Query Understanding in Web Search
- by Large Scale Log Data Mining and Statistical Learning

Hang Li
Microsoft Research Asia

Joint Work with Colleagues, Interns, Collaborators
Web Search is Part of Our Life
Search System = `Black Boxes’
Advanced Web Search Technologies Are Used...

- Natural Language Processing
- Information Retrieval
- Data Mining
- Statistical Learning
- Large Scale Distributed Computing
Web Search Relies on NLP and IR

• Query Understanding
  – Classification, structure prediction, topic modeling, similarity learning

• Document Understanding
  – Classification, structure prediction, topic modeling, learning on graph

• Query Document Matching
  – Language model, similarity learning

• Ranking
  – Learning to rank

• User Understanding
  – Classification, topic modeling
Query Understanding

• Input: query
• Output: query representation
  – Refined query (e.g., spelling error correction)
  – Similar queries
  – Categories
  – Topics
  – Key phrases
  – Named entities
This Talk = Query Understanding
Talk Outline: Three Projects

• LOGAL: Search and Browse Log Mining Platform
• Semantic Matching: Improving Tail Query Relevance
• Context aware Search: Better Search Using Context Information
PROJECT: LOGAL (LOG OBJECT GALLERY)

Joint work with Daxin Jiang, Xiaohui Sun
LOGAL

Search and Browse Log Mining Platform
Data Structure of Search/Browse Logs

• Various types of data
• Complex relationship among data objects
  – Hierarchical relationship
  – Sequential relationship

Users → Sessions → Queries

Search result pages

Srch/Ads clicks

Follow-up clicks
Rich Log Mining Applications

Log Mining Applications

Search Applications

Document Understanding

Query Understanding

Document Annotation

Query Expansion

Query Suggestion

Query Substitution

Query Classification

Keyword Generation

Document Classification

Document Summarization

Search Results Clustering

User Understanding

Search UI Design

Behavior Targeting

Personalized Search

User Satisfaction Prediction

Query-Doc Matching

Contextual Advertising

Ad Click-Through Prediction

Web Site Recommendation

Document & Ad (re-)Ranking

Search Results Diversification
The Problem

A huge gap between the data and the applications.
Each researcher or developer
1. Has to access the \textit{raw} log data directly
2. Has to build the application \textit{from scratch}

\textbf{Very difficult to build large-scale log mining applications}
Log Data Mining Platform

App Level
- Query Understanding
- Document Understanding
- User Understanding
- Query Doc Matching

Middle Level
- Data Platform
- Log Objects Gallery (LOGAL)

Raw Data Level
- Raw Logs
  - Search Log
  - Toolbar Data
  - Ads. Log
  - Web Site Log
Query Histogram

<table>
<thead>
<tr>
<th>Query</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>facebook</td>
<td>3,157 K</td>
</tr>
<tr>
<td>google</td>
<td>1,796 K</td>
</tr>
<tr>
<td>youtube</td>
<td>1,162 K</td>
</tr>
<tr>
<td>myspace</td>
<td>702 K</td>
</tr>
<tr>
<td>facebook com</td>
<td>665 K</td>
</tr>
<tr>
<td>yahoo</td>
<td>658 K</td>
</tr>
<tr>
<td>yahoo mail</td>
<td>486 K</td>
</tr>
<tr>
<td>yahoo com</td>
<td>486 K</td>
</tr>
<tr>
<td>ebay</td>
<td>486 K</td>
</tr>
<tr>
<td>facebook login</td>
<td>445 K</td>
</tr>
</tbody>
</table>

Example applications:
- Query auto completion
- Query suggestion
- Query analysis: temporal changes of query frequency
Click-through Bipartite

• Example applications
  – Document (re-)ranking
  – Search results clustering
  – Web page summarization
  – Query suggestion
Click Pattern

### Query

<table>
<thead>
<tr>
<th>Pattern 1 (count)</th>
<th>Pattern 2 (count)</th>
<th>Pattern n (count)</th>
</tr>
</thead>
<tbody>
<tr>
<td>× Doc 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>× ...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>× ...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>× Doc N</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| × Doc 1 |
| × Doc 2 |
| × ... |
| ... |
| ... |
| ... |
| × ... |
| Doc N |

