restaurants</td>
</tr>
<tr>
<td>italian restaurants idaho</td>
<td>restaurants</td>
</tr>
<tr>
<td>italian restaurants phoenix</td>
<td>mexican restaurants</td>
</tr>
</tbody>
</table>
Missing and Noisy Labels
Biased Training Data

- Most labels will have very few positive training examples

Zipf's Law
Multi-Label Prediction Costs

- Linear prediction costs are infeasible

1-vs-All Classification

geico

car insurance

iphone cases

pizza
Label and Feature Space Compression

10M Dimensional Label Space
- car
- motor
- vehicle
- auto
- iphone
cases
- iphone
case
- case
- iphone

6M Dimensional Feature Space
- Car
Ads
- iphone
Case Ads

1K Dimensional Embedding Space
Hierarchical Prediction

- Prediction in logarithmic time
Gating Tree Based Prediction

- Prediction in logarithmic time
Ensemble of Randomized Gating Trees
We seek classifiers and optimization algorithms that
- Are massively parallelizable
- Don’t need to load the feature vectors (1 Tb) into RAM
- Don’t need to load the label matrix (100 Gb) into RAM

### Efficient Training

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of training points</td>
<td>200 Million</td>
</tr>
<tr>
<td>Number of labels</td>
<td>100 Million</td>
</tr>
<tr>
<td>Dimensionality of feature vector</td>
<td>10 Million</td>
</tr>
<tr>
<td>Number of cores</td>
<td>500 – 1000</td>
</tr>
<tr>
<td>RAM per core</td>
<td>2 Gb</td>
</tr>
<tr>
<td>Training time</td>
<td>28 hours</td>
</tr>
</tbody>
</table>
Multi-Label Random Forests

- The splitting cost needs to be calculated in a $2^{10M}$ space

Is the word “insurance” present?
Learning from Positively Labeled Data

• Split condition: $x_{f^*} > t^*$

$$(f^*, t^*) = \arg\min_{f,t} n_l \sum_k p_l(l_k)(1 - p_l(l_k)) + n_r \sum_k p_r(l_k)(1 - p_r(l_k))$$

$$p(l_k) = \sum_i p(l_k|ad_i)p(ad_i)$$
Multi-Label Random Forests

\[(x_1, y_1) = \{l_2, l_3\}\]

\[(x_2, y_2) = \{l_1, l_3\}\]

\[(x_3, y_3) = \{l_1, l_2, l_3\}\]

\[p(y)\]
## Query Recommendation Data Sets

- Data set statistics

<table>
<thead>
<tr>
<th>Data Set</th>
<th># of Training Points (M)</th>
<th># of Test Points (M)</th>
<th># of Dimensions (M)</th>
<th># of Labels (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia</td>
<td>1.53</td>
<td>0.66</td>
<td>1.89</td>
<td>0.97</td>
</tr>
<tr>
<td>Ads1</td>
<td>8.00</td>
<td>0.50</td>
<td>1.58</td>
<td>1.22</td>
</tr>
<tr>
<td>Web</td>
<td>40.00</td>
<td>1.50</td>
<td>2.62</td>
<td>1.22</td>
</tr>
<tr>
<td>Ads2</td>
<td>90.00</td>
<td>5.00</td>
<td>5.80</td>
<td>9.70</td>
</tr>
</tbody>
</table>
Performance Evaluation – Precision@k

• We use loss functions where the penalty incurred for predicting the real (but unknown) ground truth is never more than that of predicting any other labelling

\[ L(y^*, y_{\text{observed}}) \leq L(y, y_{\text{observed}}) \quad \forall y \in Y \]

• Hamming Loss

• Precision at \( k \)  

• We found Precision at 10 to be robust for our application.
Query Recommendation Results

Percentage of top 10 predictions that were clicked queries
Query Recommendation Results

Percentage of top 10 predictions that were relevant
Do you know how much you could save on car insurance by switching to GEICO?

