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
Rich Log Mining Applications

Log Mining Applications

Search Applications

Document Understanding

Query Understanding

Query Clustering

Query Classification

Query Segmentation

Query Reformulation

Document Annotation

Document Classification

Document Summarization

Document Clustering

Search UI Design

Contextual Search

Personalized Search

Monitoring and Feedback

Click Prediction

Query & URL Suggestion

Document Ranking

Search Result Diversification

Ads Applications

User Understanding

Query-Doc Matching
The Problem

A huge gap between the data and the applications

- Apps
  - Query Understanding
  - Document Understanding
  - User Understanding
  - Query-Doc Matching

- Log Data
  - Search Log
  - Toolbar Data
  - Ads. Log
  - Web Site Log
The Problem

Researcher or developer
1. Has to access the raw log data directly
2. Has to build the application from scratch

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

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

Data Platform
Log Objects Gallery (LOGAL)
### Query Histogram

**Example applications:**
- Query completion
- Query suggestion
- Query analysis

<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>
Click-through Bipartite

• Example applications
  – Document ranking
  – Search result clustering
  – Web page summarization
  – Query suggestion

click-through bipartite graph
Click Pattern

Query

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

Pattern 1 (count)

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

Pattern 2 (count)

…

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

Pattern n (count)

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

- Example applications
  - Doc ranking
  - Query suggestion
  - URL suggestion
  - User satisfaction prediction

User activities in session:
- Query
- Click:
  - Srch click
  - Ads 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
Same Search Intent Different Query Representations
Example = “Distance between Sun and Earth”

- "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 earth and the sun
- distance from earth to the sun
- distance from sun to earth
- distance from sun to the earth
- distance from the earth to the sun
- distance from the sun to earth
- distance from the sun to the earth
- distance of earth from sun
- distance between earth sun
- 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 from the sun
- how far earth is from the sun
- how far from earth is the sun
- how far from earth to sun
- how far from earth to the sun
- distance between sun and earth
Matching at Different Levels

- **Match between terms in query & document**
  - *NY* $\rightarrow$ *NY*
  - *youtube* $\rightarrow$ *youtube*

- **Match between word senses in query & document**
  - *tube* $\rightarrow$ *youtube*
  - *NY* $\rightarrow$ *New York*

- **Match between topics of query & document**
  - *Microsoft Office* $\rightarrow$ *Microsoft* ... *PowerPoint, Word, Excel...

- **Match between structures of query & document title**
  - *how far is sun from earth* $\rightarrow$ *distance between sun and earth*

- **Match between phrases in query & document**
  - *hot dog* $\rightarrow$ *hot dog*

- **Match between terms in query & document**
  - *NY* $\rightarrow$ *NY*
  - *youtube* $\rightarrow$ *youtube*
Semantic Matching Is Useful for

- General Search Relevance
- Vertical Search
- Entity Search
- Task Completion
SYSTEM VIEW OF SEMANTIC MATCHING
Offline Document Processing

Named Entity Recognition in Doc.

Document Topic Identification

Key Concept Identification

Tokenization

Model

Michael Jordan is Professor in the Department of Electrical Engineering
Online Semantic Matching

Matching can be conducted at different levels
QUERY REFINEMENT USING CRF MODEL
Correcting Errors in Query

search “windows onecare”

window onecar

Query Refiner

windows onecare

Search System
Structured Prediction Problem

windows  onecare

"Ideal" word sequence

window  onecar

Observed "noisy" word sequence

\[ y^* = \text{arg max}_y \Pr(y|x) \]

"ideal" query word sequence

original query word sequence
Conditional Random Fields for Query Refinement

Introducing Refinement Operations

\[
\Pr(y, o|x) = \frac{1}{Z(x)} \prod_{i=1}^{n} \phi(y_{i-1}, y_i) \phi(y_i, o_i, x)
\]