- **Example applications**
  - Estimate relevance of document to query
  - Predict users’ satisfaction
  - Query classification (informational vs navigational)
Session Pattern

Example applications
- Doc (re-)ranking
- Query suggestion
- Site recommendation
- User satisfaction prediction

User activities in a session
- Query
  - Srch click
  - Ads click...
- Click:
- Browse
- Srch click: search click
- Ads click: advertisement click
PROJECT: SEMANTIC MATCHING

Joint work with Gu Xu, Jun Xu, Jingfang Xu
Semantic Matching

Improving Tail Query Relevance
Different Queries Can Represent Same Intent
“Distance between Sun and Earth” - Luke DeLorme

- "how far" earth sun
- "how far" sun
- "how far" sun earth
- average distance earth sun
- average distance from earth to sun
- average distance from the earth to the sun
- distance between earth & sun
- distance between earth and sun
- distance between the earth and sun
- distance between the earth and the sun
- distance between the sun and earth
- distance between the sun and the earth
- distance earth and sun
- distance earth from sun
- distance earth is from the sun
- distance earth sun
- distance earth to sun
- distance earth to the sun
- distance from earth to sun
- distance from earth to the sun
- distance from the earth to sun
- distance from the earth to the sun
- distance from earth to the sun
- distance from sun to earth
- distance from sun to the earth
- distance from the sun to earth
- distance from the sun to the earth
- distance of earth from sun
- distance of earth from the sun
- distance of earth to sun
- distance of earth to the sun
- distance of the earth from sun
- distance of the earth from the sun
- distance of the sun from earth
- distance of the sun from the earth
- distance of the sun to earth
- distance of the sun to the earth
- distance earth and sun distance
- how far away is the sun from earth
- how far away is the sun from the earth
- how far earth from sun
- how far earth is from the sun
- how far earth sun
- how far from earth is the sun
- how far from earth to sun
- how far from the earth to sun
- how far from the sun is earth
- how far from the sun is the earth
- how far is earth away from the sun
- how far is earth from sun
- how far is earth from the sun
- how far is earth to sun
- how far is earth to the sun
- how far is the earth away from the sun
- how far is the earth from sun
- how far is the earth from the sun
- how far is the earth to sun
- how far is the sun
- how far is the sun away from earth
- how far is the sun away from the earth
- how far is the sun from earth
- how far is the sun from the earth
- how far is the sun to earth
- how far is the sun to the earth
Different Levels of Semantic Matching

**Term**

- **Match exactly same terms**
  - NY $\rightarrow$ New York
  - disk $\rightarrow$ disc

**Word Sense**

- **Match terms with same meanings**
  - utube $\rightarrow$ youtube
  - NY $\rightarrow$ New York
  - motherboard $\rightarrow$ mainboard

**Topic**

- **Match topics of query and documents**
  - Microsoft Office $\xrightarrow{\times}$ ... working for Microsoft ... my office is in ...
  - Topic: PC Software
  - Topic: Personal Homepage

**Structure**

- **Match intent with answers (structures of query and document)**
  - Microsoft Office home $\rightarrow$ find homepage of Microsoft Office
  - 21 movie $\rightarrow$ find movie named 21
  - buy laptop less than 1000 $\rightarrow$ find online dealers to buy laptop with less than 1000 dollars
Semantic Matching Is Useful for

- General Search Relevance
- Vertical Search
- Entity Search
- Task Completion
SYSTEM VIEW OF SEMANTIC MATCHING
Overall System

Microsoft

Query

Online Query Processing

Query Representation

Ranked Documents

Semantic Matching

Query Knowledge

Query Index

Offline Query Processing

Document Representations

Document Index

Offline Document Processing

Search Log Data

Web Data
Online Query Processing

- **Named Entity Recognition in Query**
  - [michael I. jordan: PersonName]
  - [berkeley: Location]: academic

- **Query Topic Identification**
  - michael I. jordan berkeley: academic
  - michael I. jordan berkeley

- **Similar Query Finding**
  - michael I. jordan berkeley
  - michael jordan berkeley