Get an online quote: Enter ZIP

Get a quote by phone: 1-800-841-1588

Find a local agent: Enter ZIP

Need other insurance?
ATV
Boat
Commercial Auto
Condo/Co-op
Flood
Homeowners
ID Theft Protection
Life
Mobile Home
Motorcycle
Overseas
Renters
RV
Umbrella

Already a policyholder? Log in.

GEICO © GEICO
<table>
<thead>
<tr>
<th>KEX</th>
<th>MLRF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>geico auto insurance</td>
</tr>
<tr>
<td></td>
<td>geico car insurance</td>
</tr>
<tr>
<td></td>
<td>geico insurance</td>
</tr>
<tr>
<td></td>
<td><a href="http://www.geico.com">www.geico.com</a></td>
</tr>
<tr>
<td></td>
<td>care geicos</td>
</tr>
<tr>
<td></td>
<td>geico.com</td>
</tr>
<tr>
<td></td>
<td>need cheap auto insurance</td>
</tr>
<tr>
<td></td>
<td>wisconsin cheap car insurance quotes</td>
</tr>
<tr>
<td></td>
<td>cheap auto insurance florida</td>
</tr>
<tr>
<td></td>
<td>all state car insurance coupon code</td>
</tr>
</tbody>
</table>
### Domino’s Pizza

<table>
<thead>
<tr>
<th>KEX</th>
<th>MLRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>dominos</td>
<td>domino pizza</td>
</tr>
<tr>
<td>dominos pizza</td>
<td>domino pizza</td>
</tr>
<tr>
<td>domino pizza</td>
<td>domino pasta bowls</td>
</tr>
<tr>
<td>domino pasta bowls</td>
<td>domino pizza coupons</td>
</tr>
<tr>
<td>domino pizza coupons</td>
<td>domino pizza deals</td>
</tr>
<tr>
<td>domino pizza deals</td>
<td>domino pizza locations</td>
</tr>
<tr>
<td>domino pizza locations</td>
<td>domino pizza menu</td>
</tr>
<tr>
<td>domino pizza menu</td>
<td>domino pizza online</td>
</tr>
</tbody>
</table>
plastic pony  playground couture

Our line of infant and children's clothing and accessories feature organic and natural fabrics, designed and styled with retro inspirations that will have your children showered with compliments.

Plastic Pony clothing is of superior quality. Made from ultra-comfortable, cozy materials.

Plastic Pony fashions are perfectly suited for the active lives of children. As for their parents, Plastic Pony clothes are very easy to maintain.

Painstaking care is also taken with respect to the patterns, ensuring great freedom of movement.

We make children's clothing starting at 3 months up to size 12.
<table>
<thead>
<tr>
<th>KEX</th>
<th>MLRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>plastic ponies</td>
<td>toddlers clothes</td>
</tr>
<tr>
<td>simone</td>
<td>toddlers clothing</td>
</tr>
<tr>
<td>plastics</td>
<td>toddler costumes</td>
</tr>
<tr>
<td>clothing and accessories</td>
<td>children clothes sale</td>
</tr>
<tr>
<td>sylvia</td>
<td>children clothes</td>
</tr>
<tr>
<td>pony clothing</td>
<td>designer children clothes</td>
</tr>
<tr>
<td>couture</td>
<td>cute children clothes</td>
</tr>
<tr>
<td>playground</td>
<td>retro clothing</td>
</tr>
<tr>
<td>Plastic recycling</td>
<td>retro baby clothes</td>
</tr>
<tr>
<td>children's clothing</td>
<td>baby clothing</td>
</tr>
</tbody>
</table>
Deliveries Everyday To The Seattle Area