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

\[
Pr(y, o, z|x) = \frac{1}{Z(x)} \prod_{i=1}^{n} \varphi(y_{i-1}, y_{i}, \varphi(z_{i,j_i, 0}, z_{i,j_i, -1}))
\]
NAMED ENTITY MINING FROM QUERY LOG USING TOPIC MODEL
### Named Entity Recognition in Query

<table>
<thead>
<tr>
<th>Entity</th>
<th>Type</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>harry potter</td>
<td>Movie</td>
<td>0.5</td>
</tr>
<tr>
<td>harry potter</td>
<td>Book</td>
<td>0.4</td>
</tr>
<tr>
<td>harry potter</td>
<td>Game</td>
<td>0.1</td>
</tr>
<tr>
<td>harry potter film</td>
<td>Movie</td>
<td>0.95</td>
</tr>
<tr>
<td>harry potter author</td>
<td>Book</td>
<td>0.95</td>
</tr>
</tbody>
</table>

36
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>Query</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>final fantasy</td>
<td>300</td>
</tr>
<tr>
<td>final fantasy movie</td>
<td>120</td>
</tr>
<tr>
<td>final fantasy wallpaper</td>
<td>50</td>
</tr>
<tr>
<td>gone with the wind</td>
<td>120</td>
</tr>
<tr>
<td>gone with the wind movie</td>
<td>10</td>
</tr>
<tr>
<td>gone with the wind photos</td>
<td>10</td>
</tr>
<tr>
<td>harry potter</td>
<td>1000</td>
</tr>
<tr>
<td>harry potter book</td>
<td>650</td>
</tr>
<tr>
<td>gone with the wind book</td>
<td>80</td>
</tr>
<tr>
<td>gone with the wind summary</td>
<td>20</td>
</tr>
<tr>
<td>harry potter cheats</td>
<td>300</td>
</tr>
<tr>
<td>harry potter pics</td>
<td>200</td>
</tr>
<tr>
<td>harry potter summary</td>
<td>100</td>
</tr>
<tr>
<td>final fantasy xbox</td>
<td>10</td>
</tr>
<tr>
<td>final fantasy soundtrack</td>
<td>10</td>
</tr>
<tr>
<td>gone with the wind</td>
<td>250</td>
</tr>
<tr>
<td>harry potter movie</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></td>
<td></td>
</tr>
<tr>
<td># movie</td>
<td>120</td>
<td></td>
<td></td>
</tr>
<tr>
<td># wallpaper</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td># xbox</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td># soundtrack</td>
<td>10</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></td>
<td></td>
</tr>
<tr>
<td># movie</td>
<td>120</td>
<td></td>
<td></td>
</tr>
<tr>
<td># book</td>
<td>80</td>
<td></td>
<td></td>
</tr>
<tr>
<td># summary</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td># review</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td># photos</td>
<td>10</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></td>
<td></td>
</tr>
<tr>
<td># movie</td>
<td>800</td>
<td></td>
<td></td>
</tr>
<tr>
<td># book</td>
<td>650</td>
<td></td>
<td></td>
</tr>
<tr>
<td># cheats</td>
<td>300</td>
<td></td>
<td></td>
</tr>
<tr>
<td># pics</td>
<td>200</td>
<td></td>
<td></td>
</tr>
<tr>
<td># summary</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

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

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

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

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

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

constraints
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>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Ford → Toyota → GMC → Allstate</td>
</tr>
<tr>
<td></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>
</tr>
<tr>
<td></td>
<td><a href="http://www.autohome.com">www.autohome.com</a></td>
</tr>
<tr>
<td>S3</td>
<td>Ford cars → Toyota cars → Allstate</td>
</tr>
<tr>
<td></td>
<td><a href="http://www.allstate.com">www.allstate.com</a></td>
</tr>
<tr>
<td>S4</td>
<td>GMC → GMC dealers</td>
</tr>
<tr>
<td></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

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

\[ c_1, \ldots, c_i, \ldots, c_t \]

\[ q_1, \ldots, q_i, \ldots, q_t \]

\[ u_1, \ldots, u_i, \ldots, u_t \]
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
THANKS!