- **Query Refinement**
  - michael jordan berkeley
Michael Jordan is Professor in the Department of Electrical Engineering
Online Semantic Matching

Semantic Matching can be conducted at different levels.
Correcting Errors in Query

search “windows onecare”

window onecar

windows onecare

Query Refiner > Search System
Structured Prediction Problem

windows  onecare

```
window  onecar
```

“Ideal” word sequence

Observed “noisy” word sequence

\[ y^* = \arg \max_y \Pr(y|x) \]

“ideal” query word sequence

original query word sequence
Conditional Random Fields for Query Refinement

Introducing Refinement Operations

Operations
Spelling: insertion, deletion, substitution, transposition, ...
Word Stemming: +s/-s, +es/-es, +ed/-ed, +ing/-ing, ...
Query Refinement Using Conditional Random Fields

\[
Pr(y, \overline{o}, \overline{z}|x) = \frac{1}{Z(x)} \prod_{i=1}^{n} (\phi(y_{i-1}, y_i)) \prod_{j_l=1}^{m_l} \phi(z_{ij_l}, o_{ij_l}, z_{ij_l-1})
\]
NAMED ENTITY MINING FROM QUERY LOG USING TOPIC MODEL
Named Entity Recognition in Query

- **harry potter**
  - Harry Potter (0.5)
  - Harry Potter - Book (0.4)
  - Harry Potter - Game (0.1)

- **harry potter film**
  - Harry Potter film (0.95)

- **harry potter author**
  - Harry Potter author (0.95)
Our Approach

• Using Query Log Data (or Click-through Data)
• Using Topic Model
• Weakly Supervised Latent Dirichlet Allocation
• vs Pasca’s work (named entity mining from log data, deterministic approach)
Seed and Query Log

<table>
<thead>
<tr>
<th>Keyphrase</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>final fantasy</td>
<td>300</td>
</tr>
<tr>
<td>final fantasy</td>
<td>120</td>
</tr>
<tr>
<td>final fantasy</td>
<td>50</td>
</tr>
<tr>
<td>gone with the wind</td>
<td>120</td>
</tr>
<tr>
<td>gone with the wind</td>
<td>10</td>
</tr>
<tr>
<td>gone with the wind</td>
<td>10</td>
</tr>
<tr>
<td>harry potter</td>
<td>1000</td>
</tr>
<tr>
<td>harry potter</td>
<td>650</td>
</tr>
<tr>
<td>harry potter</td>
<td>80</td>
</tr>
<tr>
<td>harry potter</td>
<td>20</td>
</tr>
<tr>
<td>harry potter</td>
<td>300</td>
</tr>
<tr>
<td>harry potter</td>
<td>200</td>
</tr>
<tr>
<td>harry potter</td>
<td>100</td>
</tr>
<tr>
<td>final fantasy</td>
<td>10</td>
</tr>
<tr>
<td>final fantasy</td>
<td>10</td>
</tr>
<tr>
<td>gone with the wind</td>
<td>250</td>
</tr>
<tr>
<td>harry potter</td>
<td>800</td>
</tr>
<tr>
<td>......</td>
<td></td>
</tr>
</tbody>
</table>
# Pseudo Documents of Named Entities

<table>
<thead>
<tr>
<th>Entity</th>
<th>Movie, Game</th>
<th>Movie, Book</th>
<th>Movie, Book, Game</th>
</tr>
</thead>
<tbody>
<tr>
<td>final fantasy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>120</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>120</td>
<td>80</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gone with the wind</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>120</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>120</td>
<td>80</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>harry potter</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>800</td>
<td>650</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>650</td>
<td>300</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>200</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Latent Dirichlet Allocation Model

$z$: Movie, Book, Game

$w$: #, # movie, # book, ....