ORDER NOW

Business Hours
Monday-Thursday
8:30am - 5:30pm

Friday
8:30am - 6:30pm

Saturday
9am - 5pm

Sunday
10am - 2pm
## KCS Flowers

<table>
<thead>
<tr>
<th>KEX</th>
<th>MLRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>funeral flowers</td>
<td>flowers delivery</td>
</tr>
<tr>
<td>sympathy funeral flowers</td>
<td>funeral arrangements</td>
</tr>
<tr>
<td>web home</td>
<td>birthday flowers</td>
</tr>
<tr>
<td>bleitz funeral home</td>
<td>funeral flowers</td>
</tr>
<tr>
<td>funeral flowers discount</td>
<td>funeral planning</td>
</tr>
<tr>
<td>yarington's funeral home</td>
<td>flowers valentines</td>
</tr>
<tr>
<td>harvey funeral home</td>
<td>free delivery flowers</td>
</tr>
<tr>
<td>green lake funeral home</td>
<td>cheap flowers</td>
</tr>
<tr>
<td>howden kennedy funeral home</td>
<td>florists</td>
</tr>
<tr>
<td>arranging flowers</td>
<td>cheap flowers funeral</td>
</tr>
</tbody>
</table>
Customizable Apparel

The perfect fit at a perfect price.

- Great for on the job and off
- Made from superior quality fabric
- Choose from 100s of designs
- No minimum orders or setup fees

Printed T-shirts
With our fully customizable, 100% pre-shrunk cotton T-shirts, you'll be able to advertise and promote your business wherever you go.

Starting at $11.99

Get Started

Printed Ladies' T-shirts
Our most feminine fit – and the most stylish way to display your logo or message.

Starting at $13.99

Get Started

Embroidered Men's Polo Shirts
100% cotton, high-quality polos you'll be proud to put your name on. They look and feel great, wash after wash.

Starting at $19.99

Get Started

Embroidered Ladies' Polo Shirts
The same high-quality look and feel as our men's polo, but slightly more tailored for a flattering fit.

Starting at $19.99

Get Started
### Vistaprint Designer T-Shirts

<table>
<thead>
<tr>
<th>KEX</th>
<th>MLRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>embroidered apparel</td>
<td>custom t shirts</td>
</tr>
<tr>
<td>custom apparel</td>
<td>funny t shirts</td>
</tr>
<tr>
<td>readymade apparel</td>
<td>hanes beefy t shirts</td>
</tr>
<tr>
<td>customizable</td>
<td>hanes t shirts</td>
</tr>
<tr>
<td>apparel</td>
<td>long sleeve t shirts</td>
</tr>
<tr>
<td>customizable apparel</td>
<td>personalized t shirts</td>
</tr>
<tr>
<td>leading print</td>
<td>printed t shirts</td>
</tr>
<tr>
<td>online business cards</td>
<td>retro gamer t shirts</td>
</tr>
<tr>
<td>apparel and accessories</td>
<td>t shirts</td>
</tr>
<tr>
<td>own text</td>
<td>buy custom t shirts</td>
</tr>
</tbody>
</table>
OTHERS WHO SWITCHED SAVED AN AVERAGE OF

$294* PER YEAR ON AUTO INSURANCE

SEE HOW MUCH YOU CAN SAVE. GET YOUR ONLINE QUOTE NOW!

*Average annual savings based on savings reported by customers from many states who called our call center between 1/11 and 12/11 and switched to MetLife Auto & Home where our quoted premium was less than the disclosed prior carrier's premium. Source: MetLife Auto & Home internal research (2012).

Privacy Policy
# Metlife Auto Insurance

<table>
<thead>
<tr>
<th>KEX</th>
<th>MLRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>metlife auto home insurance</td>
<td>metlife auto insurance</td>
</tr>
<tr>
<td>auto home insurance</td>
<td>auto Insurance</td>
</tr>
<tr>
<td>auto insurance</td>
<td>car Insurance</td>
</tr>
<tr>
<td>massachusetts</td>
<td>automobile Insurance</td>
</tr>
<tr>
<td>metlife agent</td>
<td>geico insurance</td>
</tr>
<tr>
<td>driver discount</td>
<td>cheap car insurance</td>
</tr>
<tr>
<td>additional cost</td>
<td>metlife auto</td>
</tr>
<tr>
<td>saving benefits</td>
<td>insurance broker</td>
</tr>
<tr>
<td>car discount</td>
<td>insurance</td>
</tr>
<tr>
<td>auto quote</td>
<td>home insurance</td>
</tr>
</tbody>
</table>
WANTA THAI CUISINE

This page is under construction. Please visit us at Facebook.com/WantaThai.