$\theta$: distribution of classes for named entity

$\beta$: distribution of contexts for class
Weakly Supervised Latent Dirichlet Allocation

\[
p(D|\Theta) = \prod_{d=1}^{M} \int p(\theta_d | \alpha) \left( \prod_{n=1}^{N_d} p(z_{dn} | \theta_d)p(w_{dn} | z_{dn}, \beta) \right) d\theta_d
\]

\[
\log p(D | \Theta) + \lambda C(\Theta, y)
\]

\[
= \sum_{d=1}^{M} \log \int p(\theta_d | \alpha) \left( \prod_{n=1}^{N_d} p(z_{dn} | \theta_d)p(w_{dn} | z_{dn}, \beta) \right) d\theta_d
\]

\[
+ \sum_{d=1}^{M} \sum_{i=1}^{K} \lambda y_{di} \bar{z}_{di}
\]

\[
\bar{z}_{i} = \frac{1}{N} \sum_{n=1}^{N} z_{n}^i
\]

constraints

![Graph showing popularity of movies, books, games, and music](image)
PROJECT: CONTEXT AWARE SEARCH

Joint work with Daxin Jiang, Jian Pei, and others
Context aware Search:

Better Search Using Context Information
Search in Office

• Job related
Search at Home

- Household related
- Hobby and leisure
Search in Mobile Context

- Location
- Time
- Activity
Search in Social Context

• Community
Conventional Web Search
Search Intent and Context

• Suppose that user raises query “jaguar”

• If we know the user raises query “"BMW"” before “"jaguar"”
• Then we know that the user is likely to look for the car
Context of Search

- User usually conducts multiple related searches in a session
Context Information is Useful

- Example of search sessions

<table>
<thead>
<tr>
<th>SID</th>
<th>Search sessions</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Ford → Toyota → GMC → Allstate</td>
<td><a href="http://www.autohome.com">www.autohome.com</a></td>
</tr>
<tr>
<td>S2</td>
<td>Ford cars → Toyota cars → GMC cars → Allstate</td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>Ford cars → Toyota cars → Allstate</td>
<td><a href="http://www.allstate.com">www.allstate.com</a></td>
</tr>
<tr>
<td>S4</td>
<td>GMC → GMC dealers</td>
<td><a href="http://www.gmc.com">www.gmc.com</a></td>
</tr>
</tbody>
</table>
50% of users clicked car review site www.autohome.com after searching several car names.
Challenges in Context aware Search

- How to model context?
- How to learn context model from data?
- How to apply context model in search?
Our Approach

• Three Models
  – Sequential Model (Cao et al. KDD 2008)
  – Hidden Markov Model (Cao et al. WWW 2009)
  – Conditional Random Fields Model (Cao et al. SIGIR 2009)

• Large Scale Data Mining to Construct Models

• Using Learning Models to Make Prediction (Context aware Search)
Modeling Context by Sequential Model
Modeling Context by CRF

\[ q_1 \quad u_1 \quad c_1 \]
\[ \vdots \quad \vdots \quad \vdots \]
\[ q_i \quad u_i \quad c_i \]
\[ \vdots \quad \vdots \quad \vdots \]
\[ q_t \quad u_t \quad c_t \]
Modeling Context by HMM
TRAINING HIDDEN MARKOV MODEL
Training Very Large HMM

• **Challenge 1:**
  – EM algorithm needs determined number of hidden states.
  – However, in our problem, hidden states correspond to search intents, for which the number is unknown.

• **Our Solution:**
  – Conduct clustering on click-bipartite graph and view clusters as hidden states.
Training Very Large HMM

• Challenge 2:
  – Search log data contains hundreds of millions of sessions.
  – It is impractical to train HMM from such huge training data on single machine.

• Our Solution:
  – Deploy learning task on distributed system under map-reduce model
Training Very Large HMM

• Challenge 3:
  – Each machine needs to hold the values of all parameters.
  – Since search log data contains millions of unique queries and URLs, the space of parameters is extremely large.

• Our Solution:
  – Employ special initialization strategy based on the clusters mined from click-through bipartite
SUMMARY
Summary

• Web search relies on NLP and IR
• Query understanding = identify user search intent
• Query understanding needs
  – Large scale mining platform
  – Advanced NLP and IR technologies
  – Advanced statistical learning technologies
• Our Projects
  – LOGAL: search and browse log mining platform
  – Semantic Matching: improving tail query relevance
  – Context aware Search: better search using context information
• NLP Challenges and Technologies in Information Explosion Era