WELCOME TO WANTA THAI CUISINE

Wanta Thai Cuisine is a contemporary Thai eatery that transports you to friendly Thai culture. We serve healthy and quality food for a reasonable price, along with generally good service at a convenient location. Visit us to prove why we are one of the best Thai restaurant in Redmond (near Bellevue), WA.

Wanta is a Pali word meaning as showing respect. When Thai people meet, we greet and pay homage with gentle Wanta manner to each other (referred to as the Wai in Thai). Guests will appreciate the traditional Thai service and world-renowned hospitality. You will enjoy delicious Thai food along with humble service in a warm family-friendly atmosphere. We respect customer’s time as we prioritize that your time is money. During the lunch time, we accommodate and welcome your rush hour. We serve quickly food, not fast food. Whereas dinner time, we associate with fine relationship. We make the atmosphere comfy so customers can enjoy the excellent food as in their own living room.

BUSINESS HOURS

Mon-Thu: 11 AM - 9 PM
Fri: 11 AM - 9:30 PM
Sat: 12 PM - 9:30 PM
Sun: 12 PM - 9 PM
## Wanta Thai Restaurant

<table>
<thead>
<tr>
<th>KEX</th>
<th>MLRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>authentic thai restaurant</td>
<td>thai restaurant</td>
</tr>
<tr>
<td>delicious thai food</td>
<td>thai restaurants</td>
</tr>
<tr>
<td>thai cuisine</td>
<td>mexican restaurants</td>
</tr>
<tr>
<td>thai restaurant</td>
<td>cheap hotels</td>
</tr>
<tr>
<td>thai food</td>
<td>hotels</td>
</tr>
<tr>
<td>wanta</td>
<td>fast food restaurants</td>
</tr>
<tr>
<td>best thai restaurant</td>
<td>restaurants coupons</td>
</tr>
<tr>
<td>thai eateries</td>
<td>best web hosting restaurants</td>
</tr>
<tr>
<td>thai</td>
<td>vegetarian foods</td>
</tr>
<tr>
<td>contemporary thai</td>
<td>new york restaurants</td>
</tr>
</tbody>
</table>
Baked Ziti and Sausage

Italian sausage, Pomodoro sauce and ziti topped with a layer of provolone and parmesan cheeses.

Reserve Your Table
<table>
<thead>
<tr>
<th>best italian restaurants philadelphia</th>
<th>italian restaurant chains</th>
</tr>
</thead>
<tbody>
<tr>
<td>italian restaurants</td>
<td>italian restaurant connecticut</td>
</tr>
<tr>
<td>italian restaurant</td>
<td>italian restaurant district columbia</td>
</tr>
<tr>
<td>italian restaurants arkansas</td>
<td>thai restaurant</td>
</tr>
<tr>
<td>italian restaurants connecticut</td>
<td>thai restaurants</td>
</tr>
<tr>
<td>italian restaurants idaho</td>
<td>restaurants</td>
</tr>
<tr>
<td>italian restaurants phoenix</td>
<td>mexican restaurants</td>
</tr>
</tbody>
</table>
We have been in Waterford since 2004.
Best Thai food in racine county.
We make it fresh and tasty.

Business Hour
Monday  4:00pm........8:30pm
Tuesday-Friday   11:30am......8:30pm
Saturday  4:00pm......9:00pm
Close on Sunday
Master and Visa Card accepted
No personal check please
Compensating for Missing Labels

- Progressive insurance
- Allstate auto insurance
- American family insurance
- Case-mate phone cases
- Esurance
- Auto insurance quotes
- Maggiano’s restaurant

0.5 0.7 0.8 0.9
Training on Belief Vectors

\[(x_1, f_1), (x_2, f_2), (x_3, f_3)\]

\[x_1, y_1 = \{l_2, l_3\}, f_1\]

\[x_2, y_2 = \{l_1, l_3\}, f_2\]

\[x_3, y_3 = \{l_1, l_2, l_3\}, f_3\]

\[p(f)\]
Sparse Semi-Supervised Learning

- Graph-based SSL optimizes label belief smoothness and fidelity to original labels

\[
F^* = \min_F \frac{1}{2} \text{Tr}(F^t (I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}) F) + \frac{\lambda}{2} ||F - Y||^2
\]

s.t. \[|F|_0 \leq K\]

- **\(W_{M \times M}\)**: Document-document similarity matrix
- **\(D_{M \times M}\)**: Diagonal matrix representing the row sums of \(W\)
- **\(Y_{M \times L}\)**: 0/1 label matrix
- **\(F_{M \times L}\)**: Real valued label belief matrix
- **\(\lambda\)**: Trade-off Hyperparameter
- **\(M\)**: Number of documents
- **\(L\)**: Number of labels
- **\(K\)**: Sparsity constant
Sparse Semi-Supervised Learning

- Graph-based SSL optimizes label belief smoothness and fidelity to original labels

\[ F^* = \min_F \frac{1}{2} \sum_{i=1..L} \sum_{j=1..M} \omega_{jl} \left( \frac{F_{ij}}{\sqrt{D_{jj}}} - \frac{F_{il}}{\sqrt{D_{ll}}} \right)^2 + \frac{\lambda}{2} \sum_{i=1..M} \left( F_{ij} - Y_{ij} \right)^2 \]

s.t. \( |F|_0 \leq K \)

\[ W_{MXM} \quad \text{Document-document similarity matrix} \]
\[ D_{MXM} \quad \text{Diagonal matrix representing the row sums of } W \]
\[ Y_{MXL} \quad \text{0/1 label matrix} \]
\[ F_{MXL} \quad \text{Real valued label belief matrix} \]
\[ \lambda \quad \text{Trade-off Hyperparameter} \]
\[ M \quad \text{Number of documents} \]
\[ L \quad \text{Number of labels} \]
\[ K \quad \text{Sparsity constant} \]
Iterative Hard Thresholding

- Sparse SSL formulation

\[
F^* = \min_F J(F) = \frac{1}{2} \text{Tr}(F^t \left( I - \frac{1}{2} W D^{-\frac{1}{2}} D^{-\frac{1}{2}} \right) F) + \frac{\lambda}{2} ||F - Y||^2
\]

s.t. \( |F|_0 \leq K \)

- The iterative hard thresholding algorithm converges to a global/local optimum

\[
F_0 = Y
\]

\[
F_{t+\frac{1}{2}} = \frac{1}{\lambda+1} D^{-\frac{1}{2}} W D^{-\frac{1}{2}} F_t + \frac{\lambda}{\lambda+1} Y
\]

\[
F_{t+1} = \text{Top}_K(F_{t+\frac{1}{2}})
\]
Iterative Hard Thresholding

- If $Y_{ij} \in \{0, 1\}$ and $W$ is positive definite then
  - The sequence $F_0, F_1, ...$ converges to a stationary point $F^*$.

- $J(F_0) \geq J(F_1) \geq \cdots \geq J(F^*)$

- If $|F^*|_0 < K$ then $F^*$ is a globally optimal solution

- If $|F^*|_0 = K$ then $F^*$ is a locally optimal solution

$$J(F^*) - J(F^+) \leq \min\left(\frac{\lambda}{2}(K + |Y|_0), \frac{\lambda + 1}{2}(M - K)\alpha_K(F^*)\sqrt{|Y|_0}\right)$$
## Semi-Supervised Learning Results

- Precision@10 as judged by automatically generated click labels as well as by human experts.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Click Labels (%)</th>
<th>Human Verification (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLRF</td>
<td>MLRF+SSL</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>15.72</td>
<td>18.53</td>
</tr>
<tr>
<td>Ads1</td>
<td>18.13</td>
<td>19.88</td>
</tr>
<tr>
<td>Bing</td>
<td>22.51</td>
<td>25.32</td>
</tr>
<tr>
<td>Ads2</td>
<td>15.91</td>
<td>17.12</td>
</tr>
</tbody>
</table>
Query Expansion Results

- Query expansion techniques can help both KEX and MLRF

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Click Labels (%)</th>
<th>Human Verification (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLRF+SSL+KSP</td>
<td>KEX+KSP</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>18.01</td>
<td>10.81</td>
</tr>
<tr>
<td>Ads1</td>
<td>21.54</td>
<td>12.38</td>
</tr>
<tr>
<td>Web</td>
<td>26.66</td>
<td>19.88</td>
</tr>
<tr>
<td>Ads2</td>
<td>19.24</td>
<td>14.35</td>
</tr>
</tbody>
</table>
Query Recommendation Results

- Edit distance  [Ravi et al. WSDM 2010]

<table>
<thead>
<tr>
<th>Data Set</th>
<th>KEX</th>
<th>KEX+KSP</th>
<th>MLRF</th>
<th>MLRF+SSL</th>
<th>MLRF+SSL+KSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia</td>
<td>0.81</td>
<td>0.78</td>
<td>0.71</td>
<td>0.66</td>
<td>0.63</td>
</tr>
<tr>
<td>Ads1</td>
<td>0.83</td>
<td>0.76</td>
<td>0.71</td>
<td>0.65</td>
<td>0.61</td>
</tr>
<tr>
<td>Web</td>
<td>0.73</td>
<td>0.68</td>
<td>0.65</td>
<td>0.62</td>
<td>0.58</td>
</tr>
<tr>
<td>Ads2</td>
<td>0.77</td>
<td>0.73</td>
<td>0.69</td>
<td>0.63</td>
<td>0.59</td>
</tr>
</tbody>
</table>
Conclusions

- Query recommendation can be posed as multi-label learning.
- Learning with millions of labels can be tractable and accurate.
- Other applications
  - Query expansion.
  - Document and ad relevance and ranking.
  - Fine-grained query intent classification.
Acknowledgements

• Deepak Bapna
• Prateek Jain
• A. Kumaran
• Mehul Parsana
• Krishna Leela Poola
• Adarsh Prasad
• Varun Singla
Advantages of an ML Approach

• Can generalize to other domains such as images on Flickr or videos on YouTube.
System Architecture

- We leverage the Map/Reduce framework.
- Trees are grown in parallel breadth-wise.
- Number of compute nodes
  - Evaluators 500
  - Combiners 100
  - Maximizers 25
- Our objective is to balance the compute load across machines while minimizing data flow.

\[
\begin{align*}
X_1, Y_1 & \to X_N, Y_N \\
X_{N+1}, Y_{N+1} & \to X_{2N}, Y_{2N} \\
X_{2N+1}, Y_{2N+1} & \to X_{3N}, Y_{3N} \\
X_{3N+1}, Y_{3N+1} & \to X_{4N}, Y_{4N}
\end{align*}
\]
Evaluators

- **Input**
  - N training instances
  - Set of keys – Tree ID, Node ID, Feature ID and threshold

- **Output**
  - Partial label distributions for the keys

- **Computation**
  - N * # of keys
Combiners

• Input
  • Partial label distributions for assigned keys

• Output
  • Objective function values for the keys.

• Computation
  • # of keys * Avg # of Evaluators / key * # of labels in the distribution for the key.
Maximizers

- **Input**
  - Objective function values for assigned keys

- **Output** –
  - Optimal feature and threshold for assigned nodes in trees.

- **Computation**
  - \( \text{# of keys} \times \text{Avg # of features per key} \times \text{Avg # of thresholds per feature} \)